

Database, 2020, 1–12 doi: 10.1093/database/baaa048 Original article



Original article

CHDGKB: a knowledgebase for systematic understanding of genetic variations associated with non-syndromic congenital heart disease

Lan Yang^{1,2,†}, Yang Yang^{3,†}, Xingyun Liu^{1,4,†}, Yongquan Chen³, Yalan Chen¹, Yuxin Lin¹, Yan Sun⁴ and Bairong Shen^{4,*}

¹Center for Systems Biology, Soochow University, Suzhou 215006, China, ²Center of Prenatal Diagnosis, Wuxi Maternal and Child Health Hospital Affiliated to Nanjing Medical University, Wuxi 214002, China, ³School of Computer Science and Technology, Soochow University, Suzhou 215006, China and ⁴Institutes for Systems Genetics, West China Hospital, Sichuan University, Chengdu 610041, China

*Corresponding author: Tel/Fax: +86-28-85164199; E-mail: bairong.shen@scu.edu.cn [†]These authors contributed equally to this work.

Citation details: Yang,L., Yang,Y., Liu,X. *et al.* CHDGKB: a knowledgebase for systematic understanding of genetic variations associated with non-syndromic congenital heart disease. *Database* (2020) Vol. 2020: article ID baaa048; doi:10.1093/database/baaa048

Received 7 March 2020; Revised 18 May 2020; Accepted 27 May 2020

Abstract

Congenital heart disease (CHD) is one of the most common birth defects, with complex genetic and environmental etiologies. The reports of genetic variation associated with CHD have increased dramatically in recent years due to the revolutionary development of molecular technology. However, CHD is a heterogeneous disease, and its genetic origins remain inconclusive in most patients. Here we present a database of genetic variations for non-syndromic CHD (NS-CHD). By manually literature extraction and analyses, 5345 NS-CHD-associated genetic variations were collected, curated and stored in the public online database. The objective of our database is to provide the most comprehensive updates on NS-CHD genetic research and to aid systematic analyses of pathogenesis of NS-CHD in molecular level and the correlation between NS-CHD genotypes and phenotypes.

Database URL: http://www.sysbio.org.cn/CHDGKB/

Introduction

Congenital heart disease (CHD) is one of the most prevalent birth defects, ranging from 6.9 per 1000 births in Europe to 9.3 per 1000 births in Asia (1, 2). Although the CHD survival rate has improved due to the recent rapid development of surgical repair, not all cardiac defects can be diagnosed by a routine prenatal ultrasound. The mortality rate among patients with severe CHD remains high. While environmental factors have been shown to greatly contribute to the onset and progression of CHD (3), the huge impact of genetic defects on the pathogenesis of CHD during cardiac development has also been well

 $\ensuremath{\mathbb{C}}$ The Author(s) 2020. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Page 1 of 12

documented (4-8). When exposed to the same environment, different individuals exhibited variable susceptibilities to CHD, suggesting the effect of hereditary factors. The main known genetic factors leading to CHD include focal mutation and chromosomal abnormalities (9). Nevertheless, the genetic origin of CHD, as well as the correlation between its genotypes and phenotypes, remains unclear. As studies in recent decades have mainly focused on the effect of cardiac surgery on CHD patients (10) or the incidence of CHD subtypes with genetic variations (11) at a systematic analysis level, there is no available database associated with genetic etiology of CHD. The database presented in our current study provides investigators and the public with a systematic and comprehensive understanding of the genetic etiologies of non-syndromic CHD (NS-CHD) to improve the interpretation of the NS-CHD variants.

Methods

Data collection

Following discussion with database construction experts, molecular biologists, clinicians, medical researchers, biostatisticians and bioinformaticians, we collected all the data for our NS-CHD genetic knowledge database (CHDGKB) from PubMed, by manual text mining. The literature searches were performed on publications prior to 31 July 2019 with the following keywords: ((congenital heart disease [Title/Abstract]) AND (genetic [Title/Abstract] OR gene*[Title/Abstract])). As a result, 284 out of 2054 publications from 1998 to 2019 were selected for our NS-CHD database. The data collection work flow is depicted in Figure 1.

Inclusion and exclusion criteria

For inclusion in the CHDGKB, a study had to meet the following criteria: (i) all the patients incorporated into the CHDGKB presented with clinical features of CHD and had echocardiographic evidence or surgical records and (ii) all data in our CHDGKB was collected based on the inclusion criteria requiring that all the included studies were performed in accordance with their approved institutional guidelines and with the informed consent signed by all human study subjects. We hereby confirm ethical statements and approvals for all study data that we collected.

Genetic association data were excluded when (i) patients exhibits clinical features of CHD without echocardiographic evidence or any other examination results of heart structure; (ii) patients have any confirmed chromosomal abnormalities or syndrome-associated cardiovascular abnormalities; (iii) patients have other known complication, such as Noonan, DiGeorge, Holt-Oram, Marfan, Chat



Figure 1. The schematic flow of the NS-CHD data collection and construction. After a series of standard selections, there were 284 articles selected from PubMed and incorporated into our database construction.

and other syndromes; and (iv) patients have established obvious environmental risk factors for CHD, such as maternal illness, drug use during the first trimester of pregnancy, parental smoking or chronic exposure to toxicants and ionizing radiation.

Database construction

The CHDGKB web interface was constructed with MySQL (5.6.19) server, Apache (2.0.61), PHP (5.2.5), HTML (5) and JavaScript. All of the web operations were implemented in the Windows operation system (64). An overview of the construction of CHDGKB is shown in Figure 2.

Browsing and data retrieval

Users can browse the variation by choosing CHD type, variation type, gene type (protein coding or miRNA) or variation consequence (e.g. missense variant, stop gained and intron variant).

Users can search for the detailed information on genetic variants on query interfaces through the following processes:

- Search with key gene names: users can search for the details on a given genetic variations by entering its abbreviated gene name.
- Search with drop-down menu: users can search for any genetic variation types, CHD types/subtypes and variation consequence by selecting the terms from the drop-down menu.



