Identification of Novel Genes Associated with Atrial Fibrillation and Development of Atrial Fibrillation Predictive Models by Incorporating Polygenic Risk Scores and PheWAS-Derived Risk Factors

Shih-Yin Chen, PhD; Yu-Chia Chen, PhD; Ting-Yuan Liu, PhD; Kuan-Cheng Chang, MD, PhD; Shih-Sheng Chang, MD, PhD; Ning Wu, MD; Donald Lee Wu, DO; Rylee Kay Dunlap, BS, BA; Chia-Jung Chan, MS; Fuu-Jen Tsai, MD, PhD*.

BACKGROUND: Atrial fibrillation (AF) is the most common atrial arrhythmia and is subcategorized into numerous clinical phenotypes. Previous studies demonstrated that early-onset AF was associated with genetic loci among the certain populations.

OBJECTIVES: The objective of this study was to develop AF predictive models using AF-associated single-nucleotide polymorphisms (SNPs) selected from the Genome-Wide Association Study (GWAS) of a large cohort of Taiwanese and explore whether the models posed the prediction power for AF.

METHODS: 75,121 total subjects including 5,694 AF patients and 69,427 normal controls with the GWAS data were included in this study. The polygenic risk scores based on AF-associated SNPs were determined and then integrated with

Phenome-wide association study (PheWAS)-derived risk factors including clinical and demographic variables. The robust AF predictive models were developed through advanced statistical and machine learning techniques and then were evaluated in terms of discrimination, calibration, and clinical utility.

RESULTS: The results demonstrated that the top 30 significant SNPs associated with AF were located on chromosomes 10 and 16, which involved *NEURL1*, *SH3PXD2A*, *INA*, *NT5C2*, *STN1*, and *ZFHX3* genes with *INA*, *NT5C2*, and *STN1* being new discoveries in association with AF. The GWAS predictive power for AF had an area under the curve (AUC) of 0.626 ($P < 0.001$) and 0.851 ($P < 0.001$) before and after adjusting with age and gender, respectively. The results of PheWAS analysis showed that the top 10 diseases associated with discovered genes were all circulatory system diseases. The results of this study suggested that AF could be predicted by genetic information alone with moderate accuracy. The GWAS could be a robust and useful tool for detecting polygenic diseases by capturing the cumulative effects and genetic interactions of moderately associated but statistically significant SNPs.

CONCLUSIONS: By integrating genetic and phenotypic data, the accuracy and clinical relevance of predictive models for AF were improved. The results of this study might improve AF risk classification, enable personalized interventions, and ultimately reduce the burden of AF-related morbidity and mortality.

Keywords: genome-wide association study (GWAS), Phenome-wide association study (PheWAS), atrial fibrillation (AF), area under the curve (AUC), Ingenuity Pathway Analysis (IPA).

Short Title

AF novel SNPs and predictive models based on PRS and PheWAS-Derived Risk Factors.

Clinical Perspective

What Is New?

• Atrial fibrillation (AF) is thought to have a heritable component, but large studies, especially Asian populations, are lacking.

• Three novel loci were identified high associations to AF in this study.

• The results of this study suggested that AF could be predicted by genetic information alone with moderate accuracy.

• The GWAS could be a robust and useful tool for detecting polygenic diseases by capturing the cumulative effects and genetic interactions of moderately associated but statistically significant SNPs.

What Are the Clinical Implications?

• By integrating genetic and phenotypic data, the accuracy and clinical relevance of predictive models for AF were improved.

• The results of this study might improve AF risk classification, enable personalized interventions, and ultimately reduce the burden of AF-related morbidity and mortality.

• Additional prospective studies are needed to identify the genetic/genomic factors that contribute to these observed associations.

• The novel loci that had the high association to AF may have the potential to improve AF molecular genetical diagnosis.

Atrial fibrillation (AF) is a major cardiovascular disease affecting approximately 1.6% of the global population and is responsible for 20-25% of ischemic strokes and about 30% of heart failure. [1] AF features an irregular and fast heart rhythm, which can cause blood clots in the heart and increase the risks of stroke, ischemic heart attack, and even death. As a chronic degenerative heart disease, AF progresses from a paroxysmal to persistent type, and eventually becomes long-standing persistent and permanent AF.[2] Among the AF patients, with more than 50% of AF being asymptomatic, the early-stage low burden paroxysmal AF is hard to be detected by a

single electrocardiogram (ECG) examination. [3] However, once progressed to persistent AF, the patient's heart rhythm control becomes more difficult than that in the paroxysmal AF stage, and the AF recurrence rate is significantly increased.[2] The cause of AF is still unknown. Although it is common in people with hypertension, coronary artery disease, heart diseases, and other medical conditions such as hyperthyroidism, pneumonia, and chronic obstructive pulmonary disease, many people, especially athletes, may have it new onset without any known cause and symptoms. Recently, many studies demonstrated that AF, especially lone AF, is associated with an important genetic component.[4] Some studies found that family history was an important factor associated to AF with the risk of AF increased by more than 40% if one parent or sibling in the family had AF.[5] There is widespread sex, age, race, and ethnicity differences in AF epidemiology, risk factors, disease manifestations, clinical outcomes, and different reaction to therapies.[6] In addition, Mendelian Randomization (MR) associations analysis also suggested that the body height was associated with cardiovascular disease traits including AF.[7] All those evidence suggested that AF might be a genetically related heart disease.

Recent advancements in genomics and phenomics have provided valuable insights into the genetic and phenotypic risk factors associated with AF. Genome-wide

association studies (GWAS) is an approach to identify genes associated with a particular disease by studying the entire genome of a particular large group of patients, searching for the variants of single nucleotide polymorphisms (SNPs) that occur at multiple locations across the genome associated with a specific phenotype.[8] The GWAS for AF started in 2007.[9] Since then, more than 93,000 AF cases and more than 1 million controls have undergone the GWAS, and more than 134 distinct AF-associated loci were identified.[10] In addition to genetic factors, phenotypic data derived from phenome-wide association studies (PheWAS) can provide valuable insights into the non-genetic risk factors associated with AF. PheWAS analyze disease phenotypes and compared them to SNPs by using electronic medical record (EMR) data from EMR-linked databases to explore the relationships between a wide range of phenotypes and disease outcomes,[11] which is complementary to GWAS with the ability to replicate and validate the findings of GWAS. By analyzing the association between various clinical and demographic variables and the presence of a particular disease, PheWAS can identify potential risk factors and comorbidities associated with the development of such disease. Therefore, integrating PheWAS-derived risk factors into the predictive models for a particular disease can further enhance the accuracy and clinical utility by capturing a broader spectrum of risk factors beyond genetics alone.[12] However, there is no large sample size report of PheWAS for AF currently.

Integrating polygenic risk scores (PRS) has been used to predict the clinical phenotypes and outcomes of individuals. PRS captures the cumulative effect of multiple genetic variants across the genome, enabling the assessment of an individual's genetic susceptibility to a specific trait or disease.[13] Previous studies showed that an estimation of genetic susceptibility to AF was feasible with GWAS and could be used in the development of AF risk models.[14] Moreover, integrating PRS and PheWAS-derived risk factors also demonstrated a promising approach to develop predictive models for AF.[15] By leveraging genetic and phenotypic data, the predictive models for a particular disease have the potential to improve patient classification, identify high-risk individuals, and guide personalized prevention and management strategies. However, to take full advantage of the predictive power of the disease predictive models, a rigorous and comprehensive development and evaluation process through the combination of the data from GWAS, PheWAS, EMR, and PRS is essential.

Due to the unknown cause, insidious onset, severity of disease progression, poor prognosis, and lack of effective early diagnosis method, it is scientifically significant and clinically important to explore the potential AF mechanisms and develop an accurate method for AF prediction, early diagnosis, and prognosis evaluation. This

study employed a well-characterized clinical EMR database of AF patients to develop AF predictive models by incorporating PRS and PheWAS-derived risk factors through the data of GWAS, PheWAS, and EMR. The results of this study would greatly benefit both scientists and physicians in AF study, prevention, diagnosis, and prognosis evaluation. Further, the results would potentially improve the AF patient's risk classification, enable personalized interventions, and ultimately reduce the risks and damages to human health and the burden of AF-related morbidity and mortality.

METHODS

Data Resource and Patient Information

The GWAS data of 75,121 subjects including 5,694 patients with EKG confirmed AF and 69,427 patients with normal EKG from China Medical University Hospital (CMUH) (Taichung, Taiwan, ROC) during the time period of 1992 to 2020 were retrieved from the database of China Medical University Hospital Precision Medicine Project (Taichung, Taiwan, ROC). The clinical information of all involved patients was obtained from the electronic medical records (EMRs) of CMUH with the approval of CMUH ethics committees (Approval number: CMUH111-REC1-176, CMUH107-REC3-058, and CMUH110-REC3-005).

