Front. Med.
https://doi.org/10.1007/s11684-023-0982-1

Machine learning modeling identifies hypertrophic cardiomyopathy subtypes with genetic signature

Jiaqi Dai[1,3,*], Tao Wang[2,*], Ke Xu[1,3,*], Yang Sun[1,3], Zongzhe Li[1,3], Peng Chen[1,3], Hong Wang[3], Dongyang Wu[3], Yanghui Chen[3], Lei Xiao[3], Hao Liu[3], Haoran Wei[3], Rui Li[1,3], Liyuan Peng[1], Ting Yu[1], Yan Wang[1,3], Zhongsheng Sun (✉)[2], Dao Wen Wang (✉)[1,3]

[1]Division of Cardiology, Department of Internal Medicine, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China; [2]Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing 100101, China; [3]Hubei Key Laboratory of Genetics and Molecular Mechanism of Cardiologic Disorders, Huazhong University of Science and Technology, Wuhan 430030, China

Abstract **Previous studies have revealed that patients with hypertrophic cardiomyopathy (HCM) exhibit differences in symptom severity and prognosis, indicating potential HCM subtypes among these patients. Here, 793 patients with HCM were recruited at an average follow-up of 32.78 ± 27.58 months to identify potential HCM subtypes by performing consensus clustering on the basis of their echocardiography features. Furthermore, we proposed a systematic method for illustrating the relationship between the phenotype and genotype of each HCM subtype by using machine learning modeling and interactome network detection techniques based on whole-exome sequencing data. Another independent cohort that consisted of 414 patients with HCM was recruited to replicate the findings. Consequently, two subtypes characterized by different clinical outcomes were identified in HCM. Patients with subtype 2 presented asymmetric septal hypertrophy associated with a stable course, while those with subtype 1 displayed left ventricular systolic dysfunction and aggressive progression. Machine learning modeling based on personal whole-exome data identified 46 genes with mutation burden that could accurately predict subtype propensities. Furthermore, the patients in another cohort predicted as subtype 1 by the 46-gene model presented increased left ventricular end-diastolic diameter and reduced left ventricular ejection fraction. By employing echocardiography and genetic screening for the 46 genes, HCM can be classified into two subtypes with distinct clinical outcomes.**

Keywords machine learning methods; hypertrophic cardiomyopathy; genetic risk

Introduction

Hypertrophic cardiomyopathy (HCM) is the most common inherited heart disorder, affecting 1 in 200–500 adults worldwide [1,2]. It is characterized by left ventricular (LV) hypertrophy and measured via echocardiography or other imaging techniques. HCM is believed to be the most common cause of sudden cardiac death among adolescents and young adults. However, the majority of patients with HCM experience normal or near-normal life and typically remain clinically silent [3,4]. Significant differences in clinical manifestation and prognosis among individuals with HCM have elicited the interests of researchers to investigate the underlying mechanisms.

Several hypotheses have been proposed to illustrate extreme phenotypic variability. One such hypothesis defines a pattern of disease progression for HCM as the end-stage or “burnout” phase [5]; this process is consecutive from onset time to end stage, with adverse cardiac remodeling. Patients in the early phases are frequently asymptomatic, and hypertrophic phenotype is generally absent. As the disease evolves, advanced functional deterioration occurs in the left ventricle; this condition is defined as either a hypokinetic-dilated phase or a restrictive phenotype. However, previous cohort studies have suggested that only a small proportion of patients with HCM progressed to the end stage [6,7].

Received July 25, 2022; accepted January 5, 2023

Correspondence: Dao Wen Wang, dwwang@tjh.tjmu.edu.cn;
Zhongsheng Sun, sunzsbiols@126.com

[*] Jiaqi Dai, Tao Wang, and Ke Xu contributed equally to this work.

Why some patients undergo remodeling and progression while others do not has not yet been elucidated. Its highly diversified clinical symptoms have prompted us to speculate that HCM may have distinct subtypes.

Recent studies have revealed that patients with HCM caused by different mutations exhibit differences in symptom severity and prognosis [8,9]. Genes or modifiable risk factors reportedly influence the phenotypic severity of HCM [10]. However, modifier genes and their variants remain largely unknown. Moreover, the current approach is insufficient for quantitatively estimating risk for HCM progression, and such estimation is highly desired in clinical practice.

In the current study, a consensus clustering approach was applied to identify the clinical subtypes of HCM on the basis of echocardiography data. Interestingly, two major clinical subtypes were identified, delineating the diversity of clinical outcomes in HCM. By using machine learning methods, we identified subtype-associated genes that could effectively distinguish HCM subtypes, providing further insights into the genetic context, clinical prognosis, and potential interventions. Collectively, these findings may help bridge knowledge gaps among phenotypes, genotypes, and prognoses.

Materials and methods

Study population

The study cohort comprised 793 sporadic patients with HCM from 2007 to 2019 recruited from Tongji Hospital, Wuhan, China. HCM was diagnosed as a maximal end-diastolic LV wall thickness ⩾15 mm in echocardiographs or cardiac magnetic resonance images, in the absence of abnormal loading conditions or other cardiac or systemic diseases capable of producing the magnitude of hypertrophy, e.g., congenital heart diseases, aortic stenosis, uncontrolled hypertension, or phenocopied conditions. More limited hypertrophy (13–14 mm) was diagnosed for patients with a family history of HCM [11,12]. Peripheral blood samples were obtained from all the participants upon enrollment. All clinical variables, particularly echocardiography characteristics, 24 h Holter, and natural history, were collected from patients’ medical records that were blinded to patient genotype at the start of the study.

Clinical subtype identification

The R package ConsensusClusterPlus v1.48.0 was employed to identify clinical subtypes on the basis of echocardiography features [13]. This function provides quantitative stability evidence for determining cluster count and membership in an unsupervised analysis. In particular, the consensus clustering method involves subsampling from a set of items, and it determines clusters of specified cluster counts. Then, pairwise consensus values, i.e., the proportion in which two items occupied the same cluster out of the number of times they occurred in the same subsample, are calculated and stored in a symmetric consensus matrix for each cluster count. Fig. S1 shows all echocardiography variables that are commonly used to evaluate cardiac structure and function. Among them, three variables with more than 20% missing data were excluded from further analysis. The seven remaining variables, namely, the thickness of the interventricular septum (IVS) and left ventricular posterior wall (LVPW), left ventricular end-diastolic diameter (LVEDD), left atrial diameter (LAD), left ventricular ejection fraction (LVEF), and the ratios E/A (mitral inflow velocity curves) and septal E/E′ (annular tissue Doppler signals) ratios, were used for the clustering analysis. Agglomerative hierarchical clustering was performed with a subsampling ratio of 0.8 for 1000 iterations. Consensus matrices, subtype-consensus plots, and item-consensus plots were used to determine the optimal number of subtypes.

Simplification of clustering by creating a decision tree

The overrepresentation or underrepresentation of a variable in each subtype was calculated via v-test with the catdes function of the R package FactoMineR v2.0 on the basis of hypergeometric distribution. The contribution of each echocardiography variable to subtype clustering was measured through permutation accuracy importance by using random forest. To construct a simple decision tree model that can discriminate patient subtypes, we used the ctree function of the party v1.3-5 package, a conditional interference framework that estimates a regression relationship via binary recursive partitioning. In particular, patient subtypes determined via the aforementioned consensus clustering were used as input for decision tree modeling to identify the key parameters necessary for distinguishing among patient subtypes, and a classifier that could be applied to external cohorts was created.