Figure 2. The entity relationship diagram of the CHDGKB.

Functional enrichment analysis

Due to the complexity of subtypes of CHD, we first divided all the subtypes in our database into two main categories: isolated CHD and non-isolated CHD. Cases of isolated CHD have a single type of abnormality associated with the condition, whereas cases of non-isolated CHD have two or more types of abnormality. To further analyze the genetic factors and their correlations with isolated and nonisolated CHD, we performed Gene Ontology (GO) annotation using the database for annotation, visualization and integrated discovery (R package, ClusterProfiler). KEGG pathway was provided for an enrichment analysis and the Benjamini–Hochberg method was applied to adjust raw *P*values. Based on the top 10 significantly enriched terms (adj. *P*-value <0.05), further studies of the associations were conducted with CHD through a literature validation.

Protein positional conservation analyzing

The diversity of the distribution of amino acids in CHDGKB was measured by calculating entropies for sequences containing mutations at different positions in multiple sequence alignments. Mutual information was calculated to quantify positional co-variation (12).

Data download and submission

All of the NS-CHD data are downloadable as Excel files (http://www.sysbio.org.cn/CHDGKB/Download.html).The search interfaces of the two search methods are depicted in Figure 3A, with links to the original publications

(Figure 3B). Figure 3C shows the interface for search results using key words, with links to the original publications.

The NS-CHD research data can be submitted to community-recognized repositories at http://www.sysbio.o rg.cn/CHDGKB through the 'Submit' interface without a username or password.

Results

Database content and statistics

Our CHDGKB covers details from 284 individual studies conducted in more than 23 countries, with 4510 small variations, 531 copy number variations (CNVs), 17 methylations and 17 other genetic variations distributed in 370 NS-CHD subtypes. The small variations include 3493 items of SNPs, 981 items of gene mutations (NOT SNP), 12 haplotypes and 24 others. The CHDGKB contains comprehensive information on genetic variations of four groups: CHD type, genetic variation, publication resources and sample information.

The genetic information in our CHDGKB comprises of four variation types: CNVs, small variations, methylation and others, with majority (88.36%) of which belonging to the small variations group (Figure 4). The database contains 981 items of gene mutations, which involve 135 different genes, 3493 SNPs in 204 genes, with CNVs of 289 duplications and 237 deletions, and 34 other independent genetic variations mentioned above. The top 10 most frequently occurred genes in our database are listed in Figure 5A. Among these 10 genes, GATA4, NKX2.5, MTHFR, GDF1, MTHFD1, etc. contain not only gene mutations but also SNP variations (Figure 5B).