SNP Array Data Quality Control

The TPMv1 customized SNP array (Thermo Fisher Scientific, Santa Clara, California, USA) developed by the Academia Sinica and Taiwan Precision Medicine Initiative (TPMI) projects (Taipei, Taiwan, ROC) was employed with a total of 714,457 SNPs being included in this study. PLINK1.9 (https://www.cog-genomics.org/plink/1.9/)[16] was utilized for data analysis with the exclusions of SNPs and subjects based on the high rates of missingness per marker (geno 0.1 > 10%) for SNPs and per individual (mind 0.1 > 10%) for subjects. Variants with Hardy–Weinberg equilibrium (HWE) value less than 10-6 (HWE < 10^{-6}) and minor allele frequency (MAF) value less than 10^{-4} (MAF < 10^{-4}) were also filtered out. After such quality control process, 508,004 SNP variants and 75,121 subjects (5,694 AF and 69,427 normal control) remained in the study. Beagle 5.2 (https://faculty.washington.edu/browning/beagle/beagle.html) was applied for data imputation. The imputed data were further filtered by following the criteria of alternate allele dose less than 0.3 and genotype posterior probability less than 0.9.[17-19] After quality control and imputation, a total of 9,607,262 SNP variants and 75,121 subjects (5,694 AF and 69,427 normal control) were used for analysis (The flowchart of this study was shown in Supplementary Materials Files S1).

Genome-Wide Association Study (GWAS)

PLINK 1.9 was applied to generate summary statistics. For subjects who had EKG diagnosed AF recorded in EMR, logistic regression with multiple covariates including age and gender was performed, and the statistical significance was adjusted. Manhattan and quantile-quantile (QQ) plots with P values were generated by using R studio (https://posit.co/products/open-source/rstudio/).[20]

Biological Pathway Analysis

The canonical pathway enriched by differential metabolites was analyzed by using the QIAGEN Ingenuity Pathway Analysis (IPA) suite (http://www.ingenuity.com) to identify relevant biological pathways and functions.[21] The analysis was performed by integrating a group of different metabolites retrieved from Human Metabolome Database (HMDB), false discovery rate (FDR) value, and logarithmic fold change into IPA. Enrichment pathways of different metabolites were then generated based on the Ingenuity Pathway Knowledge Database (QIAGEN, Hilden, Germany).

Phenome-Wide Association Studies (PheWAS)

The primary PheWAS analysis used 47 SNPs identified from the GWAS of AF patients to detect the potential associations between the SNPs and phenotypes extracted from the EMR. The calculated polygenic risk score (PRS) for AF was

analyzed. Additionally, logistic regression was performed by using PLINK2 to examine the SNP association with phecode. A total of 97,735,180 International Statistical Classification of Diseases and Related Health Problems (ICD) version 9 (ICD-9) or version 10 (ICD-10) diagnosis codes were collapsed into 1,792 phecodes. The association between the PRS and each phecode was explored by using logistic regression models and the "PheWAS" R package.[22] The PheWAS results were combined in a meta-analysis of multiple populations with the significance determined by using Bonferroni correction.

AF predictive models

The regression models were then adjusted by patient's gender, age (at the enrollment), age squared, and the top 20 principal components. Ancestry-specific PheWAS was performed in all groups, and the summarized data were then put through the meta-analysis by using an inverse-variance weighted fixed-effects model implemented in the PheWAS R package.[22] The heterogeneity was assessed by using I^2 and any results with excess heterogeneity ($I^2 > 40\%$) were excluded. The association between SNP and phecodes was also performed with false discovery rate (FDR) *P* value less than 0.01 as the significant. Thus, the thresholds for significance were set as $P < 6.07 \times 10^{-5}$ for critical AF variants and $P < 4.13 \times 10^{-5}$ for hospitalized AF variants. In this study, FDR *P* value less than 0.01 was also employed as PheWAS

significant associations with increased or reduced risk for a specific condition.

Statistical Analysis

SPSS (IBM, Armonk, New York, USA) was employed for the statistical analysis of this study.[23, 24] The student's t-test was used for the comparisons between two groups, while ANOVA was applied for the one-way analysis of variance among the groups. P value less than 0.05 was defined as the significant difference while P value less than 0.01 was defined as the very significant difference.

RESULTS

Subjects and GWAS in Atrial Fibrillation Patients

A total of 75,121 patients were included in this study with 5,694 patients who were clinically confirmed AF by EKG. GWAS was conducted in all subjects and resulted in 1,715 SNPs that satisfied with significant levels ($P = 1\times10^{-5}$) of AF. (the flowchart of this study was shown in Supplementary Materials File S1 and the raw data was shown in Supplementary Materials File S2 and S3). The results showed that the SNPs on chromosomes 4, 10, and 16 even reached the significant threshold ($P < 5.0\times10^{-8}$). By comparing the results of AF group to the controls, the top 30 significant loci were identified including *NEURL1*, *SH3PXD2A*, *INA*, *NT5C2*, *STN1*, and *ZFHX3* genes

(Table 1) (raw data was shown in Supplementary Materials Files S4). The GWAS analysis results of AF were illustrated by using Manhattan plot and QQ plot (Figure 1).

Identification of Polygenetic Risk Score (PRS) with Risk Predictive Models

Based on the *P* and r2 values, a set of 47 SNPs were employed to construct the optimal PRS model. Two risk prediction models for potential AF patient were established through this study with one target model (Figure 2A and 2B) and one validation model (Figure 2C and 2D). The distributions of AF PRS were demonstrated in Figures 2A and 2C with the X-axis represented PRS of AF. The PRS percentiles among AF cases versus controls were illustrated in Figure 2B and 2D with the horizontal lines, the top and bottom of each box, and the whiskers reflected the median, quartile range, and the maximum and minimum values within each group, respectively. The area under the curve (AUC) was used in this study to evaluate the performance of PRSs and the GWAS predictive power, which required individual genotypic and phenotypic data in an independent GWAS validation dataset.[25] The GWAS predictive power for AF potential patient showed an AUC of 0.626 ($P < 0.001$) initially (Figure 3A), and after adjusted with age and gender, the GWAS predictive power for AF potential patient demonstrated an AUC of 0.851 ($P < 0.001$) (Figure 3B)

(raw data was shown in Supplementary Materials Files S5).

PheWAS in Atrial Fibrillation Patients

The association between AF PRS and 1,792 phecodes was presented by Manhattan plot. Figure 4A showed the Manhattan plot of PheWAS using top 10% PRS of AF patients *vs* the other 90% of AF patients, while Figure 4B showed the Manhattan plot of PheWAS using 4 groups of AF patients by quartile PRS. The leading top 10 diseases from the PheWAS analysis were listed in Table 2, which included atrial fibrillation and flutter, atrial fibrillation, cardiac dysrhythmias, arrhythmia (cardiac) NOS, congestive heart failure, atrial flutter, congestive heart failure (CHF) NOS, heart failure NOS, blastomycotic infection, and hypertensive heart disease (raw data was shown in Supplementary Materials Files S6 and S7).

Identification of Biological Pathways with Ingenuity Pathway Analysis (IPA)

Ingenuity pathway analysis (IPA) was employed in this study to investigate the potential biological pathways involved in the association of genetic polymorphisms and atrial fibrillation. The results demonstrated that cAMP response element-binding protein (CREB) was the major pathway associated with AF, which additionally implicated in numerous cellular pathways including gene transcription, cell growth,

proliferation hypertrophy, migration, and neointimal growth (Figure 5) (raw data was shown in Supplementary Materials Files S8).

DISCUSSION

GWAS Identification of Novel Genes Associated to AF

The results of GWAS in this study identified 1,715 SNPs that associated to AF and satisfied the predetermined significant levels, with the SNPs on chromosomes 4, 10, and 16 even surpassing the stringent threshold. Among them, the genes of *NEURL1, SH3PXD2A, INA, NT5C2, STN1,* and *ZFHX3* were found among the top 30 significant loci (SNPs) with *ZFHX3* located on chromosome 16 and the others located on chromosome 10 (Table 1). According to previous studies, *NEURL1* (Neuralized E3 Ubiquitin Protein Ligase 1) showed significant allelic and genotypic associations to AF and could significantly increase genetic susceptibility to AF.[26] *NEURL1* had been identified in AF patients of European descent and Japanese. The potential mechanism of *NEURL1* might be in prolongation of the atrial action potential duration, which had been confirmed by the knockdown of the zebrafish orthologs of *NEURL* experiment.[27] *SH3PXD2A* (SH3 And PX Domains 2A) was reported as an AF associated gene through a GWAS study including 11,300 AF cases and 153,676 controls.[28] Yang, et al. found that there was a replacement of Leu at 396 with Arg (L396R) in *SH3PXD2A*,

which enhanced migration of macrophages and their inflammatory features, resulting in enhanced susceptibility to AF.[29] In addition, disrupting the function of *SH3PXD2A* coding protein could affect axon guidance through locally inhibiting the degradation of the matrix in human neuron growth cone invadosome and disrupted motoneuron axons from exiting the spinal cord and extending into the periphery,[30] which might affect the neuronal innervation of heart. *ZFHX3* (Zinc Finger Homeobox 3) has been proved an associated gene to AF.[31] Zaw, et al. suggested that *ZFHX3* polymorphism was a risk marker for AF and AF-related phenotypes.[32] The potential mechanism of the association between *ZFHX3* and AF was examined by knock down *ZFHX3* in atrial myocytes, and the results demonstrated the dysregulated calcium homeostasis and increased atrial arrhythmogenesis, which might contribute to the occurrence of AF.[33]

