Genomics Proteomics Bioinformatics 21 (2023) 216–227

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

DATABASE

CHDbase: A Comprehensive Knowledgebase for Congenital Heart Disease-related Genes and Clinical Manifestations

Wei-Zhen Zhou [1,#], Wenke Li [1,#], Huayan Shen [1,#], Ruby W. Wang [2], Wen Chen [1], Yujing Zhang [1], Qingyi Zeng [1], Hao Wang [1], Meng Yuan [1], Ziyi Zeng [1], Jinhui Cui [1], Chuan-Yun Li [3], Fred Y. Ye [2,*], Zhou Zhou [1,*]

[1] *State Key Laboratory of Cardiovascular Disease, Beijing Key Laboratory for Molecular Diagnostics of Cardiovascular Diseases, Center of Laboratory Medicine, Fuwai Hospital, National Center for Cardiovascular Diseases, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100037, China*
[2] *International Joint Informatics Laboratory & Jiangsu Key Laboratory of Data Engineering and Knowledge Service, School of Information Management, Nanjing University, Nanjing 210023, China*
[3] *Institute of Molecular Medicine, Peking University, Beijing 100871, China*

Received 1 August 2021; revised 23 May 2022; accepted 1 August 2022
Available online 10 August 2022

Handled by Jingfa Xiao

KEYWORDS

Congenital heart disease;
Congenital heart defect;
Database;
Genetics;
Classification

Abstract **Congenital heart disease** (CHD) is one of the most common causes of major birth defects, with a prevalence of 1%. Although an increasing number of studies have reported the etiology of CHD, the findings scattered throughout the literature are difficult to retrieve and utilize in research and clinical practice. We therefore developed CHDbase, an evidence-based knowledgebase of CHD-related genes and clinical manifestations manually curated from 1114 publications, linking 1124 susceptibility genes and 3591 variations to more than 300 CHD types and related syndromes. Metadata such as the information of each publication and the selected population and samples, the strategy of studies, and the major findings of studies were integrated with each item of the research record. We also integrated functional annotations through parsing ~ 50 **databases**/tools to facilitate the interpretation of these genes and variations in disease pathogenicity. We further prioritized the significance of these CHD-related genes with a gene interaction network approach and extracted a core CHD sub-network with 163 genes. The clear genetic landscape of CHD enables the phenotype

* Corresponding authors.
E-mail: zhouzhou@fuwaihospital.org (Zhou Z), yye@nju.edu.cn (Ye FY).
# Equal contribution.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2022.08.001

classification based on the shared genetic origin. Overall, CHDbase provides a comprehensive and freely available resource to study CHD susceptibilities, supporting a wide range of users in the scientific and medical communities. CHDbase is accessible at http://chddb.fwgenetics.org.

Introduction

Congenital heart disease (CHD) refers to abnormalities in cardiocirculatory structure or function that arise during embryonic development. With a prevalence as high as ~ 1% of live births [1], CHD is one of the most common causes of major birth defects [2] and imposes enormous health and economic burdens. Specifically, 3%–5% of CHD patients have a genetic syndrome manifesting not only as CHD but also as extracardiac defects, which is termed syndromic CHD; others present abnormalities restricted to the heart, which is termed nonsyndromic CHD. Family studies have shown the strong genetic susceptibilities underlying the pathogenesis of CHD [3,4], so it is essential to clarify the CHD genetic etiology in the diagnosis and treatment of CHD. Recently, numerous genetics studies have addressed the genetic susceptibilities underlying CHD. Briefly, it has been estimated that the chromosomal aberrations and aneuploidies account for 8%–10% of CHD, the copy number variations (CNVs) for 3%–10% of CHD [5], *de novo* and rare inherited variants for 8% and 2% of CHD, respectively [6].

However, as the findings of these genetic studies are largely scattered in literature, it is difficult to incorporate them into in-depth mechanism studies and clinical practice. The holistic picture of CHD susceptibilities, such as the number and features of CHD-related genes, the strength of each item of evidence, and the relationship between genetic elements and cardiac malformations, is still to be addressed. Notably, the adequate classification of CHD types could substantially facilitate the investigation of its pathogenesis, accurate diagnosis, and effective treatments, while the current practice largely depends on its clinical manifestations with high diversity [7–9]. Efficient molecular classification could certainly supplement the current practices, while it is still to be addressed due to the lack of well-organized genotype-phenotype correlations. Notably, one pilot study recently reported a CHD-RF-KB database [10] to curate genetic risk factors associated with nonsyndromic CHD from approximately 300 studies, while it is far from complete to provide a clear genetic landscape of CHD for basic and translational studies.

Here, we present a comprehensive knowledgebase of CHD with extensive research evidence and functional annotations. We also prioritized these CHD-related genes with a gene interaction network approach. We then classified the CHD types based on their shared genetic origin, facilitating a molecular diagnosis and treatment of CHD which supplements the traditional practices with clinical manifestations.

Database implementation

Data collection and processing

To integrate current knowledge of CHD, we searched the PubMed database to retrieve CHD-related publications, using keywords including "congenital heart disease", "heart defects", "gene", "proteomics", "expression", "copy number variation", "linkage", "association", and "sequencing" (File S1). Among more than 2700 search results, we confirmed the relevance of each publication by examining its title and abstract. We excluded publications about congenital cardiomyopathies or congenital heart rhythm disorders due to their distinct clinical presentation and underlying mechanisms. After this process, a total of 1114 research articles published between February 1992 and January 2020 were collected for further curation.

For each article, we reviewed the main text and additional materials to extract evidence of an association between genes/variations and CHD. If multiple experimental approaches or different datasets were used in one study, the evidence of this study was divided into multiple independent records. In total, 1353 items of evidence were extracted, which could be classified into six types as follows: (1) 159 items of "Genetic association" evidence, (2) 452 items of "SNV/Indel" evidence reporting single-nucleotide variants/insertion-deletion variants (SNVs/Indels) in CHD patients, (3) 63 items of "Expression" evidence linking differentially expressed genes to CHD, (4) 240 items of "CNV" evidence reporting CNVs detected in patients, (5) 15 items of "Linkage" evidence reporting CHD-related genomic regions by linkage studies, and (6) 424 items of "Other" evidence, such as the evidence from transgenic animal models and cell lines (**Figure 1**). For each item of evidence, comprehensive metadata, such as the information of each publication, the selected population and samples, the strategy of studies, and the major findings of association, were extracted manually and double-checked by two researchers (Figure 1). A detailed description of the collected information is listed in Table S1.