A					
6	ายเ	\mathcal{T}	N/D	Conc	tice
			JIXD.	OTIP Gene	
		K	nowl	edgebase	
Home Browse	- Statistics	Search	Submit Document	Download About Us	
HD Type	Show 10	~ entries			Search:
tariation Type	ID A	Туре	Gene/Location	Detail	+ CHD Type
Small Variation	C0001	CNV	NA	NA	non-isolated
CNV	C0002	CNV	NA	chr1: g.144106312_145848312dup	Isolated CHD
Methylation	C0003	CNV	GIAS	chr1: 0 145658465 145658583dup	isolated CHD
Others	C0004	CNV	NA	chr1: g.144967972 146386972del	isolated CHD
lene Type					TGA
ariation consequence	C0005	CNV	NA	chr1: g.144967972_146292972del	isolated CHD ASD
	C0006	CNV	NA	chr1: g.144967972_146292972del	non-isolated CHD: MV, VS
	C0007	CNV	NA	chr1: g.144943150_146293150dup	isolated CHD TOF
	C0008	CNV	NA	chr1: g.144943150_146293150dup	isolated CHD: TOF
	C0009	CNV	NA	chr1: g.144943150_146293150dup	isolated CHD. TOF
				abril: a 144087070_146902072.due	
	C0010	CNV	NA	Gill 1. g. 14460/872_14028287200p	isolated CHD.
B mple Information	C0010 Showing 1	CNV to 10 of 2,6	NA 89 entries	Previous 1 2 3	isolated CHD. TOF
B mple information mple ID	C0010 Showing 1 1082	CNV to 10 of 2,6	NA 89 entries	Previous 1 2 3	isolated CHD TOF
mple Information mple ID spulation	C0010 Showing 1 1082 Chinese	CNV to 10 of 2,6	NA 89 entries	Previous 1 2 3	isolated CHD TOF
B mpte Information mpte ID putation urce	C0010 Showing 1 1082 Chinese peripheral 1	CNV to 10 of 2,6	NA 89 entries	Previous 1 2 3	isolate CPU TOF
B mple Information mple ID putation urce vatment Age	C0010 Showing 1 1082 Chinese peripheral I NA	CNV to 10 of 2,6	NA 89 entries	Provides 1 2 3	iodate ODA
B mple information mple ID pulation urce atatment Age infor	C0010 Showing 1 1082 Chinese perpheral I NA NA	CNV to 10 of 2,6 blood	NA 89 entries	Previous 1 2 3	isolate ODA
B mple Information mple ID pulation arce atment Age ntroi Age gion atment Number	C0010 Showing 1 1082 Chinese peripheral I NA Na Na Na 1/300	CNV to 10 of 2,6 blood	NA 89 entries	Previous 1 2 3	1004ee OD4
B mple Information mple ID pulation atment Age atment Age gion atment Number	C0010 Showing 1 1082 Chinese peripheral 1 NA NA Tartjin, Ch 1(300) 0(300)	CNV to 10 of 2,6 blood	NA 69 entries	Previous 1 2 3	indete CPU TOF
B mple Information mple ID putation urce atament Age ntrol Age gion 	C0010 Showing 1 1082 Chinese perpheral NA NA Tarigin, Ch 1(300) 0(300) PCR, DNA	CNV to 10 of 2,6 blood	NA 09 entries	Produce 1 2 3	indete CPU TOF
B mple biformation pulation urce astment Age gion autment Number autment Number theoremation	C0010 Showing 1 1082 Chinese perpheral NA NA Tarigin, Ch 1(300) 0(300) PCR, DNA	CNV to 10 of 2,6 blood ina	B9 entries	Produces 1 2 3	iodate Orb
B mpte internation mpte ID updation urree astment Age gion astment Number introl Number introl Number thod	C010 Showing 1 1082 Chinese perpheral NA NA Tanjio, Ch 1(300) 0(300) PCR, DNA 9 22764302	CNV to 10 of 2,6 blood ina	ABI3730)	Produce 1 2 3	iodete Orb
B mpte Information mpte ID updation urree urree ustanent Age gion ustanent Number uthod watanent Number thod watanent Number uthod u	C010 Showing 1 1082 Chirese perpheral NA NA Tarijo, Ch 1(300) 0(300) PCR, DNA 0 2(2764302 2012	CNV to 10 of 2,6 blood	NA 99 entries A813730)	Previous 1 2 3	Top Top 1 4 5 209 1
B mpte between and mpte 10 putation autrent Age autrent Age autren	C010 Showing 1 1082 Chinese perpheral 1 NA NA NA Tarqin, Ch 1(300) 0(300) P.CR, DNA n 22764302 2012 A novel val	CNV CNV to 10 of 2,6 blood ina sequecing(NA 69 entries AB13730)	vin i y (Helpol y zHouse brought coup Previous 1 2 3	iodate Orb
B mpte Information mpte ID putation atment Age atment A	C010 Showing 1 1082 Chinese perjaharal NA NA NA Targija, Ch 1(300) 0(300) PCR, DNA n 22764302 2012 A novel val	CNV to 10 of 2.6 blood	NA 99 entries AB33730) AGL1 In Chinese patients v	with isolated venificular septal defect	iodate Orb
B mple ID unter 10 putation	C010 Showing 1 1082 Chinese peripheral NA NA NA Targin, Ch 1(300) P.CR, DA 2012 2012 A novel var	CNV to 10 of 2,6 blood	NA 99 entries ABI3730) AGL1 In Chinese patients v	white previous fragmentation of the second defect.	hodete Chi Top b 4 5 209 1
B angle Information mple ID putation notor Age introl Age introl Age introl Age introl Mumber thod thod thod thod thod are introl Mumber introl Mumber int	C010 Showing 1 Showing 1 Chinese perpheral NA NA Tarqin, Ch 1(300) D(300) PCR, DNA 0 2012 2012 A novel vas aristion Type	CNV to 10 of 2.6 blood ina sequecing(NA 99 entries ABI3730) AGL1 in Chinese patients v	with isolated ventroutur septal defect	lioide Orb
B mpte boto-mation mpte b putation	C010 Showing 1 D02 Chinese perspheral I NA NA Tangin, Ch 1(300) 0(300) PCR, DNA 0 2012 A novel var antation Type Variation Type	CNV to 10 of 2.6 blood ina sequecing(99 entries 99 entries ABI3730)	white investment of the second	indete CPU 106 1 4 5 209 1
B mpie to formation mpie to publics urce atment Age gion atment Number ntrol Number atment Number atment States ference Information atment ference Information atment Com accession access	C000 Showing 1 ORZ Chinese peripheral NA NA Tarpin, Ch 1(300) 0(300) PCR, DNA n 22764302 2012 A novel vas aristion Type Variation Type	CNV to 10 of 2.6 bitood	NA 99 entries ABI3730) AGL1 In Chinese patients v	with isolated venitoular septal defect	v lindete CPU lindete CPU li
B mple ID = U public J public J stateet Age stateet Number stateet Number terence Information ference ID ar is Choose Choose Choose	C010 Showing 1 Object Chinese peripheral NA NA NA 1(300) 0	to 10 of 2,6	ABI3730)	white included ventioular septial defect	Violate CPU 10 4 5 209 1
B angle Information mple ID putation tarret atomer Appe atomer Appe atomer Appe atomer Appe atomer Appe atomer Appe accession acces	C010 Showing 1 Chinese perphase paratement perphase perph	CNV to 10 of 2.6 blood ina sequecing(ration of PL st st st ration of PL	NA 99 entries ABI3730) AGL1 In Chinese patients v	unity (response arg_modular) adapt Previous 1 2 1 2 1	v v v v v v v v v v v v v v
B mite latorenation mite latorenation putat	C0100 Showing 1 1082 Chinese peripheral 1 NA NA Tarqin, Cri 1(300) 0(300) P.CR, DA30 202764302 20276430 20276450 20276767650 2027676767676767676767676767676767676767	In the sequencing of PL	NA 69 entries ABI3730) AGL1 In Chinese patients v iation)	unity (Hindol & Figure 400, BF) Previous 1 2 2	
B mpie formation mpie to publics urce atment Aga gion atment Number ntrol Number ntrol Number atment Number atment States ference Information acentic Vi acentic Vi A	C000 Showing 1 ORZ Chinese peripheral NA NA Tarqin, Cri 1(300) 0(CNV to 10 of 2,6 blood ina sequecing(rtation of PL station of PL	NA 89 entries ABS3730) AGL1 In Chinese patients v tation)	with isolated venitoular septal defect	2014 000 100 100 100 100 100 100 100 100
B september of the second sec	C0000 Showing 1 Chices Chices Competition Chices Competition Chices Competition Chices Competition Competition Competition Competition Competition Competition Competition Chices ChiceS ChiceS Competition ChiceS Competition	CNV to 10 of 2,6 blood ina sequecing(riation of PL	NA 99 entries ABI3730) AGL1 In Chinese patients v IaBion)	with isolated venticular sepiral defect	
B angle Information mptel D putation turce atment Age atment Age atment Adage atment Amather atment A	C010 Showing 1 Chinese perghenal NA NA Targin, Ch 10200 0(300) 0(CNV to 10 of 2,6 blood ina sequecing(ration of PL	ABI3730) AGL1 In Chinese patients s tationy	with isolated ventioular septal defect	

Figure 3. An example of the search interface with CHD subtype and precise query. Detailed information about the genetic variations can be found in the NS-CHD database. [The search interfaces are depicted in Figure 3A, with links to the original publications (Figure 3B). Figure 3C show the interface for search results using key words, e.g. variations with GATA4].