Follow-up and clinical outcomes

Follow-up with the recruited patients was conducted by March 2019 through face-to-face interviews and/or telephone conversations. The primary end point was death due to cardiovascular diseases, including heart failure-related and sudden deaths. Other clinical outcomes included all-cause death, heart transplant, nonfatal stroke, and progression to New York Heart Association (NYHA) class III/IV.

Whole-exome sequencing (WES) for all the patients

DNA extraction from whole blood and WES was performed on an Illumina platform. The details are described in the Supplementary Methods (Table S1).

Role of HCM-associated genes in subtype classification

To compare the proportion of patients that carried mutations in known HCM-associated genes (Table S2) between subtypes [14], we identified putative causal mutations in these genes, and the recommendations of the American College of Medical Genetics and Genomics (ACMG) were adopted to determine the pathogenicity of each variant [15]. In particular, only rare nonsynonymous or truncating variants (nonsense, frameshift, and splice sites) in HCM-associated genes, with MAF ≤0.1% in the East Asian population from public databases, and labeled deleterious in functional prediction methods, were then subjected to ACMG evaluation. In particular, only truncating variants were retained for evaluating the *TTN* gene. Subsequently, we compared the proportions of patients that carried mutations in each of these genes between different subtypes to determine whether mutations in some of these genes were linked to subtype classification. Moreover, we compared the prognoses of patients that carried mutations in these genes with those that did not carry mutations to evaluate the role of HCM-associated genes in predicting the outcomes of patients. The overall effect of these mutations in predicting subtype classification and survival was assessed via the area under the receiver operating characteristic (ROC) curve (AUC), wherein the status of carrying mutations for each patient was regarded as a predictor.

Novel subtype-specific gene identification based on rare variants

We then estimated the effects of rare variants in a whole-exome scale on subtype classification, not merely in known HCM-associated genes. A total of 136 654 nonsynonymous and truncating variants with MAF < 1% in both our population and East Asian populations from public databases (1000 Genomes Project, Exome Aggregation Consortium, and Genome Aggregation Database) were subjected to subsequent analyses. After gene-based annotation, multiple *in-silico* computational methods were employed for the functional prediction of variants.

To quantify the mutation burden for each gene, we first assessed the pathogenicity of the included variants by using several *in-silico* computational methods, with the best performance in functional prediction, including REVEL, VEST3, MetaLR, and M-CAP [16]. For each variant, the average score calculated among the four algorithms was considered the combined prediction score. The variant-level prediction scores across the entire gene were accumulated as an overall mutation burden for this gene. Accordingly, a score matrix with *gene numbers* × *sample numbers* was generated, where the mutational profile for each sample was represented by 17,033 gene burden scores.

To explore the differences in genetic basis between subtypes, we attempted to model the additive effect of gene mutation burden on HCM subtype propensity. Considering that the number of genes was, relatively, considerably larger than the number of samples, which would lead to an overfitting problem and generate models with poor generalization capability, we introduced the L1-norm to penalize the weight of the model parameters; that is, we aimed to find the best compromise between model complexity and empirical risk and identify a minimum number of feature genes to best explain the observations. Consequently, we adopted a logistic regression model with L1 regularization, which can force coefficient values to be 0, generating a sparse solution to selecting the leading genes of each subtype.

Protein–protein interaction network analysis

Subsequently, we seeded the subtype-specific genes identified above into the STRING Interactome integrated protein–protein interaction (PPI) to build networks through selected connection pairs, with evidence confidence scores over 400. Then, we sought to identify modules that were tightly condensed across the entire network by using the InfoMap algorithm. Members within a module are likely to work collectively to perform biological functions. The aforementioned procedures were implemented using NetworkAnalyst v3.0 [17]. Finally, the biological functions of the observed modules were determined through enrichment analysis and annotated with the Enrichr web server [18].

Individual PPI networks for HCM with reduced LVEF

We retrieved the published proteome expression dataset PXD008934 from ProteomeXchange, which contained the proteomic changes characterized via mass spectrometry in nine human heart tissues with HCM accompanying preserved (53.12% ± 3.75%, n = 4) or reduced ejection fraction (25.00% ± 9.35%, n = 5). Individual networks for samples with reduced LVEF were built following the procedures proposed by Maron *et al.* [19]. A Pearson's correlation matrix was first calculated for each gene pair from all samples with preserved LVEF. Then, each sample with reduced LVEF was added and the correlation matrix was recalculated. Gene pairs with correlations that were significantly changed were mapped to the STRING Interactome. This procedure resulted in a network that represented the dysfunctional or perturbed system of the corresponding sample. We then used a hypergeometric test to determine whether an individualized network was significantly enriched with the identified genes and other genes associated with HCM endophenotypes [19]. The Benjamini–Hochberg procedure was applied for multiple hypothesis tests.

Validation by second independent cohort

External validation is required to ascertain the correlation between genetics and subtypes; this procedure verifies the robustness and generalization of the genetic model for clinical practice. Thus, we enrolled another independent cohort that consisted of 414 patients with HCM from the same hospital (Tongji Hospital, China). WES and genotyping were performed in accordance with the procedures described above.

Statistical analyses

Continuous variables were compared using an unpaired Student's *t*-test, while categorical variables were analyzed using the chi-square or Fisher's exact test. Survival curves were constructed in accordance with the Kaplan–Meier method, and comparisons were performed using the log-rank test. Cox proportional hazard models were used to assess the effects of multiple clinical features on the risk of outcome events. AUC was used to evaluate the performance of the binary classification model. Repeated stratified fivefold cross-validation was used to perform this evaluation. All reported probabilities were two-sided and considered significant at $P < 0.05$.

Results

Consensus clustering identified two HCM subtypes

An unsupervised consensus clustering approach was applied to determine the number of possible subtypes of all the patients with HCM by using echocardiography data. We observed that these patients were clustered into two–six subtypes. The two subtypes (k = 2) were selected for further analysis because of their better performance and stability (Fig. S2). We further measured the differences in clinical features between the two subtypes. As indicated in Table 1, more male subjects were found with subtype 1 than with subtype 2 (83.8% versus 63.7%, respectively; $P < 0.001$). The mean LVPW, LAD, and

Table 1 Characteristics of subtypes in the study population

	Subtype 1 (*n* = 229)	Subtype 2 (*n* = 564)	*P* value
Age of onset (year)	51.14 ± 13.89	51.41 ± 14.41	0.806
Age at enrollment (year)	51.79 ± 14.16	52.99 ± 14.11	0.281
Gender = male (%)	192 (83.8)	359 (63.7)	< 0.001
Smoke (%)	92 (40.2)	193 (34.2)	0.133
Drink (%)	59 (25.8)	136 (24.1)	0.69
CAD (%)	73 (31.9)	123 (21.8)	0.004
Diabetes (%)	46 (20.1)	100 (17.7)	0.5
Systolic blood pressure (mmHg)	129.30 ± 18.37	127.17 ± 15.85	0.103
Diastolic blood pressure (mmHg)	79.36 ± 12.38	75.98 ± 10.88	< 0.001
IVS (mm)	15.61 ± 2.42	17.81 ± 4.74	< 0.001
LVPW (mm)	13.35 ± 2.65	11.64 ± 3.02	< 0.001
Apex (mm)	10.20 ± 1.29	11.24 ± 3.08	< 0.001
LAD (mm)	46.22 ± 7.26	39.72 ± 7.23	< 0.001
LVEDD (mm)	55.82 ± 9.21	44.98 ± 5.03	< 0.001
LVEF (%)	44.67 ± 12.37	64.39 ± 7.87	< 0.001
LVEF < 50% (%)	142 (62.6)	20 (3.6)	< 0.001
Resting LVOTG (mmHg)	33.00 ± 42.58	41.77 ± 54.57	0.41
Valsalva LVOTG (mmHg)	29.67 ± 25.09	52.90 ± 41.89	0.104
E/A	26.41 ± 47.20	2.28 ± 10.57	< 0.001
E/E′	22.45 ± 11.94	17.38 ± 8.47	< 0.001

Values are *n* (%) or mean ± SD.