The disease name was standardized according to the 11th revision of the International Classification of Diseases (ICD-11) congenital cardiology terms [11]. Transcript-based SNVs/Indels were annotated by Ensembl Variant Effect Predictor (VEP, https://grch37.ensembl.org/Tools/VEP) [12] and confirmed by Mutalyzer [13] following the nomenclature recommendations of the Human Genome Variation Society (HGVS). Genomic coordinates are presented according to the human reference genome (GRCh37/hg19). Overall, CHDbase integrated 3813 gene–CHD associations and 3957 variation–CHD associations, which contains the association information for 1124 susceptibility genes, 1006 structural variations, and 2585 SNVs/Indels.

Prioritization of CHD-related genes

For gene prioritization, the traditional evidence-based prioritization approach typically assigns a gene to a weighted sum based on the number and type of supporting evidence. As the weight for each type of evidence is debatable, and the items of evidence from the literature may simply reflect hot research topics rather than the significance of genes, here we introduced a gene interaction network approach to prioritize the significance of CHD-related genes.

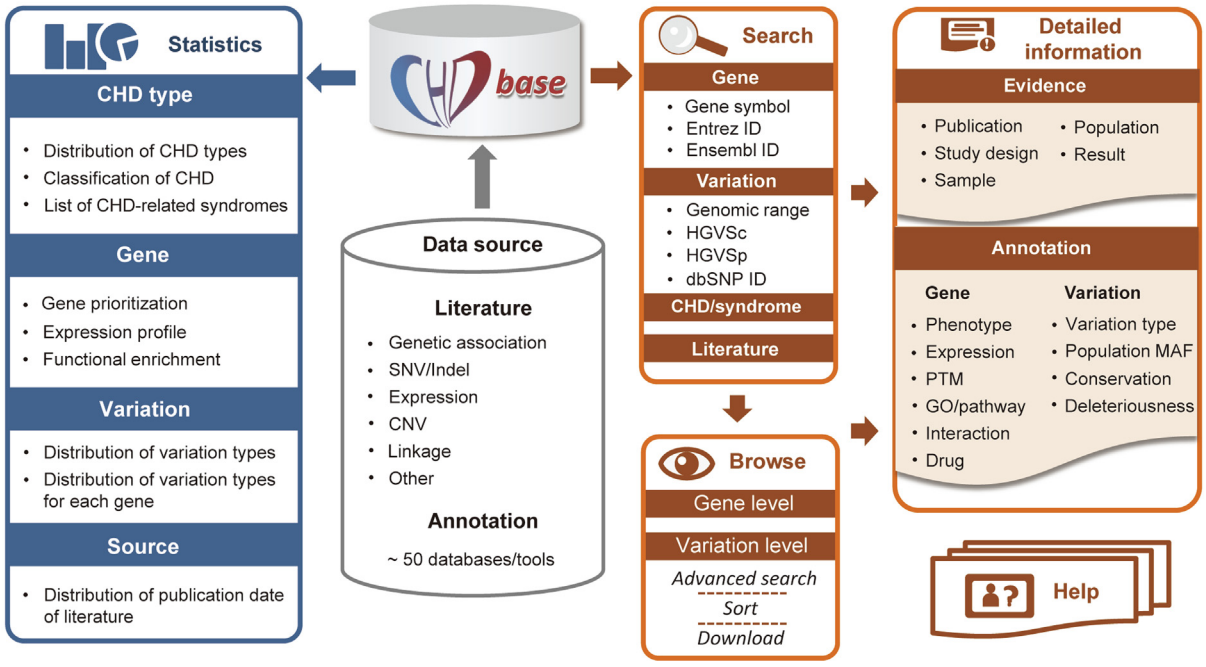


Figure 1 Overview of the framework of CHDbase

CHD, congenital heart disease; SNV, single nucleotide variant; Indel, insertion-deletion variant; CNV, copy number variation; ID, identity document; HGVSc, the Human Genome Variation Society expressions at the cDNA level; HGVSp, the Human Genome Variation Society expressions at the protein level; PTM, posttranslational modification; GO, Gene Ontology; MAF, minor allele frequency.