Comparison with existing databases

Thus far, several public genetic databases associated with disease have been well constructed, such as ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), OMIM (https://omim.org), DisGeNet (http://www.disgenet.org/home/), Decipher (https://decipher.sanger.ac.uk), etc. As illustrated in Table 1, our database (CHDGKB) has the following advantages.

- 1. As a database specific for NS-CHD, CHDGKB records genetic variation information for more than 370 NS-CHD subtypes.
- 2. CHDGKB is a comprehensive database of variation types of gene mutation, SNP variations, copy number



Figure 4. The constitution of genetic variations in the CHDGKB.





Figure 5. The top 10 genes associated with NS-CHD in the CHDGKB. (Figure 5A shows the top 10 genes with all variations for NS-CHD; Figure 5B shows the top 10 genes associated with SNP variation.)

variations and methylation variations associated with NS-CHD.

3. The annotations in our database contain specific information for future translational applications, such as patient sample information, inheritance information and risk level for risk evaluation, diagnosis, prognosis, etc.

GO analysis and pathway mapping

R package ClusterProfiler was used to the GO analysis of isolated and non-isolated CHD at three levels: biological process (BP), cellular component (CC) and molecular function (MF). The top 10 significant enriched terms (P < 0.05) of each level and number of genes for the two main kinds of CHD are summarized in Figures 6 and 7, respectively. At

	CHD subtypes	Variation Types				Methylations					Sample	
Database/ items		Gene mutation	SNP	Allel	e CNV	Aneuploidy	Location	Methylation levels	Risk levels	Locations	Inheritance	information
ClinVar	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark	×	×	×	\checkmark	×	\checkmark
OMIM							\checkmark	\checkmark	×	۰ ا	\checkmark	
DisGeNet							×	×	×	۰ ا	×	×
ISCA							\checkmark	\checkmark	×	۰ ا	\checkmark	×
DECIPHER									×		×	\checkmark
v9.31 CHD GKB	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark

Abbreviations: CHD, congenital heart disease; GKB, genetic knowledge database; InDel, insertion and deletion; SNP, single nucleotide polymorphism; CNV, copy number variation

the BP level, for both isolated and non-isolated CHD, the most significant terms were mainly related to cardiac chamber morphogenesis and development, mesenchyme development and heart morphogenesis. Researchers have demonstrated that mutations in some key genes, such as TBX20 (13), NKX2.5 (14) and CELSR1 (15), play a vital role in the heart morphogenesis process through down-regulation or up-regulation of the correlated genes. Furthermore, Bose et al. (16) revealed that mutations in non-coding regions of GATA4 could also affect the process during fetal heart development. At the CC level, the enriched terms for the two kinds of CHD were focused on adherents junction (17), contractile fiber part (18) and transcription factor complex (16). It has been shown that mutations occur in the whole coding region and splice junction sites of the PITX2c gene, which encodes paired-like home domain transcription factor 2 and is crucial for normal cardiovascular morphogenesis. Also, it is well known that GATA6, which is also a gene encoding a zinc finger transcription factor, plays a significant role in the core cardiac transcriptional factor pathway. At the MF level, results of significant enriched terms for isolated and non-isolated CHD are mapped with not only DNAbinding transcription factor activity (19), enhancer binding (20) and SMAD binding (21) but also proximal promoter sequence-specific DNA binding (22). Meanwhile, we provided KEGG pathways for the enrichment analysis. The top eight significant enriched terms of KEGG pathways along with number of genes for isolated and non-isolated CHD can be seen in Figure 8. Series of MAPK signaling pathways (23), signaling pathways regulating pluripotency of stem cells (24) and Rap1 signaling pathway (25) are essential for occurrence of both isolated and non-isolated CHD.

Protein positional conservation analyzing with 3-fold conservation algorithm

To explore the pathogenesis of NS-CHD at the molecular level, we applied this three-type conservation algorithm into analyzing some important proteins, e.g. transcription factor GATA-4, on which 32 disease-causing mutations were reported over 100 times and well collected in our database. We got MSA for this protein from protein family (https://pfam.xfam.org/) (26). It contains 313 alignments and the sequence length is 205. Three kinds of conservation values are calculated and listed in Table 2–4.

Discussion

Prior to our current study, several systematic studies had been focused on psychiatric disorders, such as the bipolar affective disorder, the attention deficit disease and autism (27-30). Researchers have also developed genetic databases on Parkinson's disease, diabetic retinopathy and hepatocellular cancer (31-33). Furthermore, there were similar diseases-associated gene/genetic variation database for hypertension, obesity (34), diabetes (35), coronary artery disease (36) and aortic aneurysm (37). Here we present, to our knowledge, the first open database for NS-CHD genetic variants, with numerous predictive functions. This database represents an up-to-date, comprehensive synopsis for the NS-CHD genetics research. It can be used for the SNP analysis, as well as meta-analysis such as ethnicity-specific meta-analysis, on the NS-CHD riskgene candidates based on all 981 items of gene mutations and 3493 items of SNP variation, following stratification for different countries.