CAD, coronary artery disease; IVS, interventricular septum; LVPW, left ventricular posterior wall; LAD, left atrial diameter; LVEDD, left ventricular end-diastolic dimension; LVEF, left ventricular ejection fraction; LVOTG, left ventricular outflow tract gradient.

LVEDD were greater in patients with subtype 1 compared with in patients with subtype 2 (LVPW: 13.35 mm versus 11.64 mm, respectively, $P < 0.001$; LAD: 46.22 mm versus 39.72 mm, respectively, $P < 0.001$; LVEDD: 55.82 mm versus 44.98 mm, respectively, $P < 0.001$), while patients with subtype 1 exhibited less thickness of IVS (15.61 mm versus 17.81 mm, $P < 0.001$). An apparent reduction in LVEF was consistently observed in subtype 1 (44.67% versus 64.39%, $P < 0.001$). Both subtypes suffered from LV diastolic dysfunction, while subtype 1 exhibited not only reduced filling function but also damage to LV compliance (E/A ratio: 26.41 versus 2.28, $P < 0.001$), suggesting reliability for two subcluster divisions. Further propensity score matching to adjust for potential bias in baseline characteristics suggested the same findings (Table S3 and Fig. S3).

Supervised decision tree modeling to enhance clinical utility

On the basis of the two identified subtypes, we further tested whether a simplified classifier with a minimal subset of these echocardiographic variables used in consensus clustering could still assign patients to their corresponding subtype. We first used random forest to measure the importance of each echocardiographic variable. The result suggested that the preceding clustering was largely driven by LVEF, LVEDD, E/A, LAD, LVPW, and IVS (Fig. 1C). We then applied decision tree modeling by using the subtypes from the preceding clustering as input to create a classifier that comprised the six variables above (Fig. 1D). The result revealed that the HCM patients could still be stratified into the two subtypes with an AUC of 0.93 (95% confidence interval 0.91–0.95).

Association of subtypes with clinical outcome

The above findings revealed two distinct subtypes of HCM on the basis of multiple methods. Subsequently, we verified whether the two subtypes were associated with different prognoses. Among the 775 (97.7%) patients included in the final evaluation, with a mean follow-up time of 32.78 ± 27.58 months, we observed higher all-cause mortality in subtype 1 compared with in subtype 2 (20.2% versus 11.4%, $P = 0.002$). The 18 patients who were lost to follow-up were excluded in the survival analysis. Further survival analysis (Fig. 2) showed that patients with subtype 1 had a higher likelihood of experiencing primary end point events (cardiac mortality: HR 2.68, $P < 0.001$; cardiac death and heart transplant rate: HR 2.83, $P < 0.001$; all-cause mortality: HR 2.11, $P < 0.001$) and developing moderate or severe congestive symptoms (NYHA class III/IV) (HR 2.69, $P < 0.001$) compared with patients with subtype 2. In accordance with an earlier study [20], age, female sex, NYHA class III/IV symptoms, and history of atrial fibrillation were predictors of cardiac mortality (Table S4). Subtype 1 remained independently associated with a higher risk of cardiovascular death (HR 2.24, $P = 0.0015$) compared with subtype 2 after using multivariable modeling inclusive of all significant univariate predictors (Table S5 and Fig. S4). When adjusted for LVEF, subtype 1 remained an independent risk factor for NYHA class III/IV (HR 1.48, $P = 0.047$) (Table S6). In general, subtype 1 patients were associated with poor overall survival probability compared with subtype 2 patients.

Effects of HCM-associated genes on subtyping and disease prognosis

The distinct clinical characteristics of the two subtypes have prompted us to explore the underlying genetic determinants. We first focused on the evaluation of HCM-associated genes (Table S7). For the majority of HCM-associated genes, the proportion of carriers was not different between the subtypes, except for *MYBPC3* and *MYH7*, whose carriers were significantly enriched in subtype 2 relative to that in subtype 1 (Figs. S5 and S6). We further compared the risk of experiencing cardiac death, and no significant difference was observed between carriers and noncarriers for most of the genes associated with HCM (Fig. S7). Consistently, the overall effect assessment suggested that mutations in these genes could hardly discriminate one subtype from the other (AUC = 0.54) or predict survival at the end of the follow-up period (AUC = 0.62) (Fig. S8).

Machine learning modeling to identify novel genetic determinants

The weak contribution of known HCM-associated genes to the subtypes and outcomes observed above has prompted speculation that other novel disease-modifying genes may be present in HCM. Given that rare variants have a relatively larger effect size but cannot be effectively captured by single-variant analyses [21–23], we constructed machine learning models based on the accumulated mutation pathogenicity of rare variants at exome-wide gene level to distinguish the subtypes. Figure illustrates the process of searching for an optimal C value (the inverse of regularization strength), which was determined by 1000 times (random shuffle) stratified fivefold cross-validation. The optimal C value was set to 0.033 for a minimum average log loss. At the given C value, 51 genes among the whole 17 033 gene set were assigned with nonzero weights, with 46 genes exhibiting an increased mutation burden in subtype 1 relative to subtype 2 (Table S8). Subsequently, we constructed