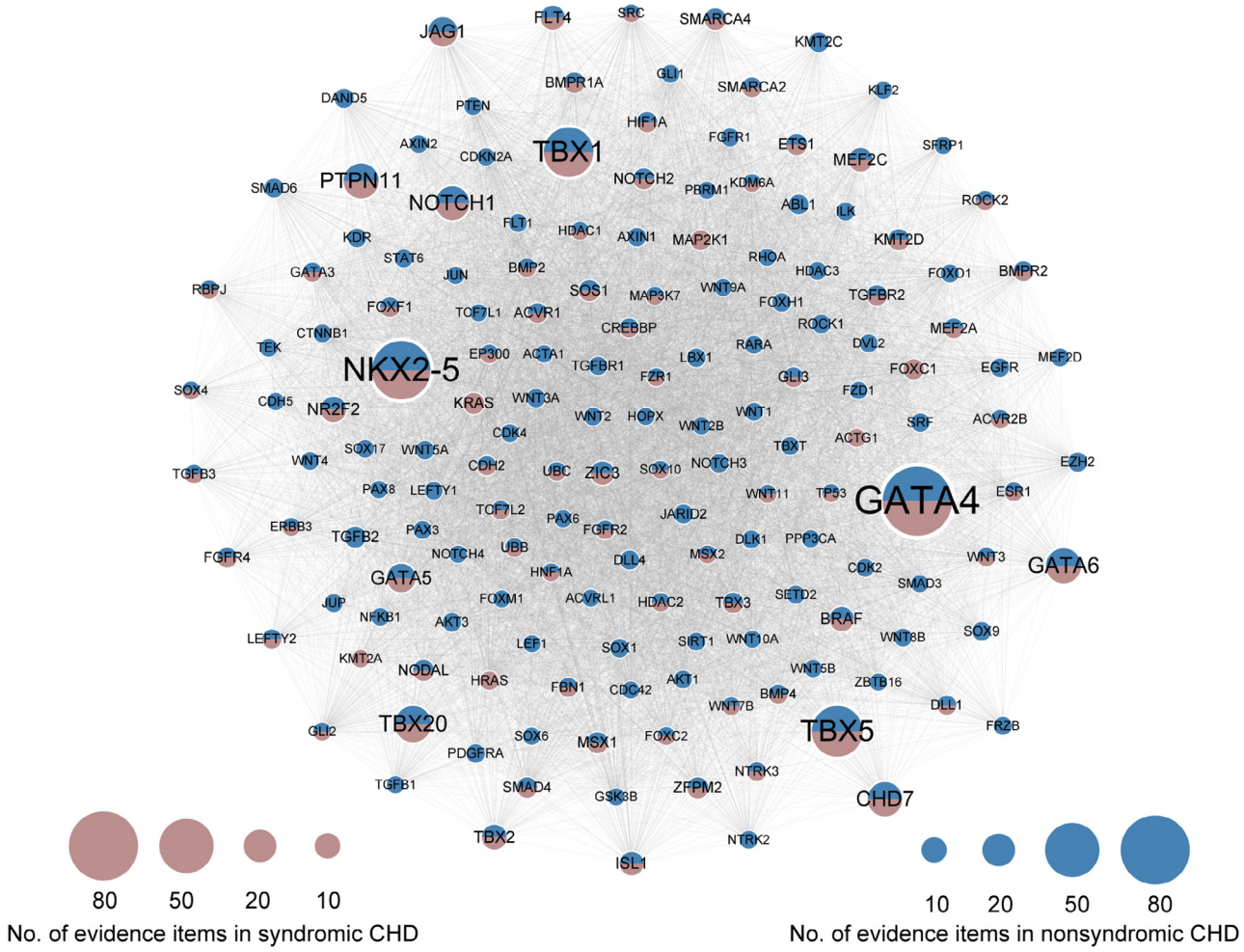
In CHDbase, 1124 genes have been linked to CHD susceptibility, including 115 reported only in syndromic CHD (syndromic genes), 597 reported only in nonsyndromic CHD (nonsyndromic genes), and 412 reported in both syndromic and nonsyndromic CHD (Figure S1). We further prioritized these susceptibility genes with a gene interaction network approach and extracted a core CHD gene set (File S1). Briefly, on the basis of the 1124 susceptibility genes and protein–protein interaction annotations from the STRING database [14], we first constructed an unweighted gene interaction network with 952 nodes and 30,777 edges. Then, we used three centrality scores, namely, degree, betweenness centrality, and eigenvector centrality, to measure the significance of each gene in the network. Specifically, a gene with a high degree could be considered a "hub" in the CHD network with more interactions with other genes. A gene with large betweenness centrality could be considered as a "bridge" in the network and the removal of such a gene could lead to a disconnected network. A gene with large eigenvector centrality has more interactions with high-degree genes. A gene with a high degree, large betweenness centrality, and large eigenvector centrality could be more important within the network.

To obtain the core gene set, we extracted a k-core with 163 genes from the CHD network using k-core decomposition (see supplementary methods in File S1). As we retained the nodes with a maximum value of k (k = 70), these 163 genes in the k-core were considered the most centrally located nodes in the original network. The core gene set contains six syndromic genes, 83 nonsyndromic genes, and 74 genes reported in both syndromic and nonsyndromic CHD, indicating that the syndromic and nonsyndromic CHD share a common gene network (**Figure 2**A). These core genes are supported by a significantly higher number of supporting evidence than that of all CHD-related genes (P = 4.55E−10, Mann–Whitney U test). Notably, for 38 high-confidence CHD genes recurrently reported in recent reviews [5,15–18], 18 are present in this core gene set (Figure 2B). To facilitate users to rank genes based on their own experiences and preferences, we provide the CHD susceptibility gene list with the metadata, such as the information indicating whether it belongs to the core gene set, its centrality scores, and the number of evidence items supporting its association with CHD, as shown in Table S2.

Functional annotations of CHD-related genes

To better understand the function of CHD-related genes, we further annotated them using public databases and data as follows: (1) HUGO Gene Nomenclature Committee (HGNC, https://www.genenames.org), National Center for Biotechnology Information (NCBI) Entrez gene (https://www.ncbi.nlm.nih.gov/gene) [19], Ensembl (https://www.ensembl.org), GeneCards (https://www.genecards.org) [20], Online Mendelian Inheritance in Man (OMIM, https://www.omim.org) [21], UniProt Knowledgebase (UniProtKB, https://www.uniprot.org), Mouse Genome Informatics (MGI, https://www.informatics.jax.org) [22], and Zebrafish Model Organism Database (ZFIN, https://www.zfin.org) [23] for basic information of the gene and its orthologs; (2) possible associated phenotypes and diseases from Human Phenotype Ontology (HPO,

A



B

	GATA4	*NKX2-5*	*TBX5*	*TBX1*	*TBX20*	*GATA6*	*NOTCH1*	*JAG1*	*GATA5*
Degree	237	203	171	163	120	221	310	238	176
Betweenness	3.8E–03	2.7E–03	2.7E–03	3.1E–03	8.2E–04	3.1E–03	1.5E–02	5.5E–03	4.2E–03
Eigenvector	0.72	0.67	0.53	0.49	0.37	0.69	0.85	0.72	0.55
No. of evidence items	83	66	54	52	31	29	27	20	19
	NR2F2	*NODAL*	*ETS1*	*ZFPM2*	*ACVR1*	*SMAD6*	*FOXH1*	*ACVR2B*	*PDGFRA*
Degree	172	114	106	155	200	153	122	179	160
Betweenness	2.5E–03	4.8E–04	4.1E–04	1.7E–03	2.2E–03	1.8E–03	7.3E–04	3.1E–03	2.4E–03
Eigenvector	0.53	0.43	0.38	0.50	0.63	0.52	0.41	0.53	0.52
No. of evidence items	14	7	7	7	5	4	3	3	3