Using the GO annotation, we performed further bioinformatic analyses on the isolated and non-isolated CHD. At the BP level, for the enriched terms 'heart morphogenesis and cardiac chamber development' shared by both two CHD types, the BP of cardiac septum morphogenesis was mainly related to isolated CHD. For example, NKX2.5 influences the process of heart growth by up-regulating its target genes, including those involved in the atrial septal defect (14) and those necessary to maintain chamber-





Figure 6. The top 10 significant enriched GO terms for isolated CHD. The statistical significance level (p.adjust, adjusted *P*-value) was depicted as different color.(*X*-axis indicated number of enriched genes; *Y*-axis indicated GO terms.)

specific identity in both the first and second heart field (38). On the other hand, studies on non-isolated CHD mostly focused on regulation of pathways via vital genes. Mutation in CELSR1 (P870L) was shown to correlate with various CHD subtypes, including septation or conotruncal defects (15) through up-regulation of the PCP pathway and the canonical WNT signaling in cells. This further demonstrates the complex functions of genes and their interactions involved in the cardiac ventricular development (39) during the BP of CHD. At the CC level, isolated

CHD was focused on the membrane region, which was probably due to that heart membrane microdomains are enriched in chaperones, cytoskeletal-associated proteins, enzymes and protein involved in signal transduction pathway (40). However, it has been demonstrated that the main morphologic of cardiocyte differences were dependent on nuclear chromatin activity/stainability and nuclear breadth (41), which can explain why the non-isolated CHD was correlated with the CC terms of nuclear chromatin. At the MF level, except for the shared significant terms, the





Figure 7. The top 10 significant enriched GO terms for non-isolated CHD. The statistical significance level (p.adjust, adjusted *P*-value) was depicted as different color. (*X*-axis indicated number of enriched genes; *Y*-axis indicated GO terms.)

enriched terms in isolated CHD are mainly correlated with transmembrane receptor protein kinase activity (40) and growth factor activity, which both play an important role in the process of message transfer to regulate and affect transcription (42, 43). The phenotypes of non-isolated CHD are more complex than those of isolated CHD in terms of nucleotide binding and interaction (44). Consequently, the distinct enriched terms of non-isolated CHD are activating transcription factor binding and protein binding.

On the basis of the associated pathways in isolated CHD, studies of immunohistochemistry for molecules in the

TGF- β signaling pathway have demonstrated that ongoing tissue remodeling of the coronary artery disease after the acute injury and confirmed the importance of the TGF- β signaling pathway in this process (45). Besides, there is another pathway involved in the process of isolated CHD, the Hippo signaling pathway, which can regulate embryonic cardiomyocyte proliferation and heart size during development through YAP (46). Studies have revealed its mechanism was that Hippo crosstalk with Wnt/ β -catenin signaling can play a critical role in mediating the positive effect of YAP on cell cycle-related gene expression and cardiac



Figure 8. Pathway enrichment analysis for genetic variations of NS-CHD. The statistical significance level (p.adjust, adjusted *P*-value) was depicted as different color. (Figure 8A, the top eight significant KEGG terms for isolated CHD; Figure 8B, the top eight significant KEGG terms for non-isolated CHD. *X*-axis indicated number of enriched genes; *Y*-axis indicated enriched pathways.)

No	Type I conse	ervation: 20 amino	acid alphabets	Type II conservation: divided amino acids with physicochemical properties (six amino acid categories)			
	Residue	Rank	Information value	Residue	Rank	Information value	
1	43R	6	2.936	154F	5	2.955	
2	6A	24	2.666	110F	11	2.920	
3	44V	25	2.659	95D	20	2.575	
4	51L	27	2.647	193P	29	1.992	
5	5L	30	2.597	116S	37	1.614	
6	193P	41	2.326	118A	45	1.390	
7	55Q	45	2.280	197N	49	1.356	
8	110F	47	2.211	150G	51	1.300	
9	46S	57	2.088	90S	66	1.093	
10	116S	59	2.073	96G	75	1.006	
11	95D	60	2.073	163P	76	0.995	
12	33A	63	2.015	93G	83	0.833	
13	118A	70	1.897	125A	86	0.781	
14	9A	94	1.741	144A	87	0.777	
15	163P	98	1.720	167A	90	0.679	

Table 2. Type I and II residue conservations on protein transcription factor GATA-4*

*The table only list reported disease-causing mutation residues with information value ranking top 100.

Table 3. Type III mutual residue conservations on protein transcription factor GATA-4*

No		20 aas alphabet		Physicochemical properties (six categories)				
	Residue1	Residue2	Mutual information value	Residue1	Residue2	Mutual information value		
1	118A	119A	182.031	163P	164A	184.663		
2	117L	144A	139.724	118A	120A	176.495		
3	117L	154F	163.434	90S	167A	120.454		
4	163P	164A	179.671	118A	139A	121.935		
5	167A	180S	138.668	163P	167A	119.326		
6	93G	106P	122.115	157S	163P	115.311		
7	154F	164A	131.230	167A	180S	126.966		
8	154F	163P	124.885					
9	117L	121A	134.008					

*The table only list pairs with top 20 high mutual information containing reported disease-causing mutation residues.

117L

154F

154F

154F

154F

154F

No		20 amino ao	cid alphabets		Physicochemical properties (6 amino acid categories)				
	Residue1	Residue2	Residue3	Triplet rank	Residue1	Residue2	Residue3	Triplet rank	
1	97A	117L	154F	2	98A	102P	163P	1	
2	97A	123A	154F	3	102P	1095	163P	3	
3	98A	102P	163P	4	102P	1595	163P	5	
4	99Y	1055	154F	7					

8

11

12

17

19

2.0

Table 4. High conserved triplet with disease-causing mutations on protein Transcription factor GATA-4*

*The table only list the triplets among top 20 most conserved one.

107R

117L

136G

107R

117L

12.3A

99Y

99Y

99Y

105S

105S

105S

5

6

7

8

9

10

overgrowth (46, 47). Compared with the influence in the initial process of heart development via hippo signaling pathway, the different significant enriched pathway correlated with non-isolated CHD focused on the cGMP-PKG signaling pathway, which are both recognized modulators of cardiac function and the chronic stress response (48). Tsai et al. (49) have revealed that when their enhancing relax, cGMP/PKG serve as a myocardial brake, countering cAMP stimulation and independently signaling alternative pathways to blunt contraction and growth. Furthermore, studies have suggested the various cGMP regulating phosphodiesterases and confirmed their proposed interactions with cGMP, cAMP and PKG myocyte target (49, 50). The complex interactions among these pathways subsequently involved in the occurrence of non-isolated CHD via reducing maladaptive hypertrophy, improving cell survival, regulating signaling and mitochondrial function, protecting against ischemia/reperfusion injury and blunting the stimulatory effects of catecholamines (48, 51).