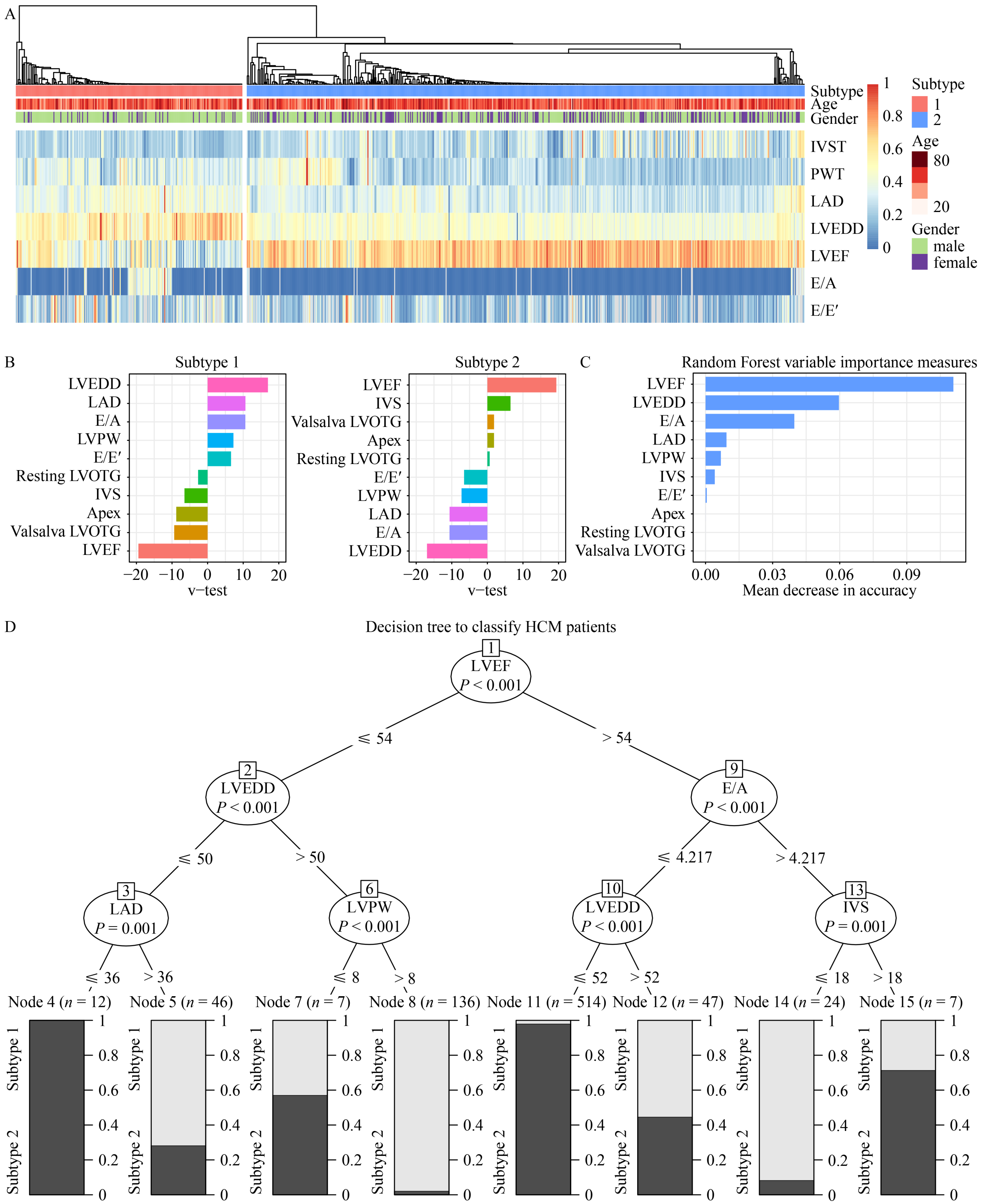

Fig. 1 Consensus clustering identified two subtypes that correlated with various clinical features. (A) Hierarchical subtype heat map shows the clinical characteristics and echocardiography features of the two subtypes. (B) Characteristic plots of the two subtypes, including their most representative echo variables. A positive value indicates overrepresentation of this variable in the applicable subtype. A negative value indicates underrepresentation of the corresponding variable. (C) Variable importance for clustering measured by random forests. (D) Supervised decision tree modeling provided availability in clinical practice.

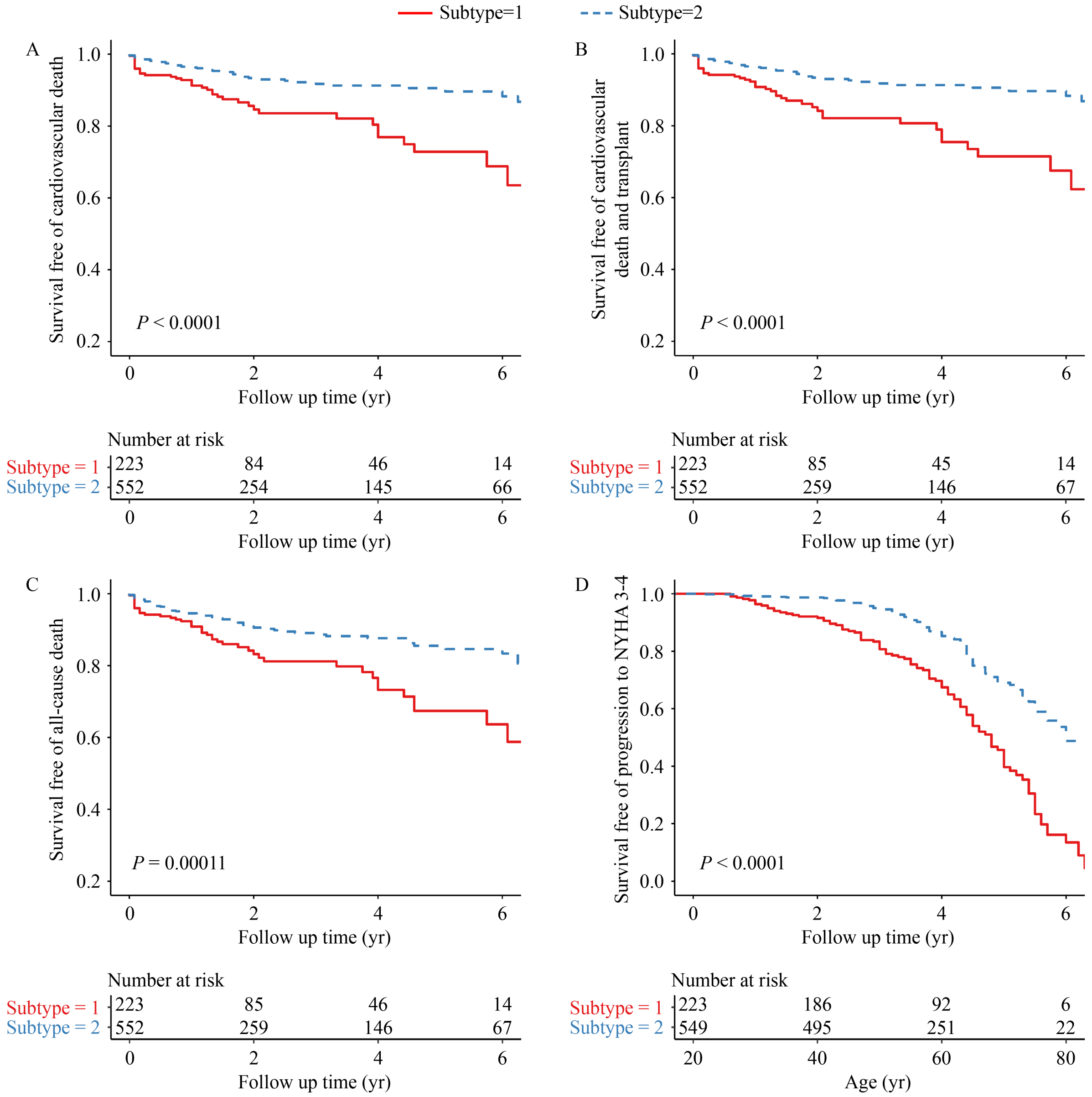

Fig. 2 Event-free survival stratified by subtypes as determined by the consensus clustering. Kaplan–Meier curves for (A) cardiovascular death, (B) the combined outcome of cardiovascular death and heart transplant, (C) all-cause death and (D) lifelong likelihood of progression to NYHA class III/IV, respectively. Age was used as the time scale, and events occurring before and during the follow-up were included. The probability values were calculated with the log-rank test.

models with 46 genes to verify whether they could accurately predict subtype classification. As shown in Fig. 3B, the machine learning model based on the 46 genes presented superior predictive power with an average AUC of approximately 0.81. Hence, these genes are probably the most distinguishing features of the subtypes.

Moreover, correlation analyses suggested a positive linear link between LVEDD and probability for subtype 1 predicted by the 46-gene model ($R = 0.25$, $P = 5.8\text{e}{-13}$) and a negative link between LVEF and probability for subtype 1 ($R = -0.34$, $P = 2.2\text{e}{-16}$) (Fig. S10). A similar trend was observed in the survival analysis, wherein the likelihood of experiencing cardiac death and progression to NYHA class III/IV increased following a rise in probability for subtype 1 (Fig. S11). Combined, these results indicate that the identified genes exerted a stronger effect on the severity and prognosis of HCM relative to known HCM-associated genes.