Figure 2 Network-based prioritization of CHD-related genes

A. The *k*-core of CHD-related genes ($n = 163$). The node size is proportional to the number of supporting evidence items for each gene. **B.** For 18 core genes overlapping with the high-confidence CHD genes recurrently reported in recent reviews, the degree, betweenness centrality, eigenvector centrality, and the number of supporting evidence items for each gene are listed.

https://hpo.jax.org/app) [24], GWAS Catalog (https://www.ebi.ac.uk/gwas) [25], and DisGeNET (https://www.disgenet.org) [26]; (3) spatiotemporal expression profiles integrated from the Genotype-Tissue Expression (GTEx) project [27], RNA-seq data for seven human organs across multiple developmental time points [28], and spatial subcellular expression

data of the developing human heart at three developmental phases [29]; (4) posttranslational modifications retrieved from UniProtKB and dbPTM (https://awi.cuhk.edu.cn/dbPTM) [30]; (5) biological pathways and protein–protein interactions annotated by Gene Ontology (GO, https://www.geneontology.org), Reactome (https://reactome.org) [31], BioCyc (https://www.biocyc.org) [32], Kyoto Encyclopedia of Genes and Genomes (KEGG, https://www.genome.jp/kegg) [33], PANTHER (https://panther.celera.com) [34], STRING (https://cn.string-db.org) [14], OmniPath (https://omnipathdb.org) [35], BioGRID (https://thebiogrid.org) [36], and the Human Reference Interactome (HuRI, http://www.interactome-atlas.org) [37]; and (6) drug–gene interactions retrieved from the Drug Gene Interaction Database (DGIdb, https://dgidb.org) [38], DrugCentral (https://drugcentral.org) [39], and Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB, https://www.pharmgkb.org) [40] (Figure 1). Additionally, to better understand the effects of variations, we classified the variations into five categories according to the conclusions of the original publications: "disease-causing", "likely disease-causing", "uncertain", "likely benign", and "benign". For SNVs/Indels, we further integrated genomic information annotated by VEP [12], including variant consequences, minor allele frequency in different populations of the Genome Aggregation Database (gnomAD) [41], conservation scores calculated with phastCons, phylogenetic P-values (phyloP), and Genomic Evolutionary Rate Profiling (GERP++), and variant deleteriousness predicted by Sorting Intolerant from Tolerant (SIFT), Polymorphism Phenotyping v2 (PolyPhen2), Combined Annotation Dependent Depletion (CADD), Deleterious Annotation of genetic variants using Neural Networks (DANN), Likelihood Ratio Test (LRT), Mendelian Clinically Applicable Pathogenicity (M-CAP), MetaLR, PrimateAI, Functional Analysis Through Hidden Markov (FATHMM) using a Multiple Kernel Learning algorithm (FATHMM-MKL) and FATHMM with an eXtended Feature set (FATHMM-XF) (Figure 1).

Database interface

We constructed a MySQL database to store and manage the data. A user-friendly web interface was further implemented in Java, HTML, CSS, and JavaScript, as powered by Spring Boot, Thymeleaf, and Bootstrap framework.

User interface and functions

CHDbase provides two search modes. The basic search mode on the Home page and the top navigation bar allows users to precisely or fuzzily search for genes, variations, CHD/syndrome, and literature of interest (**Figure 3**A and B). The basic search engine can recognize query terms such as the gene symbols, Entrez IDs, Ensembl IDs, the range of genomic coordinates, HGVS expressions at the cDNA level (HGVSc), HGVS expressions at the protein level (HGVSp), IDs in the database of single nucleotide polymorphisms (dbSNP), the full name or abbreviation of diseases, and PubMed IDs. When the user clicks on the button of "Exact Search", a result page fully matching the search term displays detailed evidence of gene–CHD or variation–CHD association. If the user clicks on the "Fuzzy Search" button, the search engine will attempt to find partial matches for the search term, and return the results on the Browse page. The Browse page not only enables users to flexibly browse associations at the gene level or variation level, but enables the users to filter data with an advanced search mode (Figure 3C). Specifically, the user can enter multiple search terms in corresponding search boxes under column names to search for information of interest. Search results can be sorted by multiple columns in a customized order, and are downloadable in the format of JavaScript Object Notation (JSON) or Comma Separated Values (CSV). The user could further click on the gene or variation links on the Browse page to view the detailed information of research evidence and annotations as described below.

Detailed information about gene–CHD or variation–CHD associations is provided in a similar manner on the Gene Evidence Page or Variation Evidence Page (Figure 3D). Taking Gene Evidence Page as an example (Figure S2), in the upper panel, the list of research evidence for a specific gene relevant to CHD is summarized, including evidence ID, PubMed ID, title, abstract, study type, species, associated CHD types, associated syndromes, and conclusions. Furthermore, the user can click on the "Detail" label on the left of the "Evidence ID" column to view details in the lower panel. For detailed information on each item of evidence, we first provide population information, study design, and a summary of results, followed by the details of the results. According to different categories of evidence, the detailed information is summarized in varied formats (Figure S3). For example, for the evidence of "SNV/Indel" or "CNV", variation, variation type, associated CHD types, associated syndromes, the number of variations in affected individuals, the number of variations in unaffected individuals, and variation pathogenicity are provided. If the sample information is available, the user can click on the plus sign at the end of the row to view the information of individuals carrying this variation, including sample ID, family name, ethnicity, sex, age, diagnosis, other phenotypes, transmission (*de novo*, paternal, maternal, and familial), and zygosity (homozygous, heterozygous, and mosaic). For the evidence of "Genetic association", "Expression", and "Linkage", besides the associated CHD types and syndromes, statistical methods and results are also integrated. For further functional information about the gene or variation, the user can click on the label "Gene Annotation" or "Variation Annotation" in the top left to view annotations from external databases and the data mentioned above (Figures S2 and S4).

Finally, to display a comprehensive genetic landscape of CHD, the database also shows the statistics at disease, gene, variation, and study levels, in charts or forms with embedded hyperlinks (Figures 1 and 3E). A Help Page with detailed introductions to the database is also provided (Figures 1 and 3F).