Based on our previous study (12), we made further analysis of function mechanism via 3-fold conservation algorithm toward proteins at molecular level. Entropy calculations are used to identify types I and II conservation. For type I conservation, the normal alphabet of 20 amino acids is used while the amino acid alphabet is divided into six physiochemically in type II conservation (52). The six categories are hydrophobic (VILFMWYC), negatively charged (D E), positively charged (R K H), conformational (G P), polar (N Q S) and (A T).

For the types I conservation or single residue conservation (Table 1), there are 15 residues with reported mutations are regarded as highly conserved (top 10). The residue 43R (ranking 6th of all 205 residues) is seven times reported as disease related in our database; the mutation in this residue is R43W. If we divided amino acids with physicochemical properties (type II conservation), also 15 residues with disease-causing mutations are regarded as highly conserved (top 100), six of which are also regarded

as highly conserved in type I (110F, 95D, 193P, 116S, 118A and 163P). The residue 154F is found to be most highly conserved, ranking fifth of all positions, deletion mutation in this residue is three times reported. For the type II conservation or the mutual conservation, we calculated information values focused on specific pairs of residues. The disease-causing mutations related to type II conservation are listed in Table 2. There are nine such pairs of residues for the normal alphabet of 20 amino acids and eight pairs for the six physicochemical alphabet which have top 20 mutual information. These pairs contain nine different disease-causing mutations reported in the database. The pair 118A and 119A has third biggest mutual information value among all pairs, and the diseaserelated mutation between the two residues is an insertion. For physicochemical properties, the pair 163P and 164A has the second highest mutual values, and there was a substitution P163S reported. Furthermore, we can find conserved triplets, which means co-variation among three residues, according to the mutual information. Those highly conserved triplets with six disease-causing mutations are listed in Table 3. The two residues, 154F and 163P, play most important roles in triplet conservation in 20 amino acids alphabet and six physicochemical properties group, separately.

We made an analysis of GATA4 variation as an example, which was reported as one of the most genetic variation associated with NS-CHD. We found that some single residues with high conservation value might not be important in co-variation or triplet, which means co-variation with three different residues. Some residues might need further analyzation in future, like 117L, 154F and 118A, because these residues are included in several pairs of co-variation and triplets.

In brief, our CHDGKB provides comprehensive results, with which we can perform statistical and systematic analyses to further our understanding of how genetic factors influence the pathogenesis of CHD. For example, GATA factors regulate a large number of cardiac genes, including NPPA, NPPB, MEF2c, NKX2.5, BMP4, MYH6 and MYH2 (53), and the complex gene interactions, regulations and the resulting functional variation of proteins all affect the progression of CHD. Our future direction will be to explore the different enriched gene network and study the mechanisms involved in the various subtypes of NS-CHD. Moreover, our CHDGKB also provides homogeneous results using a statistical analysis to enable investigations into the correlations between genotypes and phenotypes of NS-CHD. For example, researchers can examine mutations in GATA4 and NKX2.5 and their correlations to specific NS-CHD subtypes. Through browsers such as the Human Genetic Variation Browser (54), we can study the allele frequency of genetic variations that cause non-synonymous amino acid changes in patients included in our database. Using statistical methods such as the weighted genetic score, we can calculate the genetic score (55) and logistic regression model (56) on the information pulled from our CHDGKB. This will greatly improve risk assessment and prediction of NS-CHD in the near future.

There are some limitations in our NS-CHD database, and future improvements have already been planned. The current version documents 5345 variations in populations from 24 countries until July 2019, and it will be continually updated. Most of the data in the current version came from the Asian population, which was the actual distribution in PubMed. Furthermore, the genetic information in our database will be expanded to include gene expression, functional variations and other more comprehensive genetic parameters in future, in order to further investigate the relationships between NS-CHD and genetic variations. Using our database, additional systematic analyses on the molecular mechanisms of NS-CHD are also planned to improve and expand its clinical applications.

Author Contributions

The authors' responsibilities were as follows: Bairong Shen, Lan Yang and Xingyun Liu designed the research; Lan Yang, Xingyun Liu, Yalan Chen and Yuxin Lin performed literature search, selection and data extraction; Xingyun Liu constructed the database; Yang Yang and Yongquan Chen performed the functional analysis of the amino acid mutations. Lan Yang, Yan Sun and Bairong Shen write the manuscript; Bairong Shen conceived and supervised the work. All the authors completely consented with all the data in the study, critically revised the manuscript for important intellectual content and approved the final version.

Additional Information

Data resource access: http://www.sysbio.org.cn/CHDGKB/

Conflict of interests

None declared.

Funding

National Natural Science Foundation of China (31670851, 61602332); National Key Research and Development Program of China (2016YFC1306605); Natural Science Foundation of the Jiangsu Higher Education Institutions of China (18KJD520003).

References

1. Writing Group Members, Mozaffarian, D., Benjamin, E.J. *et al.* (2016) Heart disease and stroke statistics-2016 update: a report

from the American Heart Association. Circulation, 133, e38-e360.