Network analyses to unravel underlying pathobiology

To further explore the pathobiology that accounted for subtype, we subsequently mapped the 46 machine-

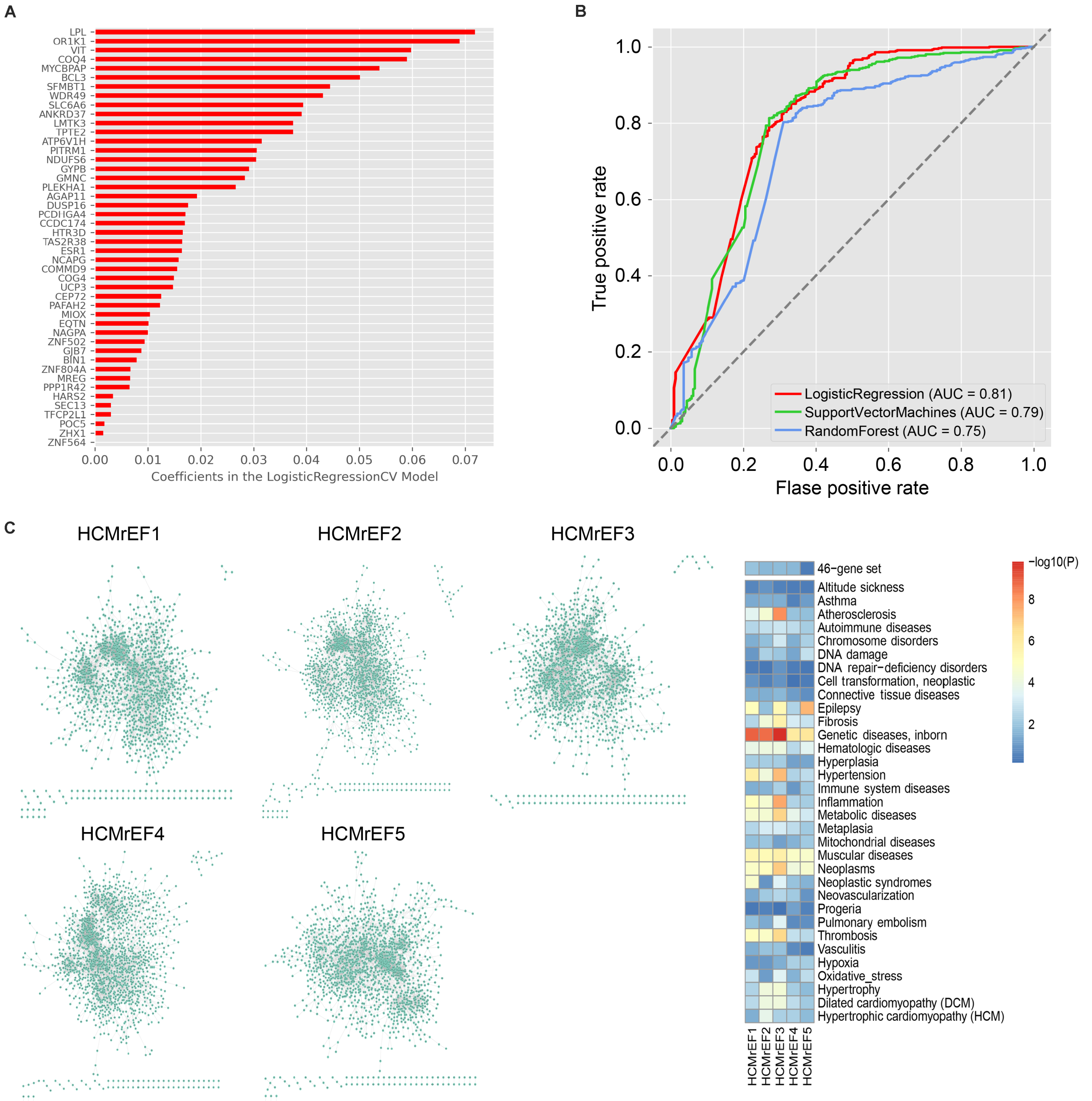

Fig. 3 Machine learning model construction. (A) Machine-identified genes with increased mutation burden in subtype 1. (B) ROC curves for the models based on the 46 feature genes with different classifiers. AUC was determined via stratified fivefold cross-validation. (C) Individual networks for HCM patients with reduced LVEF and enrichment analysis of the 46-gene set and endophenotype for each patient network. The rows correspond to the gene ontology (GO) classifications for genes, and the columns denote samples.

identified genes onto human PPI networks to determine associated biological pathways. Subsequent community detection identified 36 modules that were tightly condensed internally. Expectedly, the GO term annotation for these modules suggested links with the cardiovascular system to a certain extent (Fig. S12). Given the dominant role of LVEF in subtyping, published proteomic expression profiles from HCM patients with reduced LVEF compared with those with preserved LVEF were used to generate individual PPI networks (Table S9). Enrichment analysis showed that the 46-gene set was significantly enriched across the patient networks, except for sample HCMrEF5 (Fig. 3C, Table S10). In addition, we determined that some of the individual networks were also enriched for the HCM endophenotypes identified in a previous study (Fig. 3C) [19]. Combined, these results provided insights into the pathobiological complexity of HCM subtypes at the network medicine level.

Second cohort validation

To validate the correlation of identified genes with phenotypic variability, we enrolled for another independent cohort that comprised 414 patients with HCM recruited from Tongji Hospital (Wuhan, China) and diagnosed with the same criterion, and performed WES. The same mutation burden weighting that used rare variants for the 46 genes was followed. The subtype status of these patients was then predicted using the aforementioned genetic model, which was fitted by the first cohort based on the 46 genes. To avoid confusion, the predicted subtype for each individual was labeled as “group” rather than “subtype”. As presented in Table 2, 101 patients were predicted for Group 1, while the rest were labeled for Group 2. Significant differences still existed between the two groups in terms of IVS, LVEDD, and LVEF. Compared with the characteristics summarized in the first cohort, Group 1 presented increased LVEDD (53.47 mm versus 49.57 mm, $P < 0.001$) and impaired LVEF (53.00% versus 57.92%, $P = 0.002$), while Group 2 was characterized by more severe IVS (15.73 mm versus 17.09 mm, $P = 0.002$). Moreover, we ranked samples into quartiles in accordance with their predicted possibilities for subtype 1 and observed the same progression trends across quartiles (Fig. S13).

To test the clinical utility of the genetic model, we applied the previous decision tree derived from echo-based clustering (Fig. 1D) to the second cohort and determined the corresponding clinical subtypes, namely, true labels. The predictive power of the genetic model in the second cohort is depicted in Fig. S14, with an AUC of 0.64. Accounting for the effects of traditional risk factors on cardiovascular diseases, we collected 12 other clinical variables of these patients to construct an integrated model. These clinical variables were as follows: sex, age, smoking, alcohol intake, systolic blood pressure, diastolic blood pressure, serum triglycerides, total cholesterol, high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, coronary atherosclerosis, and diabetes. By integrating these factors into our genetic model, we achieved a better interpretation of HCM subtyping with significantly increased AUCs of 0.84 in the first cohort and 0.70 in the second cohort. Overall, these results supported the assumption that the 46 genes, although with limited coverage in all the patients, were associated with HCM phenotypic variability. The integration of genetic and nongenetic factors is capable of recognizing patients who tend to suffer from adverse remodeling, albeit only partially.