Features of CHD-related genes

To clarify the molecular mechanisms of CHD and facilitate the prioritization of novel susceptibility genes, we systematically characterized the features of CHD-related genes. As syndromic CHD is usually accompanied by abnormal brain development, we explored the expression profiles of three CHD gene categories in the brain and heart, across different time points of human development [28]. Compared with a random

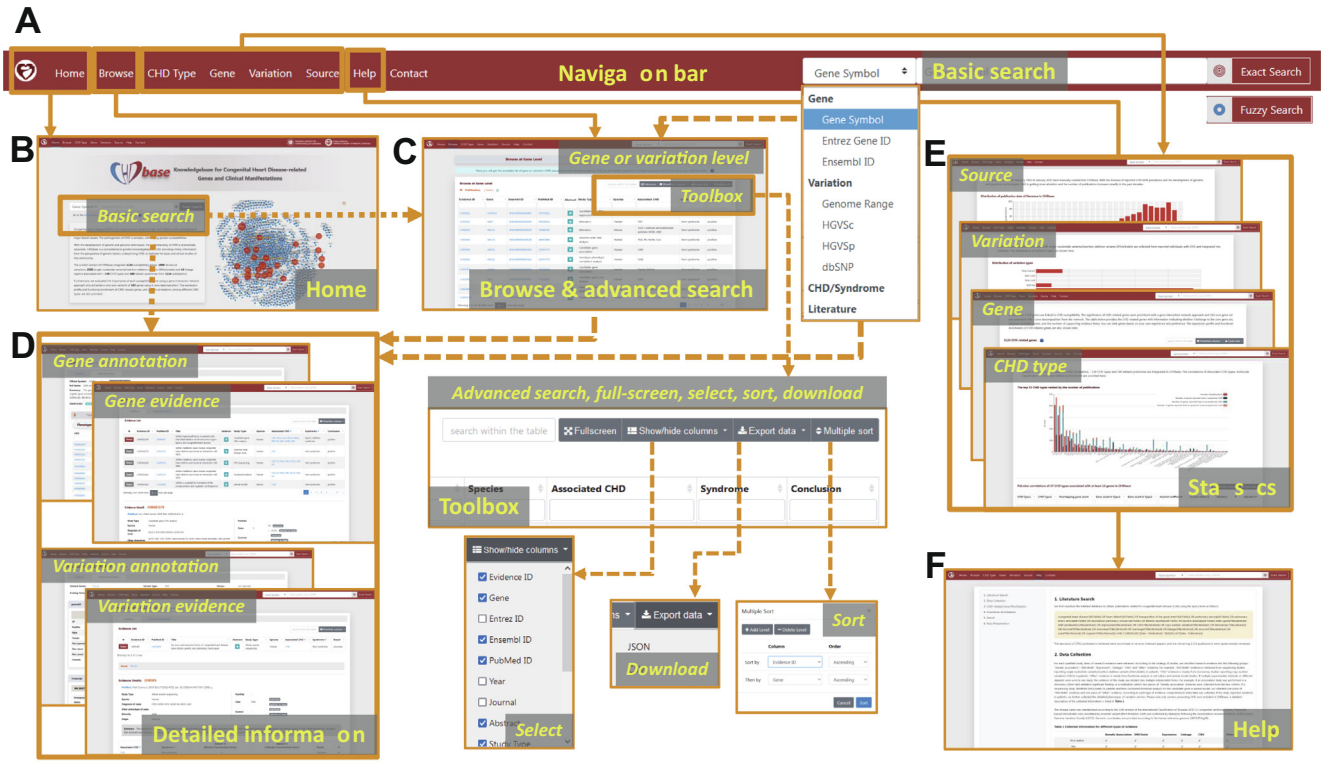


Figure 3 The web interface of CHDbase

The major components of the web interface of CHDbase are shown. **A.** The navigation bar. **B.** The Home Page. **C.** The Browse Page. **D.** The Evidence and Annotation Page for each CHD-related gene or variation. **E.** The Statistics Page for the diseases, genes, variations, and studies. **F.** The Help Page.

gene set unrelated to CHD as a background, all of the three CHD categories showed higher expression levels in the brain and heart, especially in fetal organs, indicating their roles in embryonic development. Notably, at varied developmental stages, the syndromic genes showed higher expression levels than nonsyndromic genes in the brain ($P = 0.13$, one-sided Mann–Whitney U test) (**Figure 4**A); however, in the heart, nonsyndromic genes expressed at a significantly higher level than syndromic genes ($P < 2.20E{-}16$, one-sided Mann–Whitney U test), suggesting a more specific role of nonsyndromic genes in heart development (Figure 4B). Furthermore, CHD-related genes tend to broadly express across tissues and developmental stages compared with the background, according to tissue- and time-specificity indexes as defined by Cardoso-Moreira et al., indicating CHD-related genes are more intolerant to functional variations [28] (Figure 4C and D).

We then identified the enriched GO terms and pathways for CHD-related genes with ClueGO [42] (**Figure 5**; Table S3). Briefly, the GO terms of heart development, embryonic morphogenesis, pattern specification process, cell differentiation, muscle structure development, regulation of cellular component movement, and cell surface receptor signaling pathway are significantly enriched. Consistently, on the pathway level, the cell surface receptor signaling pathways involved in heart development, such as the signaling pathways of wingless-type MMTV integration site family (Wnt), Notch, and transforming growth factor -beta (TGF-beta), are significantly enriched. The enrichment of CHD-related genes in terms of heart development is expected since the initial keywords for retrieving CHD-related genes were chosen to find these genes. While the enriched terms and pathways identified here provide a comprehensive molecular picture to understand CHD. Moreover, several cancer-related signaling pathways are also identified, providing a perspective to understand the previous epidemiological data that CHD patients typically show higher cancer prevalence [43,44]. Interestingly, the generic transcription pathway is also enriched, with 112 of the 1124 CHD-related genes cataloged as transcription factors in hTFtarget [45] ($P < 2.20E{-}16$, Fisher's exact test). Notably, among the top ten genes with the strongest supporting evidence, six are transcription factor genes (*GATA4*, *NKX2-5*, *TBX5*, *GATA6*, *CHD7*, and *NOTCH1*). In addition, our data also recapitulated several pathways previously linked to CHD etiology [18], such as muscle contraction, extracellular matrix organization, and cilium assembly.