Besides the known associations of *NEURL1, SH3PXD2A,* and *ZFHX*3 to AF, this study identified three novel loci, near *INA, NT5C2,* and *STN1* genes, that also demonstrated high associations to AF. INA, also known as alpha-internexin, is a highly specific structural component of neuronal axons as a scaffolding protein in axonal and dendritic branching and growth,[34] which plays an important role in neurite outgrowth and regulates the expression of other neurofilaments during neuronal development. However, overexpression of *INA* could induce apoptosis-like cell

death,[35] which might impact the autonomic neuronal regulation network of heart. *NT5C2* (5'-nucleotidase cytosolic II) has been found to have a significant association with coronary heart disease and hypertension.[36, 37] Cunningham, *et al.* reported 15 significant loci associated with the risks of incident myocardial infarction or coronary artery disease, stroke, heart failure, and aortic stenosis including *NT5C2* gene.[38] The expression of *NT5C2* is related to adenosine levels in cardiac endothelial cells and cardiomyocytes,[39] and adenosine results in a biphasic effect on heart rate, an initial period of sinus bradycardia followed within seconds by sinus tachycardia by increasing heart cell potassium conductance,[40, 41] which might be the potential cause of AF. *STN1* coding protein is a subunit of CST complex for telomere maintenance. The length of telomeres has been confirmed to be associated with a higher risk of ischemic heart disease both observationally and genetically.[42] Zheng, *et al.* reported that leucocyte telomere shortening was an independent predictor of AF, and even significantly associated with recurrent AF.[43] The study performed by Liu, *et al.* found that the leukocyte telomere length was inversely correlated with the occurrence of aging-related AF and that mitochondrial dysfunction played a role.[44] Therefore, *STN1* could be a potential risk factor of AF, especially in aged people.

Development of AF Risk Predictive Models

The risk predictive models of AF are useful tools for the prevention of AF episode and will help to reduce the complications and even life-threatening consequences of AF. PRS has been confirmed as a promising approach to accurately predict an individual's risk of developing disease, which makes it possible to be employed to develop AF predictive models for the assessment of individual genetic susceptibility of AF. Two risk prediction models including the target model and the validation model were developed through this study. The distribution of PRS scores between AF patients and normal controls demonstrated significant difference, indicating the potential application values of PRS as the AF predictive markers. The results of the area under the curve (AUC) for GWAS AF predictive power showed significant results both before and after adjusting by age and gender with the AUCs as 0.626 ($P < 0.001$) and 0.851 ($P < 0.001$), respectively. Comparing the results of this study to that of previous research, the results of Wong, et al. showed that the AUCs for the maximum clumping and thresholding PRSs was 0.599 for AF after integrating age and gender into the model designing,[45] while Torres, *et al.* reported that the AUC of age-related disease (ARD) – PRS was 0.57 for AF.[46] The comparison results indicated that the models developed through this study showed higher AUC values than that in the past studies even in the model before adjusting by age and gender, which suggested the potential clinical applications of these models in evaluation of individuals with high risks to

develop AF.

PheWAS Analysis

PheWAS analysis of this study revealed significant associations between AF and various phenotypic traits. The Manhattan plot displayed the associations between the top 10% PRS of AF patients and other 90% of AF patients, as well as the associations between quartile PRS groups based on polygenic risk scores. Notably, the PheWAS analysis results showed that 90% (9 out of 10) of the top 10 AF PRS associated diseases were cardiovascular system diseases with the only exception of blastomycotic infection. In addition, 50% (5 out of 10) of the top 10 AF PRS associated diseases were abnormal heart rate related diseases including atrial fibrillation and flutter, atrial fibrillation, cardiac dysrhythmias, arrhythmia, and atrial flutter. These findings provided further evidence of the broad impacts of AF PRS on the health of cardiovascular system and might be the explanation of the potential comorbidities associated with the AF.[47]

Ingenuity Pathway Analysis of Biological Pathways Associated with AF

Ingenuity Pathway Analysis (IPA) was applied to explore the biological pathways associated with AF for the purpose to identify the potential molecular signal

transduction mechanisms of AF. The IPA results identified CREB pathway as the major one that associated with AF, which additionally implicated in various cellular processes including gene transcription, cell growth, proliferation, migration, and neointimal growth. Previous study suggested that AF susceptibility was associated with decreased expression of the targets of activating transcription factors (ATF)/CREB family that included a group of transcription factors related to inflammation, oxidation, and cellular stress responses.[48] *ZFHX3*, identified as one of the 6 genes among the top 30 SNPs associated with AF in this study, was confirmed linking to CREB.[49] In addition, the expression of *NT5C2* was related to adenosine levels in cardiac endothelial cells and cardiomyocytes,[39] and cellular ATP production was related to the activation of CREB. Therefore, *NT5C2* and CREB together might affect the heart cell mitochondrial function[50] and cause cardiac cell function abnormality. On the other hand, these findings also conformed the results from previous studies that AF might have a significant influence on cellular responses and played a role in modulating cell growth processes.[51, 52] The results of this study suggested that AF associated genes might play crucial roles in regulating and controlling heart and/or neuronal cells, and potentially served as key modulators of cell growth and function. Further investigations were warranted to elucidate the precise mechanisms underlying these AF associated genes and their clinical

implications in the context of AF.

Limitations

1. Ethnic specificity: This study only included a large population from Taiwan. There are genetic variations and differences in disease risks among different ethnicities, requiring further research to validate the accuracy and reliability of these models and newly identified loci in other populations.

2. Gene-environment interactions: This study primarily focused on the role of genes to AF, while the effects of environmental factors on those genomic changes were not considered. The onset of atrial fibrillation may result from complex interactions between genes and the environment. Therefore, future research needs to comprehensively consider environmental factors to obtain more accurate predictive models.

3. Limitations of model evaluation: Although statistical and machine learning techniques were used to develop predictive models in this study, there were still some limitations in the evaluation process. While the models' discriminative power, calibration, and clinical utility were assessed, further studies with additional independent datasets is necessary to validate the performance of these models in real-world clinical applications.

4. Limitations in explaining genetic variations: Although the results of this study demonstrated that genetic information could be used to predict atrial fibrillation, genetic variations could only explain a portion of the disease risk. Other unknown genetic and non-genetic factors might also have significant influences on atrial fibrillation, which were not included in this study.

5. Lack of Functional Validation: Despite this study identified genetic loci associated to AF with statistical significance, further functional validation experiments were still needed to confirm the biological relevance of the identified genes and pathways and to elucidate the mechanisms through which these genetic variants contribute to AF development.

6. Limited Clinical Utility Assessment: Although the study evaluated the discrimination and calibration of the predictive models, the assessment of clinical utility, such as impact on patient outcomes or cost-effectiveness, was not investigated. Future studies should consider evaluating the real-world impact of implementing these models in clinical practice.

7. Potential biases: The data used in this study was derived from electronic medical records. Despite efforts to control confounding variables, the presence of unmeasured confounders and selection bias might not be fully ruled out which could influence the associations observed in the study and affect the validity of the predictive models.

Conclusions

This comprehensive study integrated GWAS, PRS, PheWAS, and IPA to unravel the genetic and phenotypic aspects of AF and developed potential AF risk prediction models. The findings of this study shed lights on the genetic variants, phenotypic associations, and potential biological pathways involved in AF. The results of this study indicated the important implications for risk classification, personalized interventions, and the understanding of the underlying mechanisms of AF. Further research is needed to validate these findings and explore the clinical applications in predicting AF risks, managing potential AF patients, and improving AF patient care.

ARTICLE INFORMATION

Affiliations

School of Chinese Medicine, China Medical University, Taichung, 404, Taiwan (S.Y.C., F.J.T.); Genetics Center, Department of Medical Research, China Medical University Hospital, Taichung 404, Taiwan (S.Y.C., G.J.C., F.J.T.); Million-Person Precision Medicine Initiative, Department of Medical Research, China Medical University Hospital, Taichung, 404, Taiwan (Y.C.C., T.Y.L.); Division of Cardiovascular Medicine, Department of Medicine, China Medical University Hospital,

Taichung 404, Taiwan (K.C.C., S.S.C.); School of Medicine, College of Medicine, China Medical University, Taichung 404, Taiwan (K.C.C., S.S.C.); Department of Biological Sciences, Southeastern Oklahoma State University, Durant, OK 74701, USA (N.W.); Department of Internal Medicine, University of Oklahoma Health Sciences Center, Tulsa, OK 74104, USA (D.L.W.); College of Osteopathic Medicine, Oklahoma State University Center for Health Sciences, Tulsa, OK 74107, USA (R.K.D.); Department of Medical Genetics, China Medical University Hospital, Taichung 404, Taiwan (F.J.T.).

Acknowledgments

Authors would like to thank the data exploration, statistical analysis, and the support of the iHi Clinical Research Platform from the Big Data Center of CMUH, and all colleagues at the Genetic Center, Department of Medical Research, China Medical University for their feedback and technical support.

Sources of Funding

This work was supported by China Medical University and China Medical University Hospital in Taiwan (grant nos. CMU110-N-30 and DMR-112-126).

Declaration of Interests

The authors declare no competing interests.