Table 2 Characteristics of second population subtyping by the genetic model

	Group 1 (*n* = 101)	Group 2 (*n* = 313)	*P* value
Age at enrollment (year)	53.42 ± 12.58	53.47 ± 11.72	0.97
Gender = male (%)	85 (84.2)	233 (74.4)	0.044
Smoke (%)	51 (51.0)	135 (43.8)	0.211
Drink (%)	24 (24.0)	85 (27.6)	0.48
CAD (%)	48 (48.0)	130 (42.2)	0.31
Diabetes (%)	21 (21.0)	72 (23.3)	0.633
IVS (mm)	15.73 ± 3.12	17.09 ± 3.93	0.002
LVPW (mm)	12.10 ± 2.60	12.41 ± 2.55	0.291
Apex (mm)	10.86 ± 2.76	10.97 ± 2.71	0.725
LAD (mm)	43.36 ± 7.47	43.01 ± 7.62	0.685
LVEDD (mm)	53.47 ± 9.04	49.57 ± 7.74	< 0.001
LVEF (%)	53.00 ± 15.46	57.92 ± 12.80	0.002
Resting LVOTG (mmHg)	26.41 ± 26.65	43.99 ± 58.51	0.124
Valsalva LVOTG (mmHg)	59.00 ± 38.17	61.49 ± 60.49	0.911
E/A	1.14 ± 0.81	1.12 ± 0.90	0.914
E/E′	18.81 ± 10.52	17.68 ± 8.01	0.273

Values are *n* (%) or mean ± SD.
IVS, interventricular septum; LVPW, left ventricular posterior wall; LAD, left atrial diameter; LVEDD, left ventricular end-diastolic dimension; LVEF, left ventricular ejection fraction; LVOTG, left ventricular outflow tract gradient.

Discussion

Overall, the comprehensive clustering analysis of echocardiography features from 793 HCM cases uncovered two major clinical subtypes, which exhibited distinct manifestation and genetic basis. Patients with subtype 2 presented a form of asymmetric septal hypertrophy and were associated with a stable course. By contrast, posterior free wall involvement, LV systolic dysfunction, and unfavorable outcomes were more common in subtype 1. The subsequent machine learning model construction identified 46 most distinguishing genes with increased mutation burden in subtype 1. Network analysis revealed functional modules and biological pathways involved in subtypes, along with the enrichment of the identified genes in individual PPI networks for HCM patients with reduced LVEF. External validation in a second cohort of 414 cases provided evidence in favor of the correlation between genetics and subtypes. We intended to draw an overall picture of the genetic basis accounting for subtypes at the levels of variant, gene, and network, without subjective choices in any steps.

Previous studies have noted that the lifelong process of LV remodeling and progressive dysfunction occurred in some HCM patients [24]. The results of large-scale cohort studies implied the inadequacy that not all, but only a small proportion, of patients developed to this end stage [7]. Otherwise, the substantial heterogeneity that drives for such different progression is less clear. Therefore, we discussed the possibility that HCM subtypes exist naturally and their differences in a genetic context, intending to view this disease as inclusive of its clinical, morphological, and molecular diversities. In contrast with previous studies that were based on natural history observation and subjective division [25], we conducted unsupervised clustering in a large-scale HCM cohort by taking advantage of machine learning approaches, wherein relevant structural and functional data recorded via echocardiography were used as input to the clustering algorithm. Therefore, patients were automatically clustered into several groups in accordance with their similarities in echocardiography features.

Considering the limitations in viewing HCM through the narrow prism of a single sarcomere gene mutation, we comprehensively inspected all genes based on WES data from different hierarchies. In contrast with classical burden tests and SKAT [26], which are based on allele frequency or variance component score tests, we weighted rare variants with pathogenicity instead of regression coefficients and aggregated them into a combined mutation burden for each gene. Given the limited power for detecting genetic susceptibility with a relatively small sample size, we adopted a feature selection algorithm based on a penalized linear classification model that measured the contribution of genes to subtype status. The initial objective could be substantially interpreted to explain and predict the morphological abnormalities of HCM with personal genetics. The minimal subset of 46 genes identified by the L1-penalized regression model achieved the most accurate prediction of HCM subtypes. HCM has been widely regarded as a monogenic disease, in which causal mutation in sarcomere genes is believed to be the prerequisite and a major determinant of the phenotype [27]. In contrast with this hypothesis, a poor predictive performance was observed in the model based on HCM-associated genes. These observations further support the new perspective that HCM clinical phenotype may be defined by the genetic context rather than solely by a single genetic event [28]. Similarly, significant increases in the predictive power for both cohorts after integrating the genetic model with clinical risk factors reflected nongenetic contributions in disease processes. However, a considerable proportion of patients with adverse remodeling in the second cohort could not be captured by the 46-gene model. Such limited coverage emphasized the variability in molecular mechanisms among patients with the same HCM diagnosis. Combined, these points underscore the need to expand the spectrum of determinants and modifying factors in HCM remodeling.

Apart from applying echo-based decision trees to match patients to their corresponding subtypes, personal genetics seems more applicable in offering an early evaluation of HCM progression risk. Genetic testing is recommended for patients fulfilling the diagnostic criteria for HCM due to an increased understanding of the genetic basis of HCM and the rapidly evolving high-throughput sequencing technologies. Our results suggest that patients may benefit from genetic testing in other aspects, not only in the diagnosis of HCM. A widening range for genetic testing should be recommended, because sequencing and analysis should not be limited to HCM-associated genes. Evaluating the risk for different progression based on personal genetics and traditional risk factors is possible and may provide valuable advice for early intervention and disease management.

In the absence of experimental evidence, 46 genes were agnostically and automatically selected and considered subtype-related genes. The machine learning algorithm only considered genes whose increased mutational burden in subtype 1 would contribute to prediction accuracy, which might carry a risk of false positives. With the aim of testing whether these selected genes are involved in the pathogenesis of HCM and determine their functional context, we mapped them onto a human PPI network and identified tightly clustered topological modules linked with subtype 1. Expectedly, the subsequent GO and Kyoto Encyclopedia of Genes and Genomes enrichment analyses for these modules indicated that some modules were directly involved in the cardiovascular system, such as cholesterol metabolism, mitochondrial oxidative metabolism, and sarcomere organization (Fig. S12). We also noted some novel or less reported pathways, such as the mTOR signaling pathway, PI3K-Akt signaling pathway, and aminoacyl-tRNA biosynthesis, which may promote new perspectives for HCM [29,30]. Previous studies have proposed that individualized PPI networks can provide critical insight into determining patient-specific and clinically relevant HCM pathophenotypic characteristics [19,31]. Thus, we utilized the information provided by individual networks of LVEF-reduced patients to check the role of these feature genes and relevant endophenotypes that were unique to specific patients. Our results revealed that the 46-gene set was enriched across the individual networks of HCM patients with reduced LVEF and provided further support for the involvement of these genes in HCM. These results also suggested that mutation signatures in these genes were implicated in phenotypic heterogeneity. Further work is required to establish the relationship between these modules and HCM subtypes and to elucidate their exact mechanisms.