Molecular classification of CHD

CHD is a highly heterogeneous disease with a large set of cardiac malformations. Adequately clustering these malformations could classify this complex disease into several groups with shared pathophysiological mechanisms, facilitating in-depth mechanism study and clinical management. As CHDbase integrates comprehensive phenotype–genotype correlation data, we then attempted to classify CHD based on the genetic background (File S1). We first used Jaccard coefficient [26] to measure pairwise genetic similarity between all

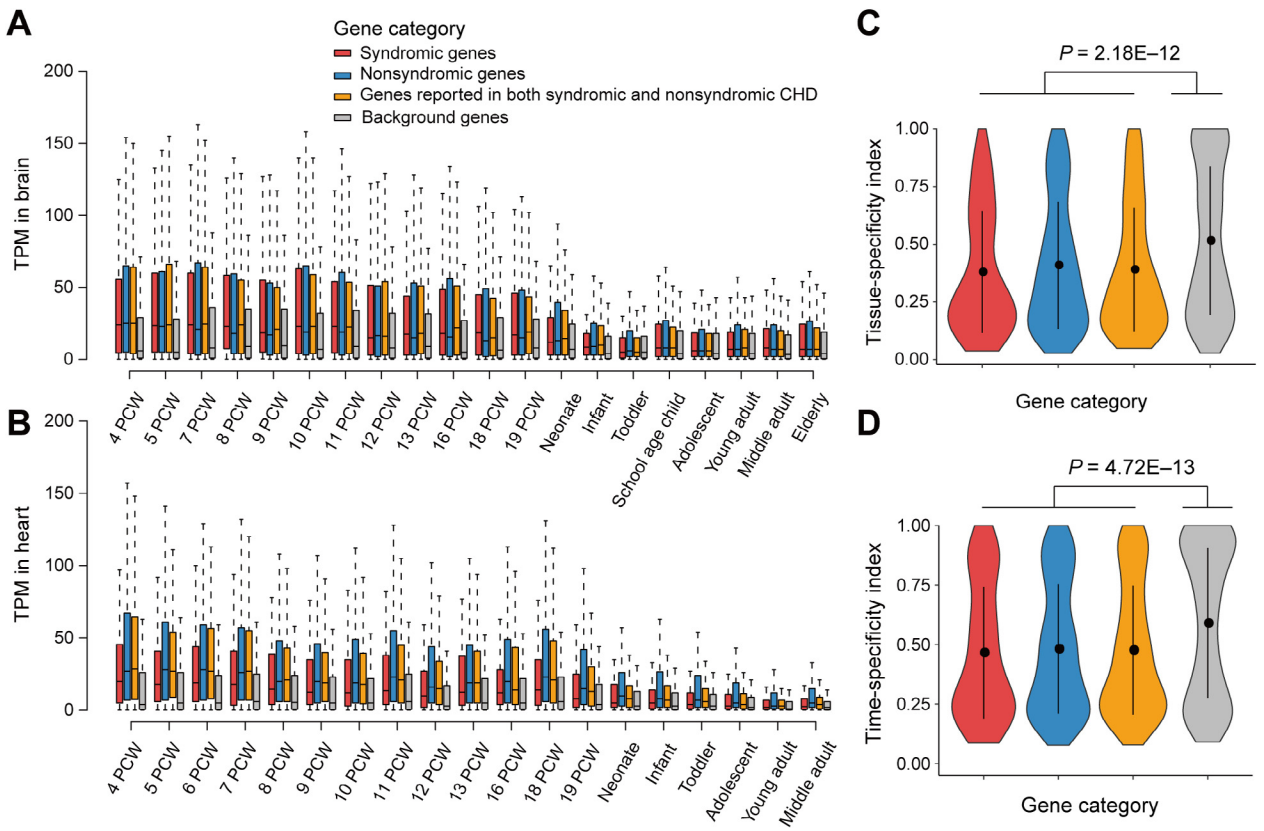


Figure 4 The expression profiles of CHD-related genes

A. and **B.** Boxplots showing the expression levels for three categories of CHD-related genes in the brain (A) and heart (B) across different developmental time points. **C.** and **D.** Violin plot showing the difference between CHD-related genes and background genes in tissue-specificity (C) and time-specificity (D). The *P* values were obtained with Mann–Whitney U test. In all panels, CHD-related genes are classified into three categories: syndromic genes (red, $n = 115$), nonsyndromic genes (blue, $n = 597$), and genes reported in both syndromic and nonsyndromic CHD (orange, $n = 414$). A non-CHD gene set ($n = 1124$) was randomly selected as the background (gray). TPM, transcripts per million; PCW, post-conception week.

pairs of 27 CHD types. In total, 196 pairs of CHD types showed significantly more shared genes than expected (adjusted $P < 0.05$) (**Figure 6**A; Table S4). Based on the matrix of pairwise Jaccard distances, hierarchical clustering was further performed to reveal genetic patterns shared by CHD types. Seven distinct groups were identified (Figure 6B), which are largely consistent with known developmental processes and previous CHD classifications, such as Botto's taxonomy [7] and Ellesøe's classification based on familial co-occurrence [9]. To control for the effects of false positives in identifying disease-causing variants, we repeated the classification after excluding 21 variants that were reclassified as "benign" or "likely benign" by InterVar [46] according to the American College of Medical Genetics and Genomics/Association for Molecular Pathology (ACMG/AMP) guidelines [47]. The new results are consistent, indicating the reliability of this classification.