Supplemental Material

1. S1-flowchart
2. S2-AF cohort
3. S3-AF_GWAS data
4. S4-Table 1 data
5. S5-AUC analysis
6. S6-PheWAS data-1_10PRS
7. S7-PheWAS data-2_QuartilePRS
8. S8-IPA data

REFERENCES

1. Kim D, Yang PS, Jang E, Yu HT, Kim TH, Uhm JS, Kim JY, Pak HN, Lee MH, Joung B, Lip GY. 10-Year nationwide trends of the incidence, prevalence, and adverse outcomes of non-valvular atrial fibrillation nationwide health insurance data covering the entire Korean population. *Am Heart J* 2018;202:20–26. doi: 10.1016/j.ahj.2018.04.017.
2. Kirchhof P, Benussi S, Kotecha D, Ahlsson A, Atar D, Casadei B, Castella M,

Diener HC, Heidbuchel H, Hendriks J, et al. 2016 ESC guidelines for the management of atrial fibrillation developed in collaboration with EACTS. *Eur Heart J* 2016;37:2893–2962. doi: 10.1093/eurheartj/ehw210.

3. Steg PG, Alam S, Chiang CE, Gamra H, Goethals M, Inoue H, Krapf L, Lewalter T, Merioua I, Murin J, et al. Symptoms, functional status and quality of life in patients with controlled and uncontrolled atrial fibrillation: data from the RealiseAF cross-sectional international registry. *Heart* 2012;98:195–201. doi: 10.1136/heartjnl-2011-300550.
4. Lozano-Velasco E, Franco D, Aranega A, Daimi H. Genetics and Epigenetics of Atrial Fibrillation. *Int J Mol Sci*. 2020;21:5717. doi: 10.3390/ijms21165717.
5. Lubitz SA, Yin X, Fontes JD, Magnani JW, Rienstra M, Pai M, Villalon ML, Vasan RS, Pencina MJ, Levy D, et al. Association between familial atrial fibrillation and risk of new-onset atrial fibrillation. *JAMA* 2010;304:2263–2269. doi: 10.1001/jama.2010.1690.
6. Volgman AS, Bairey Merz CN, Benjamin EJ, Curtis AB, Fang MC, Lindley KJ , Pepine CJ, Vaseghi M, Waldo AL, Wenger NK, et al. Sex and race/ethnicity differences in atrial fibrillation. *J Am Coll Cardiol* 2019;74:2812-2815. doi: 10.1016/j.jacc.2019.09.045.
7. Raghavan S, Huang J, Tcheandjieu C, Huffman JE, Litkowski E, Liu C , Ho YA, Hunter-Zinck H, Zhao H, Marouli E, et al. A multi-population phenome-wide association study of genetically-predicted height in the Million Veteran Program. *PLoS Genet* 2022;18:e1010193. doi: 10.1371/journal.pgen.1010193.
8. Dehghan A. Genome-wide association studies. *Methods Mol Biol* 2018;1793:37-49. doi: 10.1007/978-1-4939-7868-7_4.
9. Gudbjartsson DF, Arnar DO, Helgadottir A, Gretarsdottir S, Holm H, Sigurdsson A, Jonasdottir A, Baker A, Thorleifsson G, Kristjansson K, et al. Variants

conferring risk of atrial fibrillation on chromosome 4q25. *Nature* 2007;448:353–357. doi: 10.1038/nature06007.

10. Roselli C, Chaffin MD, Weng LC, Aeschbacher S, Ahlberg G, Albert CM, Almgren P, Alonso A, Anderson CD, Aragam KG, et al. Multi-ethnic genome-wide association study for atrial fibrillation. *Nat Genet* 2018;50:1225–1233. doi: 10.1038/s41588-018-0133-9.
11. Bastarache L, Denny JC, Roden DM. Phenome-wide association studies. *JAMA* 2022;327:75-76. doi: 10.1001/jama.2021.20356.
12. Maclean RH, Ahmed F, Ong VH, Murray CD, Denton CP. A Phenome-Wide Association Study of Drugs and Comorbidities Associated With Gastrointestinal Dysfunction in Systemic Sclerosis. *J Rheumatol.* 2023;15:jrheum.220990. doi: 10.3899/jrheum.220990.
13. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. *Genome Med.* 2020;12:44. doi: 10.1186/s13073-020-00742-5.
14. Lubitz SA, Yin X, Lin HJ, Kolek M, Smith JG, Trompet S, Rienstra M, Rost NS, Teixeira PL, Almgren P, et al. Genetic risk prediction of atrial fibrillation. *Circulation* 2017;135:1311–1320. doi: 10.1161/CIRCULATIONAHA.116.024143.
15. Phulka JS, Ashraf M, Bajwa BK, Pare G, Laksman Z. Current State and Future of Polygenic Risk Scores in Cardiometabolic Disease: A Scoping Review. *Circ Genom Precis Med.* 2023;10:e003834. doi: 10.1161/CIRCGEN.122.003834.
16. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007;81:559-75. doi: 10.1086/519795.
17. Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from

Next-Generation Reference Panels. *Am J Hum Genet.* 2018;103:338-348. doi: 10.1016/j.ajhg.2018.07.015.

18. Liu TY, Lin CF, Wu HT, Wu YL, Chen YC, Liao CC, Chou YP, Chao D, Chang YS, Lu HF, Chang JG, et al. Comparison of multiple imputation algorithms and verification using whole-genome sequencing in the CMUH genetic biobank. *Biomedicine (Taipei)*. 2021;11:57-65. doi: 10.37796/2211-8039.1302.
19. Liao WL, Tsai FJ. Personalized medicine in Type 2 Diabetes. *Biomedicine (Taipei)*. 2014;4:8. doi: 10.7603/s40681-014-0008-z.
20. Liu TY, Liao WL, Wang TY, Chan CJ, Chang JG, Chen YC, Lu HF, Yang HH, Chen SY, Tsai FJ. Genome-wide association study of hyperthyroidism based on electronic medical record from Taiwan. *Front Med (Lausanne)*. 2022;9:830621. doi: 10.3389/fmed.2022.830621.
21. Chu Y, Jiang H, Ju J, Li Y, Gong L, Wang X, Yang W, Deng Y. A metabolomic study using HPLC-TOF/MS coupled with ingenuity pathway analysis: Intervention effects of Rhizoma Alismatis on spontaneous hypertensive rats. *J Pharm Biomed Anal.* 2016;117:446-52. doi: 10.1016/j.jpba.2015.09.026.
22. Carroll RJ, Bastarache L, Denny JC. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics.* 2014;30:2375–2376. doi: 10.1093/bioinformatics/btu197.
23. Liu SC, Tsai CH, Wu TY, Tsai CH, Tsai FJ, Chung JG, Huang CY, Yang JS, Hsu YM, Yin MC, et al. Soya-cerebroside reduces IL-1β-induced MMP-1 production in chondrocytes and inhibits cartilage degradation: implications for the treatment of osteoarthritis. *Food and Agric. Immunol.* 2019;30:620-632. doi: 10.1080/09540105.2019.1611745.
24. Liao WL, Liu TY, Cheng CF, Chou YP, Wang TY, Chang YW, Chen SY, Tsai FJ. Analysis of HLA Variants and Graves' Disease and Its Comorbidities Using a

High Resolution Imputation System to Examine Electronic Medical Health Records. *Front Endocrinol (Lausanne).* 2022;13:842673. doi: 10.3389/fendo.2022.842673.

25. Song L, Liu A, Shi J; Molecular Genetics of Schizophrenia Consortium. SummaryAUC: a tool for evaluating the performance of polygenic risk prediction models in validation datasets with only summary level statistics. *Bioinformatics* 2019;35:4038-4044. doi: 10.1093/bioinformatics/btz176.
26. Wang P, Qin W, Wang P, Huang Y, Liu Y, Zhang R, Li S, Yang Q, Wang X, Chen F, et al. Genomic variants in NEURL, GJA1 and CUX2 significantly increase genetic susceptibility to atrial fibrillation. *Sci Rep* 2018;8:3297. doi: 10.1038/s41598-018-21611-7.
27. Sinner MF, Tucker NR, Lunetta KL, Ozaki K, Smith JG, Trompet S, Bis JC, Lin H, Chung MK, Nielsen JB, et al. Integrating genetic, transcriptional, and functional analyses to identify 5 novel genes for atrial fibrillation. *Circulation* 2014;130:1225-1235. doi: 10.1161/CIRCULATIONAHA.114.009892.
28. Low SK, Takahashi A, Ebana Y, Ozaki K, Christophersen IE, Ellinor PT; AFGen Consortium; Ogishima S, Yamamoto M, Satoh M, et al. Identification of six new genetic loci associated with atrial fibrillation in the Japanese population. *Nat Genet* 2017;49:953-958. doi: 10.1038/ng.3842.
29. Yang X, Sasano T, Ebana Y, Takeuchi JK, Ihara K, Yamazoe M, Furukawa T. Functional role of the L396R mutation of Tks5 identified by an Exome-Wide Association Study in atrial fibrillation. *Circ J* 2020;84:2148-2157. doi: 10.1253/circj.CJ-20-0101.
30. Santiago-Medina M, Gregus KA, Nichol RH, O'Toole SM, Gomez TM. Regulation of ECM degradation and axon guidance by growth cone invadosomes. *Development* 2015;142:486-496. doi: 10.1242/dev.108266.