Study limitations

Our study was based on the echocardiographic features of 793 patients with HCM from a large-scale cohort. Compared with magnetic resonance imaging, echocardiography may be limited in providing detailed information for patients with poor acoustic windows or in detecting LV apical and anterolateral hypertrophy. In addition, our patients were recruited from a single center and an age span existed in the cohort. A higher rate of progression to heart failure and mortality was observed in this study compared with previously published cohorts, which might be explained by the inadequate attention given to HCM in China. Patients with HCM only visit a hospital when evident symptoms emerge. Meanwhile, the links of genes and pathways obtained from machine learning modeling with HCM subtypes should be further confirmed by animal and cytological experiments. In addition, limited proteins were available for proteome analysis, resulting in incomplete individual network construction.

Conclusions

This study was designed to explore the potential subtypes of HCM in a large-scale cohort. On the basis of echocardiography features, we propose a new classification scheme based on a distinct genetic context. Personal whole exome-based machine learning methods have been used to identify HCM subtype-associated genes and subtype prediction model construction. These findings may contribute to our understanding of the correlations among phenotypes, genotypes, and prognoses in HCM.

Acknowledgements

We gratefully acknowledge all the participants and funding sources. This study was funded by the National Key R&D Program of China (No. 2017YFC0909400), the National Natural Science Foundation of China (Nos. 91439203, 91839302, and 81700413), Shanghai Municipal Science and Technology Major Project (No. 2017SHZDZX01), and the fundamental Research Funds for the Central Universities, HUST (No. 2016JCTD117).

Compliance with ethics guidelines

Jiaqi Dai, Tao Wang, Ke Xu, Yang Sun, Zongzhe Li, Peng Chen, Hong Wang, Dongyang Wu, Yanghui Chen, Lei Xiao, Hao Liu, Haoran Wei, Rui Li, Liyuan Peng, Ting Yu, Yan Wang, Zhongsheng Sun, and Dao Wen Wang declare no conflict of interest. All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the *Helsinki Declaration* of 1975, as revised in 2000. Informed consent was obtained from all the patients for being included in the study.

Electronic Supplementary Material Supplementary material is available in the online version of this article at https://doi.org/10.1007/s11684-023-0982-1 and is accessible for authorized users.

References

1. Maron BJ, Mathenge R, Casey SA, Poliac LC, Longe TF. Clinical profile of hypertrophic cardiomyopathy identified *de novo* in rural communities. J Am Coll Cardiol 1999; 33(6): 1590–1595
2. Zou Y, Song L, Wang Z, Ma A, Liu T, Gu H, Lu S, Wu P, Zhang dagger Y, Shen dagger L, Cai Y, Zhen double dagger Y, Liu Y, Hui R. Prevalence of idiopathic hypertrophic cardiomyopathy in China: a population-based echocardiographic analysis of 8080 adults. Am J Med 2004; 116(1): 14–18
3. Eriksson MJ, Sonnenberg B, Woo A, Rakowski P, Parker TG, Wigle ED, Rakowski H. Long-term outcome in patients with apical hypertrophic cardiomyopathy. J Am Coll Cardiol 2002; 39(4): 638–645
4. Maron BJ, Rowin EJ, Casey SA, Link MS, Lesser JR, Chan RH, Garberich RF, Udelson JE, Maron MS. Hypertrophic cardiomyopathy in adulthood associated with low cardiovascular mortality with contemporary management strategies. J Am Coll Cardiol 2015; 65(18): 1915–1928
5. Olivotto I, Cecchi F, Poggesi C, Yacoub MH. Patterns of disease progression in hypertrophic cardiomyopathy: an individualized approach to clinical staging. Circ Heart Fail 2012; 5(4): 535–546
6. Harris KM, Spirito P, Maron MS, Zenovich AG, Formisano F, Lesser JR, Mackey-Bojack S, Manning WJ, Udelson JE, Maron BJ. Prevalence, clinical profile, and significance of left ventricular remodeling in the end-stage phase of hypertrophic cardiomyopathy. Circulation 2006; 114(3): 216–225
7. Melacini P, Basso C, Angelini A, Calore C, Bobbo F, Tokajuk B, Bellini N, Smaniotto G, Zucchetto M, Iliceto S, Thiene G, Maron BJ. Clinicopathological profiles of progressive heart failure in hypertrophic cardiomyopathy. Eur Heart J 2010; 31(17): 2111–2123
8. Watkins H, McKenna WJ, Thierfelder L, Suk HJ, Anan R, O' Donoghue A, Spirito P, Matsumori A, Moravec CS, Seidman JG, Seidman CE. Mutations in the genes for cardiac troponin T and α-tropomyosin in hypertrophic cardiomyopathy. N Engl J Med 1995; 332(16): 1058–1065
9. Coppini R, Ho CY, Ashley E, Day S, Ferrantini C, Girolami F, Tomberli B, Bardi S, Torricelli F, Cecchi F, Mugelli A, Poggesi C, Tardiff J, Olivotto I. Clinical phenotype and outcome of hypertrophic cardiomyopathy associated with thin-filament gene mutations. J Am Coll Cardiol 2014; 64(24): 2589–2600
10. Harper AR, Goel A, Grace C, Thomson KL, Petersen SE, Xu X, Waring A, Ormondroyd E, Kramer CM, Ho CY, Neubauer S; HCMR Investigators; Tadros R, Ware JS, Bezzina CR, Farrall M, Watkins H. Common genetic variants and modifiable risk factors underpin hypertrophic cardiomyopathy susceptibility and expressivity. Nat Genet 2021; 53(2): 135–142
11. Ommen SR, Mital S, Burke MA, Day SM, Deswal A, Elliott P, Evanovich LL, Hung J, Joglar JA, Kantor P, Kimmelstiel C, Kittleson M, Link MS, Maron MS, Martinez MW, Miyake CY, Schaff HV, Semsarian C, Sorajja P. 2020 AHA/ACC guideline for the diagnosis and treatment of patients with hypertrophic cardiomyopathy: a report of the American College of Cardiology/American Heart Association Joint Committee on clinical practice guidelines. Circulation 2020; 142(25): e558–e631
12. Dai J, Li Z, Huang W, Chen P, Sun Y, Wang H, Wu D, Chen Y, Li C, Xiao L, Liu H, Wei H, Li R, Duan Q, Peng L, Song X, Yu T, Wang Y, Wang DW. Rbm20 is a candidate gene for hypertrophic cardiomyopathy. Can J Cardiol 2021; 37(11): 1751–1759
13. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics 2010; 26(12): 1572–1573