As depicted in Figure 6B, cardiac malformations of abnormal rotation and septation of the outflow tract, such as transposition of the great arteries (TGA), double-outlet right ventricle (DORV), persistent truncus arteriosus (PTA), interrupted aortic arch (IAA), and right aortic arch (RAA), were grouped into "Outflow tract malformations", which represents the spectrum of conotruncal defects. This group highly correlates with "Atrioventricular canal and septal abnormality" because it is usually accompanied by septal abnormalities. In addition to "Outflow tract malformations", ventricular outflow tract obstruction (VOTO) clustered strongly and was divided into the left (LVOTO) and right (RVOTO) groups. VOTO refers to any anatomic or functional obstruction of flow out of the ventricle into the aorta or the pulmonary artery, usually caused by valvular stenosis or atresia. Our LVOTO and RVOTO groups largely overlap with Botto's and Ellesøe's classifications. Exceptions are pulmonary atresia (PA) and coarctation of the aorta (CoA). In our classification, PA and CoA share the most genes with dextrocardia (Jaccard coefficient = 0.20, adjusted $P = 0$) and vena cava abnormalities (Jaccard coefficient = 0.18, adjusted $P = 0$), respectively (Figure 6A; Table S4); both classified as "left–right abnormality". As previously reported, PA manifests in 23.2% of patients with heterotaxy, an abnormality of formation of the left–right axis of the body [48], and CoA has been associated with an isolated persistent left superior vena cava, a common vena cava abnormality in 21.3% of patients [49]. These findings support the correlation between the development of these two cardiac malformations and abnormal left–right

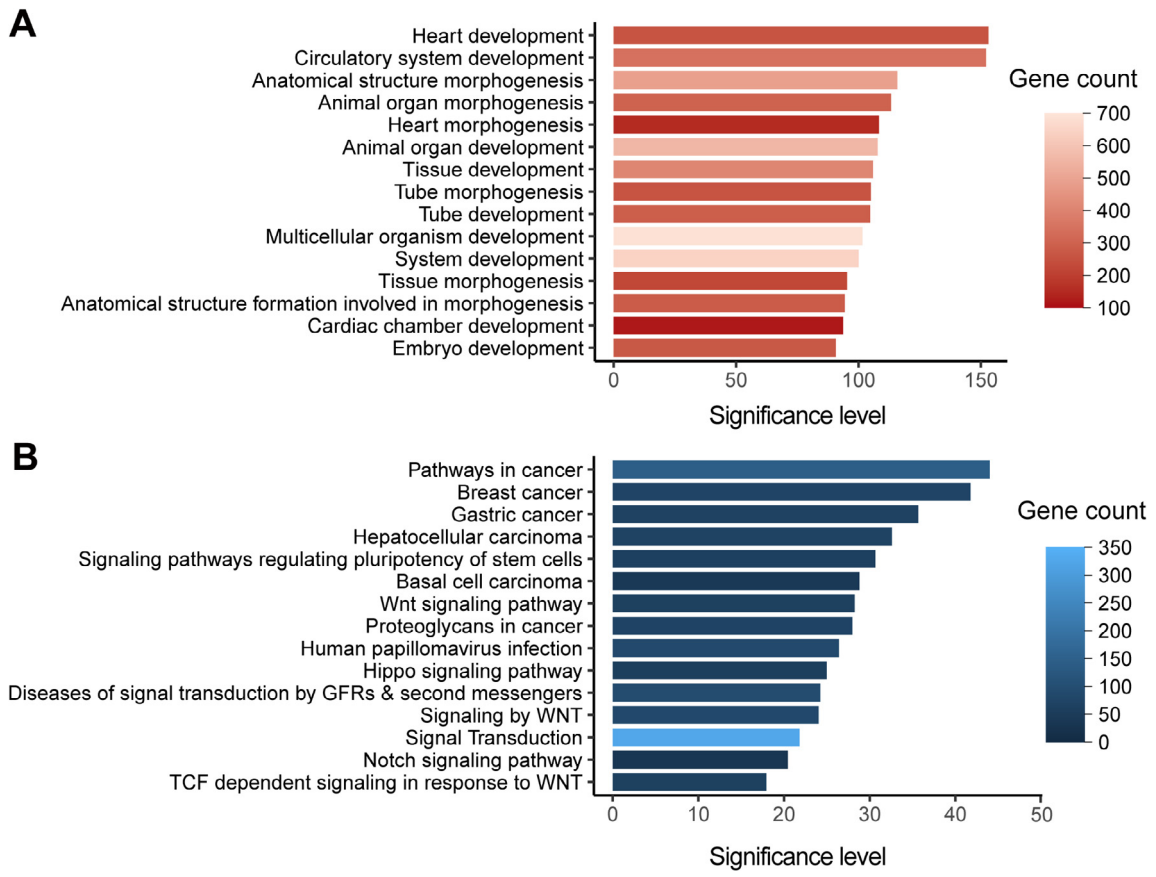


Figure 5 Functional enrichments of CHD-related genes

A. The top 15 significantly enriched GO terms for CHD-related genes (n = 1124) are shown with significance levels. **B.** The top 15 significantly enriched KEGG and Reactome pathways for CHD-related genes (n = 1124) are shown with significance levels. The adjusted P values corrected with Bonferroni step down were $-\log_{10}$ transformed to denote the significance levels. GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; TCF, T-cell factor; Wnt, wingless-type MMTV integration site family; GFR, growth factor receptor.

arrangement. The other four CHD types in our "left–right anomaly" group have been recognized to be related to the left–right arrangement [7,9]. Taken together, these major groups we defined are highly related to anatomy, embryology, and especially genetic origin. This molecular classification would thus serve as a supplement to the traditional CHD classifications and clinical practice.

Comparison with existing databases

To evaluate the utility and uniqueness of CHDbase, we compared it with CHD-RF-KB [10], a recently published database to compile risk factors associated with nonsyndromic CHD but not syndromic CHD, and two widely-used variant databases, ClinVar [50] and the Human Gene Mutation Database (HGMD) [51], that focus on a wide range of inherited diseases. As shown in **Table 1**, CHDbase compiles a larger set of CHD-related genes and variations, from a wider scope and a larger number of publications, compared with the other three databases. In comparison to the three databases which only reported information at the variation level, CHDbase provides 60 types of metadata, such as the study design, major associations, and sample information (Table S1). Moreover, CHDbase provides comprehensive functional annotations and analyses of the genes, showing a comprehensive landscape of CHD susceptibilities and facilitating a "one-stop" solution to clarify the gene–CHD associations.