31. Wu L, Chu M, Zhuang W. Association between ZFHX3 and PRRX1 polymorphisms and atrial fibrillation susceptibility from meta-analysis. *Int J Hypertens* 2021;2021:9423576. doi: 10.1155/2021/9423576.

32. Zaw KTT, Sato N, Ikeda S, Thu KS, Mieno MN, Arai T, Mori S, Furukawa T, Sasano T, Sawabe M, et al. Association of ZFHX3 gene variation with atrial fibrillation, cerebral infarction, and lung thromboembolism: An autopsy study. *J Cardiol* 2017;70:180-184. doi: 10.1016/j.jjcc.2016.11.005.

33. Kao YH, Hsu JC, Chen YC, Lin YK, Lkhagva B, Chen SA, Chen YJ. ZFHX3 knockdown increases arrhythmogenesis and dysregulates calcium homeostasis in HL-1 atrial myocytes. *Int J Cardiol* 2016;210:85-92. doi: 10.1016/j.ijcard.2016.02.091.

34. Gaiottino J, Norgren N, Dobson R, Topping J, Nissim A, Malaspina A, Bestwick JP, Monsch AU, Regeniter A, Lindberg RL, et al. Increased neurofilament light chain blood levels in neurodegenerative neurological diseases. *PLoS One* 2013;8:e75091. doi: 10.1371/journal.pone.0075091.

35. Chien CL, Liu TC, Ho CL, Lu KS. Overexpression of neuronal intermediate filament protein alpha-internexin in PC12 cells. *J Neurosci Res* 2005;80:693-706. doi: 10.1002/jnr.20506.

36. Chen X, Zhang Z, Wang X, Chen Y, Wang C. NT5C2 Gene polymorphisms and the risk of coronary heart disease. *Public Health Genomics* 2020;23:90-99. doi: 10.1159/000507714.

37. Vishnolia KK, Hoene C, Tarhbalouti K, Revenstorff J, Aherrahrou Z, Erdmann J. Studies in Zebrafish demonstrate that CNNM2 and NT5C2 are most likely the causal genes at the blood pressure-associated locus on human chromosome 10q24.32. *Front Cardiovasc Med* 2020;7:135. doi: 10.3389/fcvm.2020.00135.

38. Cunningham JW, Di Achille P, Morrill VN, Weng LC, Choi SH, Khurshid S,

Nauffal V, Pirruccello JP, Solomon SD, Batra P, et al. Machine learning to understand genetic and clinical factors associated with the pulse waveform dicrotic notch. *Circ Genom Precis Med* 2023;16:e003676. doi: 10.1161/CIRCGEN.121.003676.

39. Le DE, Davis CM, Wei K, Zhao Y, Cao Z, Nugent M, Scott KLL, Liu L, Nagarajan S, Alkayed NJ, et al. Ranolazine may exert its beneficial effects by increasing myocardial adenosine levels. *Am J Physiol Heart Circ Physiol* 2020;318:H189-H202. doi: 10.1152/ajpheart.00217.2019.
40. Gupta A, Lokhandwala Y, Rai N, Malviya A. Adenosine-A drug with myriad utility in the diagnosis and treatment of arrhythmias. *J Arrhythm* 2020;37:103-112. doi: 10.1002/joa3.12453.
41. Belardinelli L, Isenberg G. Isolated atrial myocytes: adenosine and acetylcholine increase potassium conductance. *Am J Physiol* 1983;244:H734-737. doi: 10.1152/ajpheart.1983.244.5.H734.
42. Scheller Madrid A, Rode L, Nordestgaard BG, Bojesen SE. Short telomere length and ischemic heart disease: observational and genetic studies in 290 022 individuals. *Clin Chem* 2016;62:1140-9. doi: 10.1373/clinchem.2016.258566.
43. Zheng Y, Zhang N, Wang Y, Wang F, Li G, Tse G, Liu T. Association between leucocyte telomere length and the risk of atrial fibrillation: An updated systematic review and meta-analysis. *Ageing Res Rev* 2022;81:101707. doi: 10.1016/j.arr.2022.101707.
44. Liu C, Bai J, Dan Q, Yang X, Lin K, Fu Z, Lu X, Xie X, Liu J, Fan L, et al. Mitochondrial dysfunction contributes to aging-related atrial fibrillation. *Oxid Med Cell Longev* 2021;2021:5530293. doi: 10.1155/2021/5530293.
45. Wong CK, Makalic E, Dite GS, Whiting L, Murphy NM, Hopper JL, Allman R. Polygenic risk scores for cardiovascular diseases and type 2 diabetes. *PLoS One*

2022;17:e0278764. doi: 10.1371/journal.pone.0278764.

46. Torres GG, Dose J, Hasenbein TP, Nygaard M, Krause-Kyora B, Mengel-From J, Christensen K, Andersen-Ranberg K, Kolbe D, Lieb W, et al. Long-lived individuals show a lower burden of variants predisposing to age-related diseases and a higher polygenic longevity score. *Int J Mol Sci* 2022;23:10949. doi: 10.3390/ijms231810949.
47. Larsson SC, Wang L, Li X, Jiang F, Chen X, Mantzoros CS. Circulating lipoprotein(a) levels and health outcomes: Phenome-wide Mendelian randomization and disease-trajectory analyses. *Metabolism*. 2022;137:155347. doi: 10.1016/j.metabol.2022.155347.
48. Deshmukh A, Barnard J, Sun H, Newton D, Castel L, Pettersson G, Johnston D, Roselli E, Gillinov AM, McCurry K, et al. Left atrial transcriptional changes associated with atrial fibrillation susceptibility and persistence. *Circ Arrhythm Electrophysiol* 2015;8:32-41. doi: 10.1161/CIRCEP.114.001632.
49. Kim TS, Kawaguchi M, Suzuki M, Jung CG, Asai K, Shibamoto Y, Lavin MF, Khanna KK, Miura Y. The ZFHX3 (ATBF1) transcription factor induces PDGFRB, which activates ATM in the cytoplasm to protect cerebellar neurons from oxidative stress. *Dis Model Mech* 2010;3:752-762. doi: 10.1242/dmm.004689.
50. Niu Z, Tang J, Ren Y, Feng W. Ropivacaine impairs mitochondrial biogenesis by reducing PGC-1α. *Biochem Biophys Res Commun* 2018;504:513-518. doi: 10.1016/j.bbrc.2018.08.186.
51. Ngwa JS, Manning AK, Grimsby JL, Lu C, Zhuang WV, Destefano AL. Pathway analysis following association study. *BMC Proc*. 2011;5 Suppl 9:S18. doi: 10.1186/1753-6561-5-S9-S18.
52. Yu F, Shen XY, Fan L, Yu ZC. Genome-wide analysis of genetic variations

assisted by Ingenuity Pathway Analysis to comprehensively investigate potential genetic targets associated with the progression of hepatocellular carcinoma. *Eur Rev Med Pharmacol Sci.* 2014;18:2102-2108.

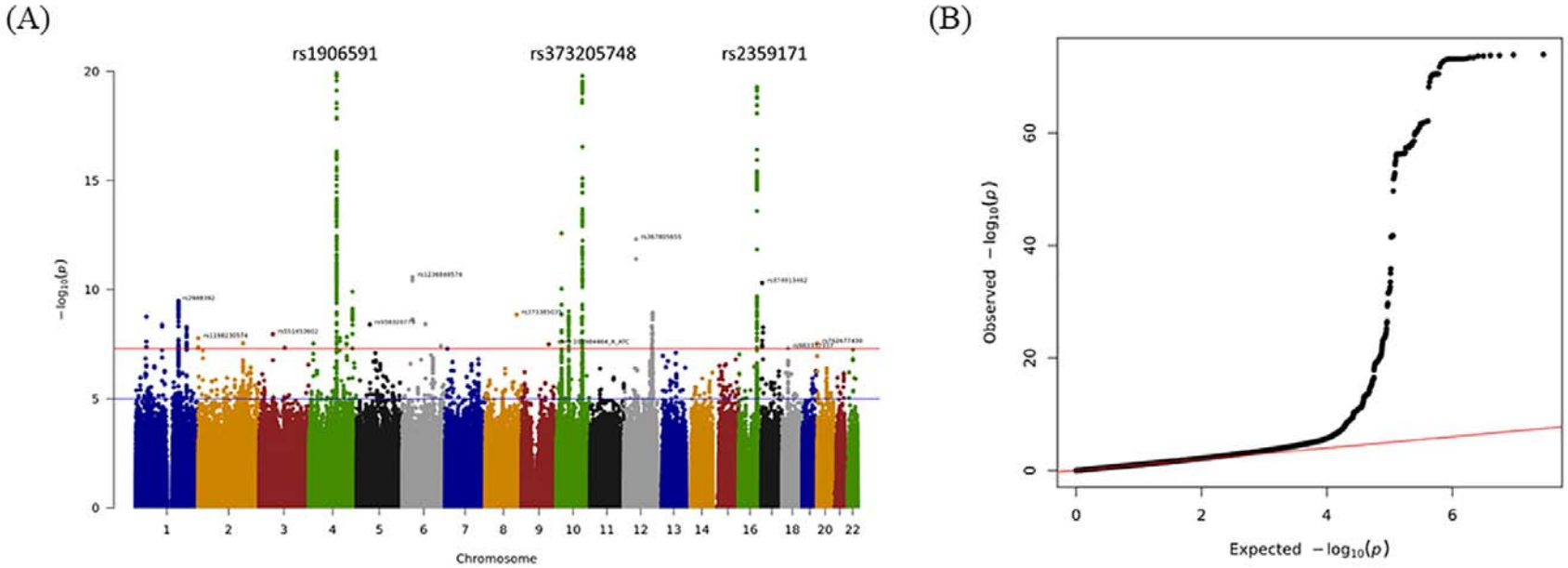


Figure 1. Manhattan plot (A) and quantile-quantile plot (B) for the genome-wide association study (GWAS) of ICD-10 code 148 (Ischemic Atrial Fibrillation). The upper and lower lines indicate the genome-wide significance threshold ($p = 5.0 \times 10^{-8}$) and the cut-off level for selecting SNPs ($p = 1 \times 10^{-5}$), respectively (A). The plot shows no significant deviation from the expected line, suggesting a lack of systematic biases or confounding effects in the analysis (B).