14. Ingles J, Goldstein J, Thaxton C, Caleshu C, Corty EW, Crowley SB, Dougherty K, Harrison SM, McGlaughon J, Milko LV, Morales A, Seifert BA, Strande N, Thomson K, Peter van Tintelen J, Wallace K, Walsh R, Wells Q, Whiffin N, Witkowski L, Semsarian C, Ware JS, Hershberger RE, Funke B. Evaluating the clinical validity of hypertrophic cardiomyopathy genes. Circ Genom Precis Med 2019; 12(2): e002460
15. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL; ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 2015; 17(5): 405–424
16. Li J, Zhao T, Zhang Y, Zhang K, Shi L, Chen Y, Wang X, Sun Z. Performance evaluation of pathogenicity-computation methods for missense variants. Nucleic Acids Res 2018; 46(15): 7793–7804
17. Zhou G, Soufan O, Ewald J, Hancock REW, Basu N, Xia J. NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. Nucleic Acids Res 2019; 47(W1): W234–W241
18. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, McDermott MG, Monteiro CD, Gundersen GW, Ma'ayan A. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 2016; 44(W1): W90–W97
19. Maron BA, Wang RS, Shevtsov S, Drakos SG, Arons E, Wever-Pinzon O, Huggins GS, Samokhin AO, Oldham WM, Aguib Y, Yacoub MH, Rowin EJ, Maron BJ, Maron MS, Loscalzo J. Individualized interactomes for network-based precision medicine in hypertrophic cardiomyopathy with implications for other clinical pathophenotypes. Nat Commun 2021; 12(1): 873
20. Geske JB, Ong KC, Siontis KC, Hebl VB, Ackerman MJ, Hodge DO, Miller VM, Nishimura RA, Oh JK, Schaff HV, Gersh BJ, Ommen SR. Women with hypertrophic cardiomyopathy have worse survival. Eur Heart J 2017; 38(46): 3434–3440
21. Rhee EP, Yang Q, Yu B, Liu X, Cheng S, Deik A, Pierce KA, Bullock K, Ho JE, Levy D, Florez JC, Kathiresan S, Larson MG, Vasan RS, Clish CB, Wang TJ, Boerwinkle E, O'Donnell CJ, Gerszten RE. An exome array study of the plasma metabolome. Nat Commun 2016; 7(1): 12360
22. Wessel J, Chu AY, Willems SM, Wang S, Yaghootkar H, Brody JA, Dauriz M, Hivert MF, Raghavan S, Lipovich L, Hidalgo B, Fox K, Huffman JE, An P, Lu Y, Rasmussen-Torvik LJ, Grarup N, Ehm MG, Li L, Baldridge AS, Stančáková A, Abrol R, Besse C, Boland A, Bork-Jensen J, Fornage M, Freitag DF, Garcia ME, Guo X, Hara K, Isaacs A, Jakobsdottir J, Lange LA, Layton JC, Li M, Hua Zhao J, Meidtner K, Morrison AC, Nalls MA, Peters MJ, Sabater-Lleal M, Schurmann C, Silveira A, Smith AV, Southam L, Stoiber MH, Strawbridge RJ, Taylor KD, Varga TV, Allin KH, Amin N, Aponte JL, Aung T, Barbieri C, Bihlmeyer NA, Boehnke M, Bombieri C, Bowden DW, Burns SM, Chen Y, Chen YD, Cheng CY, Correa A, Czajkowski J, Dehghan A, Ehret GB, Eiriksdottir G, Escher SA, Farmaki AE, Frånberg M, Gambaro G, Giulianini F, Goddard WA 3rd, Goel A, Gottesman O, Grove ML, Gustafsson S, Hai Y, Hallmans G, Heo J, Hoffmann P, Ikram MK, Jensen RA, Jørgensen ME, Jørgensen T, Karaleftheri M, Khor CC, Kirkpatrick A, Kraja AT, Kuusisto J, Lange EM, Lee IT, Lee WJ, Leong A, Liao J, Liu C, Liu Y, Lindgren CM, Linneberg A, Malerba G, Mamakou V, Marouli E, Maruthur NM, Matchan A, McKean-Cowdin R, McLeod O, Metcalf GA, Mohlke KL, Muzny DM, Ntalla I, Palmer ND, Pasko D, Peter A, Rayner NW, Renström F, Rice K, Sala CF, Sennblad B, Serafetinidis I, Smith JA, Soranzo N, Speliotes EK, Stahl EA, Stirrups K, Tentolouris N, Thanopoulou A, Torres M, Traglia M, Tsafantakis E, Javad S, Yanek LR, Zengini E, Becker DM, Bis JC, Brown JB, Cupples LA, Hansen T, Ingelsson E, Karter AJ, Lorenzo C, Mathias RA, Norris JM, Peloso GM, Sheu WH, Toniolo D, Vaidya D, Varma R, Wagenknecht LE, Boeing H, Bottinger EP, Dedoussis G, Deloukas P, Ferrannini E, Franco OH, Franks PW, Gibbs RA, Gudnason V, Hamsten A, Harris TB, Hattersley AT, Hayward C, Hofman A, Jansson JH, Langenberg C, Launer LJ, Levy D, Oostra BA, O'Donnell CJ, O'Rahilly S, Padmanabhan S, Pankow JS, Polasek O, Province MA, Rich SS, Ridker PM, Rudan I, Schulze MB, Smith BH, Uitterlinden AG, Walker M, Watkins H, Wong TY, Zeggini E; EPIC-InterAct Consortium; Laakso M, Borecki IB, Chasman DI, Pedersen O, Psaty BM, Tai ES, van Duijn CM, Wareham NJ, Waterworth DM, Boerwinkle E, Kao WH, Florez JC, Loos RJ, Wilson JG, Frayling TM, Siscovick DS, Dupuis J, Rotter JI, Meigs JB, Scott RA, Goodarzi MO. Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. Nat Commun 2015; 6(1): 5897
23. Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, Daly MJ, Neale BM, Sunyaev SR, Lander ES. Searching for missing heritability: designing rare variant association studies. Proc Natl Acad Sci USA 2014; 111(4): E455–E464
24. Maron BJ, Spirito P. Implications of left ventricular remodeling in hypertrophic cardiomyopathy. Am J Cardiol 1998; 81(11): 1339–1344
25. Klues HG, Schiffers A, Maron BJ. Phenotypic spectrum and patterns of left ventricular hypertrophy in hypertrophic cardiomyopathy: morphologic observations and significance as assessed by two-dimensional echocardiography in 600 patients. J Am Coll Cardiol 1995; 26(7): 1699–1708
26. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 2011; 89(1): 82–93
27. Marian AJ, Braunwald E. Hypertrophic cardiomyopathy: genetics, pathogenesis, clinical manifestations, diagnosis, and therapy. Circ Res 2017; 121(7): 749–770
28. Maron BJ, Maron MS, Maron BA, Loscalzo J. Moving beyond the sarcomere to explain heterogeneity in hypertrophic cardiomyopathy: Jacc review topic of the week. J Am Coll Cardiol 2019; 73(15): 1978–1986
29. Friederich MW, Timal S, Powell CA, Dallabona C, Kurolap A, Palacios-Zambrano S, Bratkovic D, Derks TGJ, Bick D, Bouman K, Chatfield KC, Damouny-Naoum N, Dishop MK, Falik-Zaccai TC, Fares F, Fedida A, Ferrero I, Gallagher RC, Garesse R, Gilberti M, González C, Gowan K, Habib C, Halligan RK, Kalfon L, Knight K, Lefeber D, Mamblona L, Mandel H, Mory A, Ottoson J, Paperna T, Pruijn GJM, Rebelo-Guiomar PF, Saada A, Sainz B Jr, Salvemini H, Schoots MH, Smeitink JA, Szukszto MJ, Ter Horst HJ, van den Brandt F, van Spronsen FJ, Veltman JA, Wartchow E, Wintjes LT, Zohar Y, Fernández-Moreno MA, Baris HN, Donnini C, Minczuk M, Rodenburg RJ, Van Hove JLK. Pathogenic variants in glutamyl-tRNA^{Gln} amidotransferase

subunits cause a lethal mitochondrial cardiomyopathy disorder. Nat Commun 2018; 9(1): 4065

30. Marin TM, Keith K, Davies B, Conner DA, Guha P, Kalaitzidis D, Wu X, Lauriol J, Wang B, Bauer M, Bronson R, Franchini KG, Neel BG, Kontaridis MI. Rapamycin reverses hypertrophic cardiomyopathy in a mouse model of LEOPARD syndrome-associated PTPN11 mutation. J Clin Invest 2011; 121(3): 1026–1043
31. Lee LY, Loscalzo J. Network medicine in pathobiology. Am J Pathol 2019; 189(7): 1311–1326