- 2. van der Bom,T., Zomer,A.C., Zwinderman,A.H. *et al.* (2011) The changing epidemiology of congenital heart disease. *Nat. Rev. Cardiol.*, **8**, 50–60.
- 3. Patel,S.S. and Burns,T.L. (2013) Nongenetic risk factors and congenital heart defects. *Pediatr. Cardiol.*, 34, 1535–1555.
- Luo,Z., Shen,Y., Chen,W. *et al.* (2017) Association analysis of nonsyndromic congenital heart disease and tag single nucleotide polymorphisms of TBX20 and genes in the Ras-MAPK pathway. *Genet. Test. Mol. Biomarkers*, 21, 440–444.
- 5. An,Y., Duan,W., Huang,G. *et al.* (2016) Genome-wide copy number variant analysis for congenital ventricular septal defects in Chinese Han population. *BMC Med. Genomics*, 9, 2.
- Huang,R.T., Xue,S., Xu,Y.J. et al. (2014) Somatic GATA5 mutations in sporadic tetralogy of Fallot. Int. J. Mol. Med., 33, 1227–1235.
- Wang, J., Hu, X.Q., Guo, Y.H. *et al.* (2017) HAND1 loss-offunction mutation causes tetralogy of Fallot. *Pediatr. Cardiol.*, 38, 547–557.
- Yang,X.F., Zhang,Y.F., Zhao,C.F. *et al.* (2016) Relationship between TBX20 gene polymorphism and congenital heart disease. *Genet. Mol. Res.*, 15, 1–9.
- Thienpont,B., Mertens,L., de Ravel,T. *et al.* (2007) Submicroscopic chromosomal imbalances detected by array-CGH are a frequent cause of congenital heart defects in selected patients. *Eur. Heart J.*, 28, 2778–2784.
- St Louis, J.D., Jodhka, U., Jacobs, J.P. *et al.* (2014) Contemporary outcomes of complete atrioventricular septal defect repair: analysis of the Society of Thoracic Surgeons Congenital Heart Surgery Database. *J. Thorac. Cardiovasc. Surg.*, 148, 2526–2531.
- Patel,A., Costello,J.M., Backer,C.L. *et al.* (2016) Prevalence of noncardiac and genetic abnormalities in neonates undergoing cardiac operations: analysis of The Society of Thoracic Surgeons Congenital Heart Surgery Database. *Ann. Thorac. Surg.*, 102, 1607–1614.
- 12. Shen,B. and Vihinen,M. (2004) Conservation and covariance in PH domain sequences: physicochemical profile and information theoretical analysis of XLA-causing mutations in the Btk PH domain. *Protein Eng. Des. Sel.*, **17**, 267–276.
- Boogerd,C.J., Zhu,X., Aneas,I. *et al.* (2018) Tbx20 is required in mid-gestation cardiomyocytes and plays a central role in atrial development. *Circ. Res.*, 123, 428–442.
- Zakariyah,A.F., Rajgara,R.F., Veinot,J.P. et al. (2017) Congenital heart defect causing mutation in Nkx2.5 displays in vivo functional deficit. J. Mol. Cell. Cardiol., 105, 89–98.
- Qiao,X., Liu,Y., Li,P. *et al.* (2016) Genetic analysis of rare coding mutations of CELSR1-3 in congenital heart and neural tube defects in Chinese people. *Clin. Sci. (Lond.)*, 130, 2329–2340.
- Bose,D., Vaigundan,D., Shetty,M. *et al.* (2017) Identification of intronic-splice site mutations in GATA4 gene in Indian patients with congenital heart disease. *Mutat. Res.*, 803–805, 26–34.
- Sinn,H.W., Balsamo,J., Lilien,J. *et al.* (2002) Localization of the novel Xin protein to the adherens junction complex in cardiac and skeletal muscle during development. *Dev. Dyn.*, 225, 1–13.

- Henze,M., Patrick,S.E., Hinken,A. *et al.* (2013) New insights into the functional significance of the acidic region of the unique N-terminal extension of cardiac troponin I. *Biochim. Biophys. Acta*, 1833, 823–832.
- Tromp,J., Westenbrink,B.D., Ouwerkerk,W. *et al.* (2018) Identifying pathophysiological mechanisms in heart failure with reduced versus preserved ejection fraction. *J. Am. Coll. Cardiol.*, 72, 1081–1090.
- Kim,J.O., Kwon,E.J., Song,D.W. *et al.* (2016) miR-185 inhibits endoplasmic reticulum stress-induced apoptosis by targeting Na+/H+ exchanger-1 in the heart. *BMB Rep.*, 49, 208–213.
- 21. Mandel,E.M., Kaltenbrun,E., Callis,T.E. *et al.* (2010) The BMP pathway acts to directly regulate Tbx20 in the developing heart. *Development*, 137, 1919–1929.
- 22. Sindhu,K.V., Rani,V., Gupta,M.K. *et al.* (2004) Isolation of a library of target-sites for sequence specific DNA binding proteins from chick embryonic heart: a potential tool for identifying novel transcriptional regulators involved in embryonic development. *Biochem. Biophys. Res. Commun.*, 323, 912–919.
- 23. Li,Y., Zhao,W., Shi,R. *et al.* (2018) Rs4759314 polymorphism located in HOTAIR is associated with the risk of congenital heart disease by alternating downstream signaling via reducing its expression. *J. Cell. Biochem.*, **119**, 8112–8122.
- Chimenti, I., Forte, E., Angelini, F. et al. (2013) Biochemistry and biology: heart-to-heart to investigate cardiac progenitor cells. Biochim. Biophys. Acta, 1830, 2459–2469.
- Zhang,Q., Feng,M., Zhang,H. *et al.* (2018) Long noncoding RNA uc.4 inhibits cell differentiation in heart development by altering DNA methylation. *J. Cell. Biochem* 120, 8061–8068.
- El-Gebali,S., Mistry,J., Bateman,A. *et al.* (2019) The Pfam protein families database in 2019. *Nucleic Acids Res.*, 47, D427–D432.
- Chang,S.H., Gao,L., Li,Z. *et al.* (2013) BDgene: a genetic database for bipolar disorder and its overlap with schizophrenia and major depressive disorder. *Biol. Psychiatry*, 74, 727–733.
- 28. Matuszek,G. and Talebizadeh,Z. (2009) Autism Genetic Database (AGD): a comprehensive database including autism susceptibility gene-CNVs integrated with known noncoding RNAs and fragile sites. *BMC Med. Genet.*, 10, 102.
- 29. Mazoyer,B., Mellet,E., Perchey,G. *et al.* (2016) BIL&GIN: a neuroimaging, cognitive, behavioral, and genetic database for the study of human brain lateralization. *Neuroimage*, **124**, 1225–1231.
- Zhang,L., Chang,S., Li,Z. *et al.* (2012) ADHDgene: a genetic database for attention deficit hyperactivity disorder. *Nucleic Acids Res.*, 40, D1003–D1009.
- Ramana, J. (2012) RCDB: renal cancer gene database. BMC. Res. Notes, 5, 246.
- 32. Vidhya,G. and Anusha,B. (2014) Diaretinopathy database a gene database for diabetic retinopathy. *Bioinformation*, 10, 235–240.
- Mok,S.R., Mohan,S., Grewal,N. *et al.* (2016) A genetic database can be utilized to identify potential biomarkers for biphenotypic hepatocellular carcinoma-cholangiocarcinoma. *J. Gastrointest. Oncol.*, 7, 570–579.
- Dai,H.J., Wu,J.C., Tsai,R.T. *et al.* (2013) T-HOD: a literaturebased candidate gene database for hypertension, obesity and diabetes. *Database (Oxford)*, 2013, bas061.