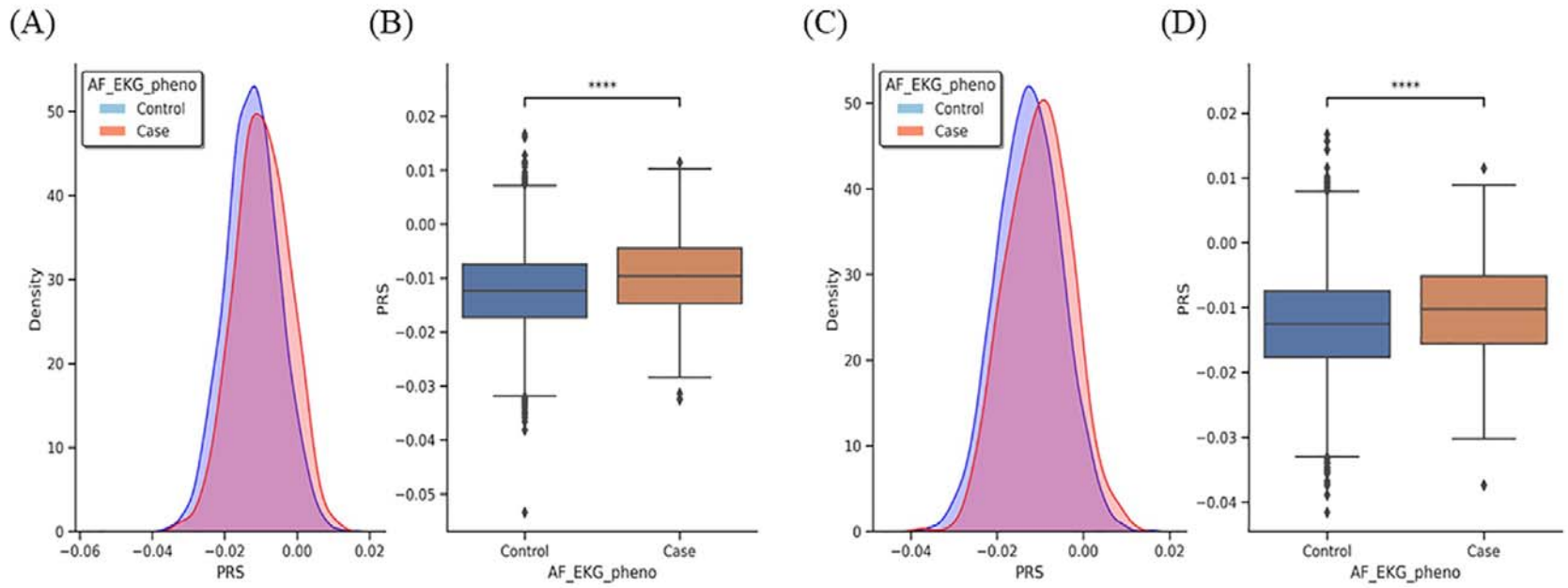


Figure 2. Risk analysis for atrial fibrillation (AF) based on polygenic risk score (PRS). Two AF risk predictive models were established as target model (A and B) and validation model (C and D). A and C: the distribution of AF PRS. B and D: PRS percentile among AF cases versus controls (the horizontal lines in each boxplot were the median; the top and bottom of each box were the quartile range; the whiskers were the maximum and minimum values within each group). **** represented *p*-value < 0.001.

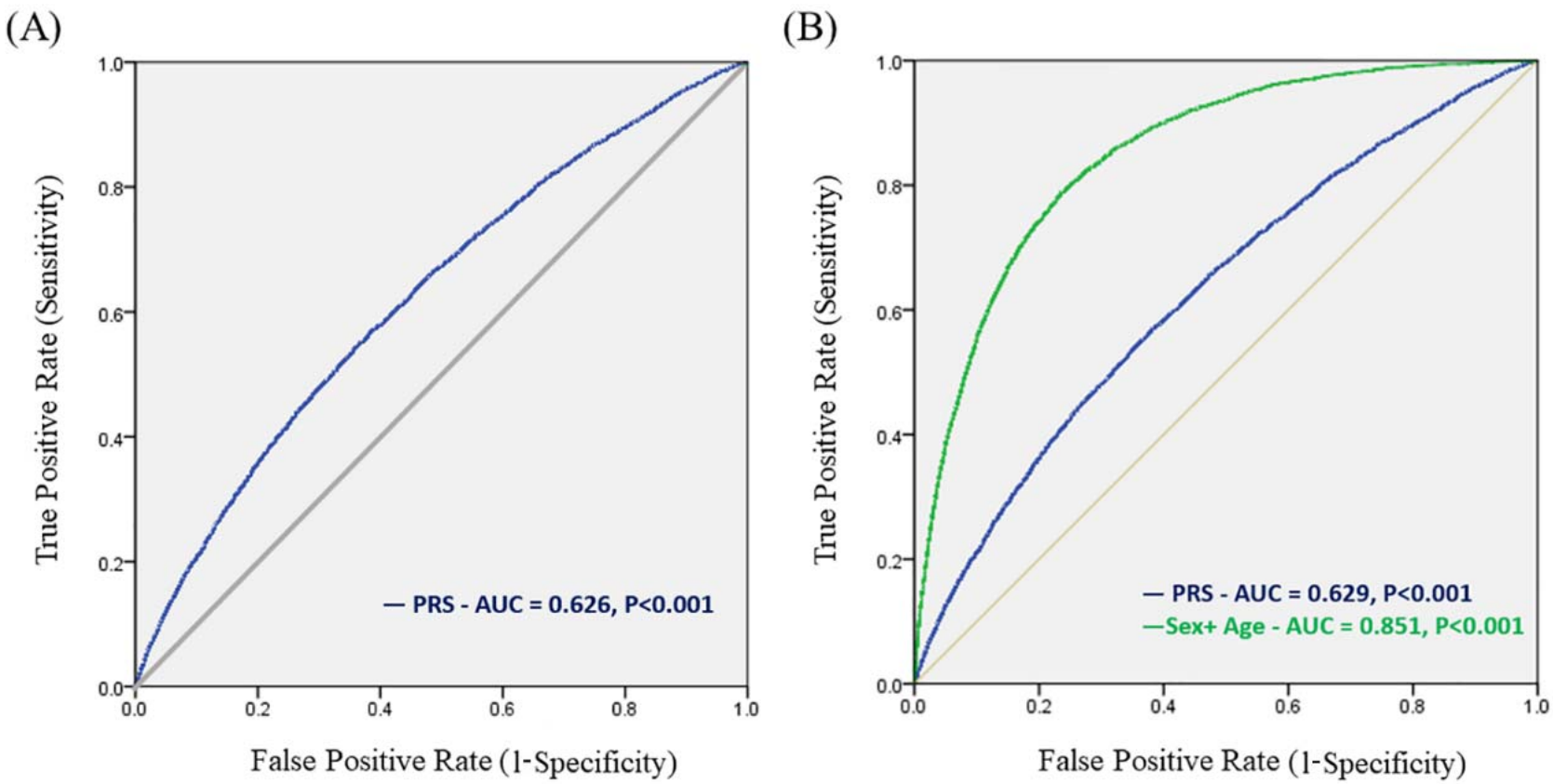


Figure 3. Area under the receiver-operator curves (AUC) for AF risk predictive models. (A). The GWAS predictive power for AF potential patient showed an AUC of 0.626 ($P < 0.001$). (B). The GWAS predictive power, after adjusted with age and gender, for AF potential patient demonstrated an AUC of 0.851 ($P < 0.001$).

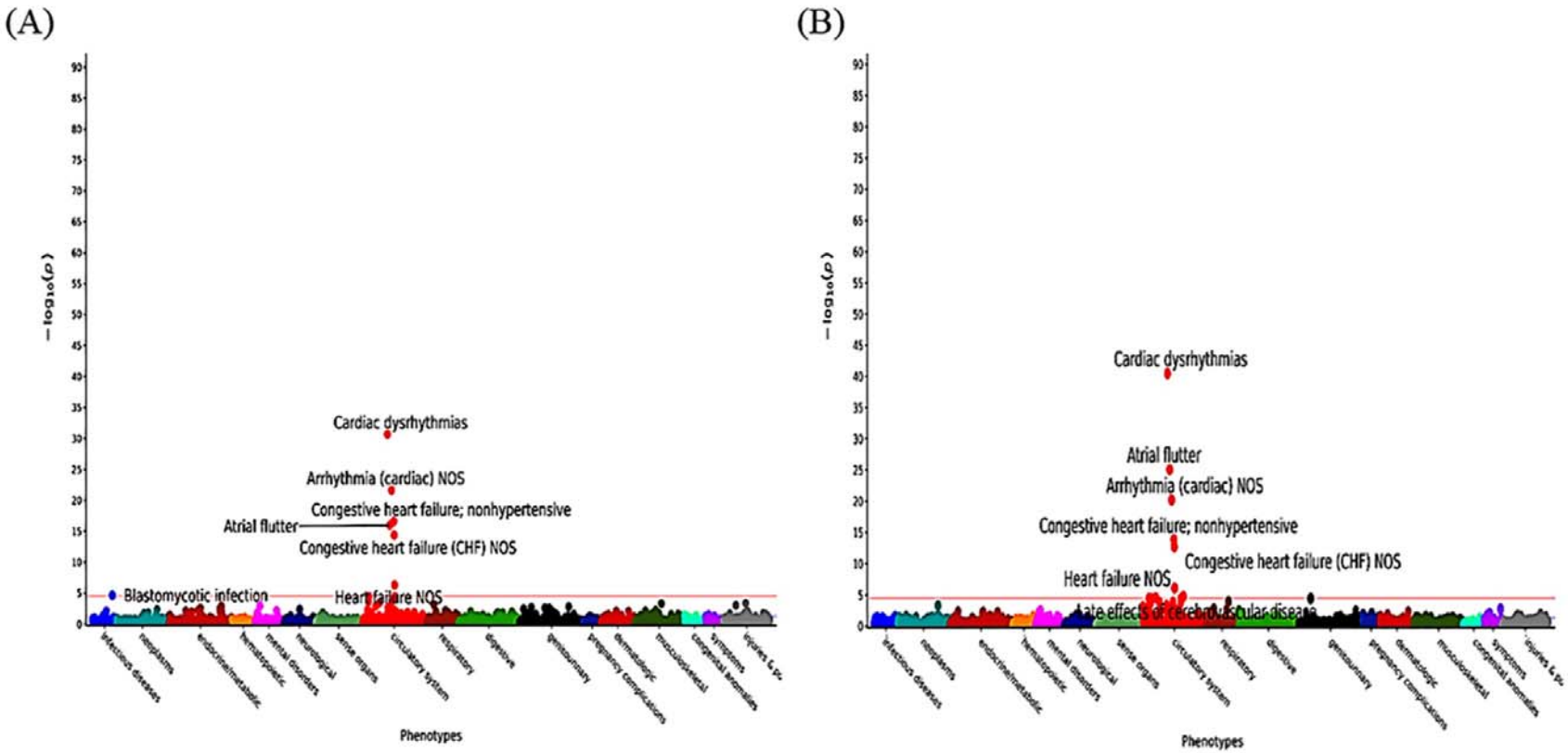


Figure 4. Manhattan plot of AF PRS and PheWAS presenting the association between AF PRS and 1,792 phecodes. (A): Manhattan plot of PheWAS using top 10% AF PRS patients *vs* the other 90% of AF patients. (B): Manhattan plot of PheWAS using 4 groups of AF patients by quartile PRS.

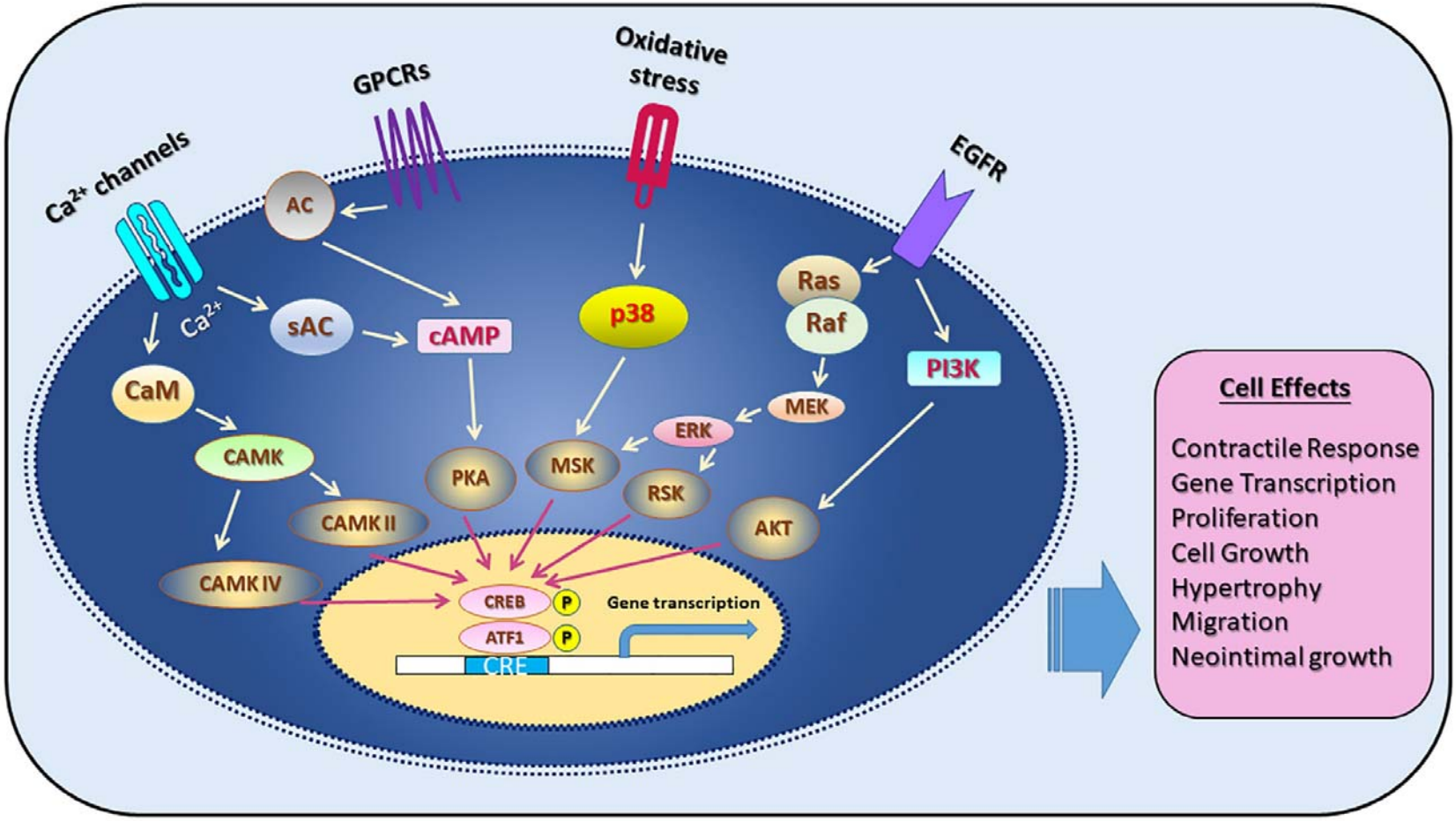


Figure 5. Identification and functional analysis of genes and relevant biological pathways that associated with AF by using GWAS and Ingenuity Pathway Analysis (IPA).

Table 1. The top 30 SNPs from the discovery- and replication- GWAS.