- 35. Lim, J.E., Hong, K.W., Jin, H.S. *et al.* (2010) Type 2 diabetes genetic association database manually curated for the study design and odds ratio. *BMC Med. Inform. Decis. Mak.*, 10, 76.
- Liu,H., Liu,W., Liao,Y. *et al.* (2011) CADgene: a comprehensive database for coronary artery disease genes. *Nucleic Acids Res.*, 39, D991–D996.
- Sun, R., Cui, C., Zhou, Y. et al. (2018) AGD: Aneurysm Gene Database. Database (Oxford), 2018, 1–6
- George, V., Colombo, S. and Targoff, K.L. (2015) An early requirement for nkx2.5 ensures the first and second heart field ventricular identity and cardiac function into adulthood. *Dev. Biol.*, 400, 10–22.
- Wang,Z., Zhai,W., Richardson,J.A. *et al.* (2004) Polybromo protein BAF180 functions in mammalian cardiac chamber maturation. *Genes Dev.*, 18, 3106–3116.
- 40. Banfi, C., Brioschi, M., Wait, R. *et al.* (2006) Proteomic analysis of membrane microdomains derived from both failing and non-failing human hearts. *Proteomics*, **6**, 1976–1988.
- Nozynski, J., Zakliczynski, M., Konecka-Mrowka, D. et al. (2009) Pivotal karyometric measurements in different types of cardiomyopathic morphology: study of hearts explanted from transplant recipients. *Transplant. Proc.*, 41, 3179–3184.
- Liu,Y., Wang,F., Wu,Y. *et al.* (2014) Variations of CITED2 are associated with congenital heart disease (CHD) in Chinese population. *PLoS One*, 9, e98157.
- Zhang,W., Shen,L., Deng,Z. *et al.* (2014) Novel missense variants of ZFPM2/FOG2 identified in conotruncal heart defect patients do not impair interaction with GATA4. *PLoS One*, 9, e102379.
- Mahmut,A., Boulanger,M.C., Bouchareb,R. *et al.* (2015) Adenosine derived from ecto-nucleotidases in calcific aortic valve disease promotes mineralization through A2a adenosine receptor. *Cardiovasc. Res.*, 106, 109–120.
- Lee, A.M., Shimizu, C., Oharaseki, T. *et al.* (2015) Role of TGFbeta signaling in remodeling of noncoronary artery aneurysms in Kawasaki disease. *Pediatr. Dev. Pathol.*, 18, 310–317.

- 46. Heallen, T., Zhang, M., Wang, J. *et al.* (2011) Hippo pathway inhibits Wnt signaling to restrain cardiomyocyte proliferation and heart size. *Science*, 332, 458–461.
- Xin,M., Kim,Y., Sutherland,L.B. *et al.* (2011) Regulation of insulin-like growth factor signaling by Yap governs cardiomyocyte proliferation and embryonic heart size. *Sci. Signal.*, 4, ra70.
- Zhang,M. and Kass,D.A. (2011) Phosphodiesterases and cardiac cGMP: evolving roles and controversies. *Trends Pharmacol. Sci.*, 32, 360–365.
- Tsai,E.J. and Kass,D.A. (2009) Cyclic GMP signaling in cardiovascular pathophysiology and therapeutics. *Pharmacol. Ther.*, 122, 216–238.
- Takimoto, E., Champion, H.C., Li, M. *et al.* (2005) Chronic inhibition of cyclic GMP phosphodiesterase 5A prevents and reverses cardiac hypertrophy. *Nat. Med.*, 11, 214–222.
- Lu,Z., Xu,X., Hu,X. *et al.* (2010) Oxidative stress regulates left ventricular PDE5 expression in the failing heart. *Circulation*, 121, 1474–1483.
- 52. Shen,B. and Vihinen,M. (2004) Conservation and covariance in PH domain sequences: physicochemical profile and information theoretical analysis of XLA-causing mutations in the Btk PH domain. *Protein Eng. Des. Sel.*, 17, 267–276
- 53. Laforest,B. and Nemer,M. (2011) GATA5 interacts with GATA4 and GATA6 in outflow tract development. *Dev. Biol.*, 358, 368–378.
- Yoshida, A., Morisaki, H., Nakaji, M. *et al.* (2016) Genetic mutation analysis in Japanese patients with non-syndromic congenital heart disease. *J. Hum. Genet.*, 61, 157–162.
- 55. Zhu,M., Cheng,Y., Dai,J. et al. (2015) Genome-wide association study based risk prediction model in predicting lung cancer risk in Chinese. Zhonghua Liu Xing Bing Xue Za Zhi, 36, 1047–1052.
- Katsigiannis, S., Hamisch, C., Krischek, B. *et al.* (2020) Independent predictors for functional outcome after drainage of chronic subdural hematoma identified using a logistic regression model. *J. Neurosurg. Sci.*, 64, 133–140