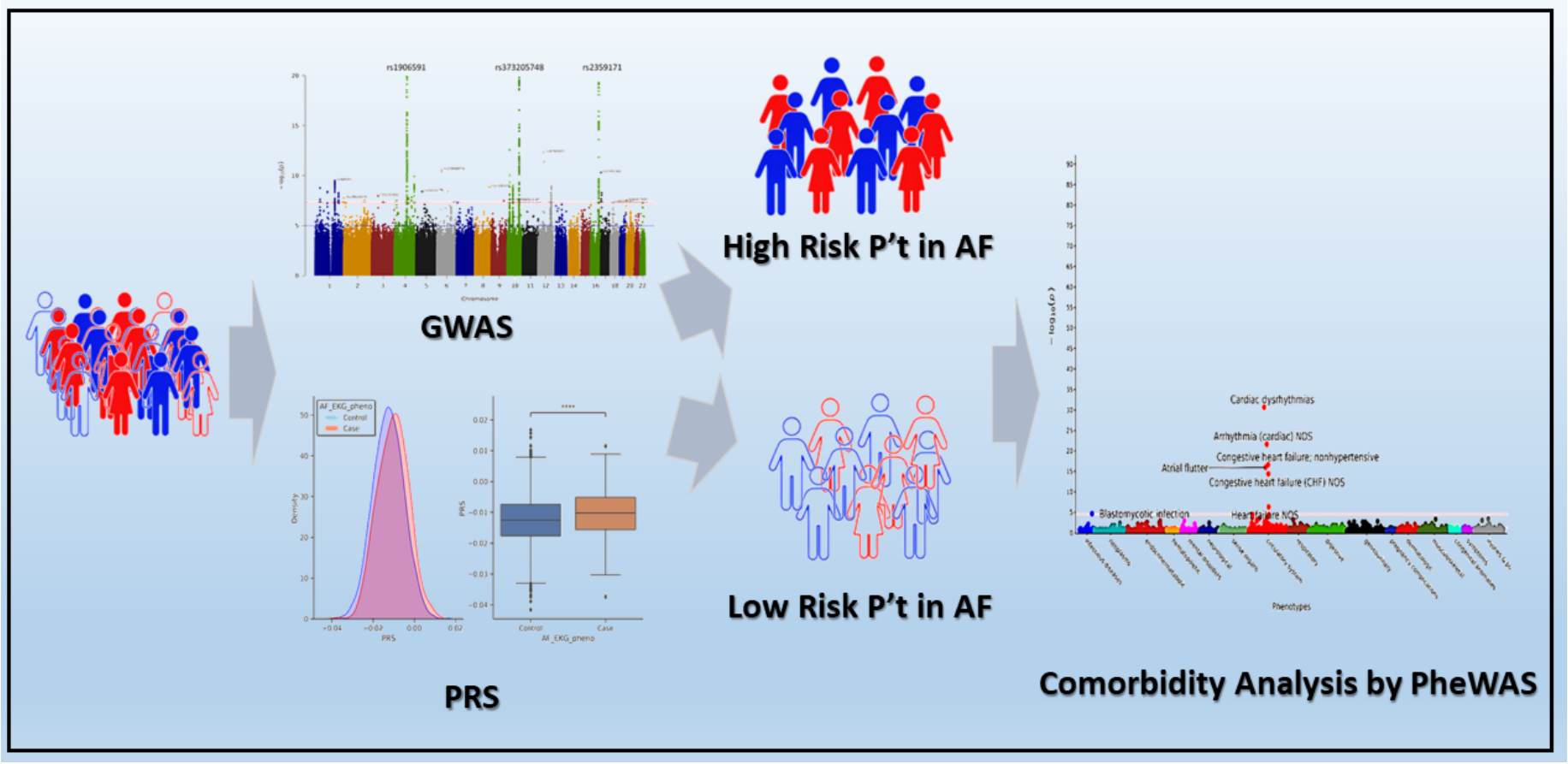
No.	Rank*	SNP ID	Reference sequence	Altered sequence	Gene symbol	Gene Name	Chromosome
1	1	rs373205748	C	T	*NEURL1*	Neuralized E3 ubiquitin protein ligase 1	10
2	2	rs185158502	A	G			
3	4	rs184385002	G	C			
4	12	rs12415501	C	T			
5	15	rs11815048	G	A			
6	17	rs151023098	TTTTG	T			
7	18	rs148177794	G	GT			
8	19	rs60469668	A	G			
9	20	rs12411463	C	T			
10	21	rs11598047	A	G			
11	24	rs12253987	T	A			
12	3	rs202011870	A	C	*SH3PXD2A*	SH3 and PX domains 2A	
13	5	rs186141893	C	A			
14	6	rs377070929	C	T			
15	7	rs183474824	T	A			
16	8	rs148633574	G	T	*INA*	Alpha-internexin	
17	10	rs188256230	G	A	*NT5C2*	5'-nucleotidase cytosolic II	
18	16	rs183689570	C	T	*STN1*	STN1 subunit of CST complex	
19	9	rs2359171	T	A	*ZFHX3*	Zinc finger homeobox 3	16
20	11	rs2106261	C	T			
21	13	rs4499262	C	A			
22	14	rs67402452	C	T			
23	22	rs12051512	G	A			
24	23	rs879324	G	A			
25	25	rs4404097	G	A			
26	26	rs4309429	C	T			
27	27	rs59275616	G	C			
28	28	rs67329386	C	T			
29	29	rs57828240	T	G			
30	30	rs8057081	C	T			

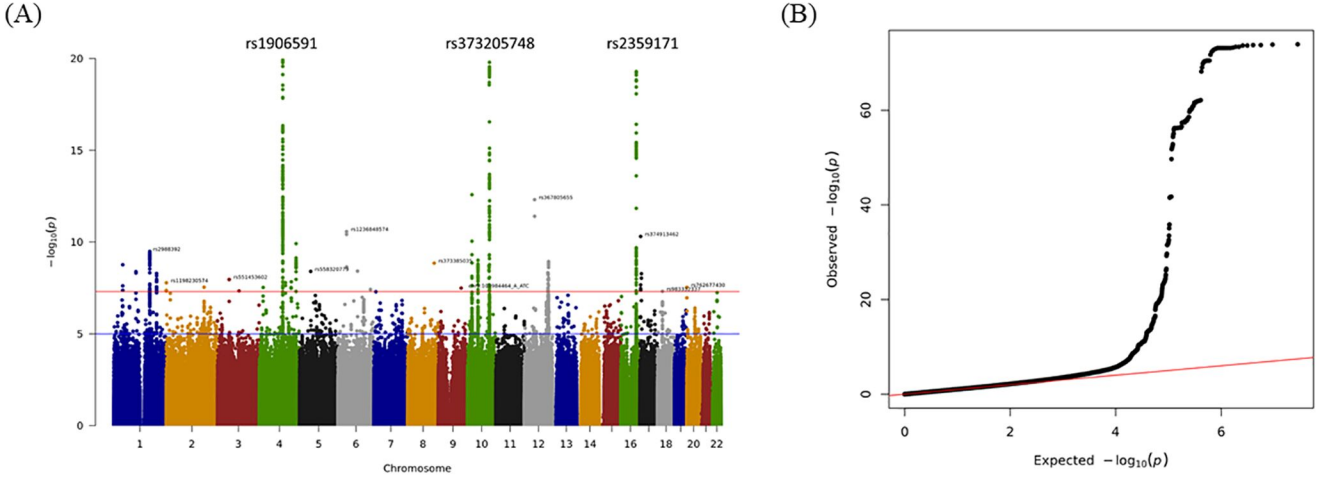
*Rankings are based on the *P* values.

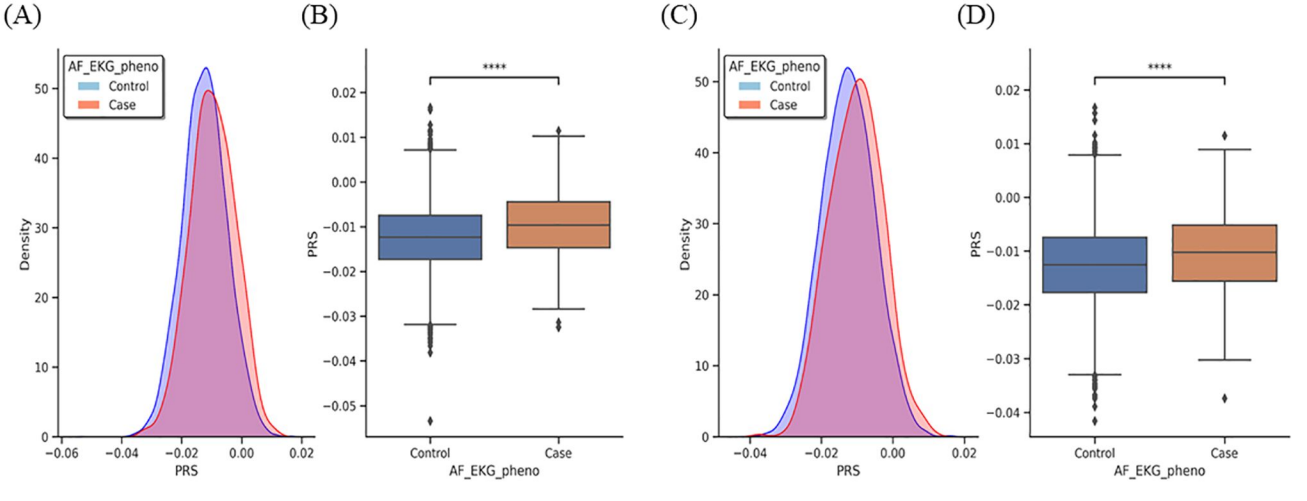
Table 2. The top 10 diseases from the PheWAS analysis.

Phecode	Description	Group	*p* value
427.2	Atrial fibrillation and flutter	circulatory system	9.33E-140
427.21	Atrial fibrillation	circulatory system	4.47E-138
427	Cardiac dysrhythmias	circulatory system	2.17E-31
427.5	Arrhythmia (cardiac) NOS	circulatory system	2.39E-22
428	Congestive heart failure; nonhypertensive	circulatory system	2.36E-17
427.22	Atrial flutter	circulatory system	1.10E-16
428.1	Congestive heart failure (CHF) NOS	circulatory system	4.12E-15
428.2	Heart failure NOS	circulatory system	4.77E-07
117.3	Blastomycotic infection	infectious diseases	2.18E-05
401.21	Hypertensive heart disease	circulatory system	3.41E-05

ABSTRACT GRAPHICS




(A)
rs1906591
rs373205748
rs2359171
rs2968392
rs1198230574
rs551453602
rs1236849574
rs373385032
rs367805655
rs374913462
rs2267430
−log10(p)
Chromosome
(B)
Observed −log10(p)
Expected −log10(p)


(A)
(B)
(C)
(D)
AF_EKG_pheno
Control
Case
Density
PRS
****
Control
Case
AF_EKG_pheno