Discussion

The genetic heterogeneity of CHD has been increasingly recognized in recent decades. However, several fundamental questions, such as the number of genes associated with CHD, and the reliability and reproducibility of each association, are still to be addressed, hindering the applications of current knowledge to etiological studies and clinical practice. In this study, we developed an evidence-based, manually curated knowledgebase for CHD, the CHDbase, linking 1124 susceptibility genes and 3591 variations with ~ 310 CHD types and related syndromes. We also propose a new network-based approach to prioritize the contribution of each gene to CHD pathogenesis. The users could thus simply select their candidate genes for in-depth mechanism study through our user-friendly data retrieval interfaces. CHDbase thus provides a "one-stop" solution from a comprehensive landscape of CHD susceptibilities to promising candidate genes or even drug targets.

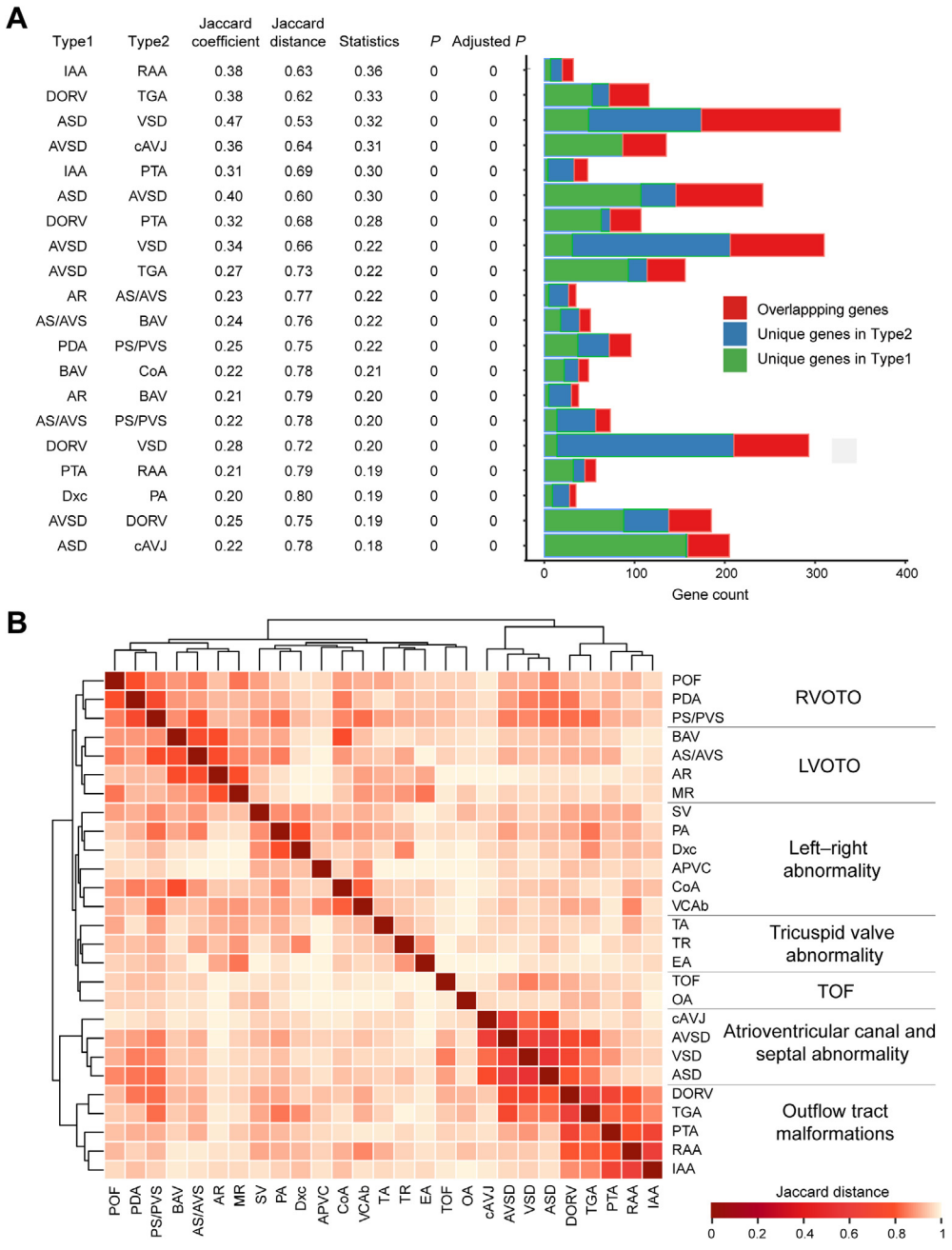


Figure 6 CHD classification based on genotype–phenotype correlations

A. Jaccard coefficient, Jaccard distance, statistical significance for Jaccard coefficient, and gene distribution are shown for 20 pairs of CHD types with the strongest correlations. The centered Jaccard coefficient equals to the Jaccard coefficient minus the unbiased estimation of expectation. *P* and Adjusted *P* were calculated with the Jaccard test and corrected for multiple testing using the Benjamini–Hochberg false discovery rate. **B.** Classification of 27 CHD types in CHDbase. Only CHD types associated with at least ten genes were included in this analysis. Jaccard distances for all pairs of CHD types are shown in a heatmap, which were further used to classify these CHD types into seven major groups through hierarchical cluster analysis. For each group, the anatomical term is defined and shown. POF, patent oval foramen; PDA, patent arterial duct; PS/PVS, pulmonary stenosis/pulmonary valve stenosis; BAV, bicuspid aortic valve; AS/AVS, aortic stenosis/aortic valvar stenosis; AR, aortic regurgitation; MR, mitral regurgitation; SV, single ventricle; PA, pulmonary atresia; Dxc, dextrocardia; APVC, anomalous pulmonary venous connection; CoA, coarctation of the aorta; VCAb, vena cava abnormality; TA, tricuspid atresia; TR, tricuspid regurgitation; EA, Ebstein malformation of the tricuspid valve; TOF, tetralogy of Fallot; OA, overriding aorta; cAVJ, common atrioventricular junction; AVSD, atrioventricular septal defect; VSD, ventricular septal defect; ASD, atrial septal defect; DORV, double-outlet right ventricle; TGA, transposition of the great arteries; PTA, persistent truncus arteriosus; RAA, right aortic arch; IAA, interrupted aortic arch; RVOTO, right ventricular outflow tract obstruction; LVOTO, left ventricular outflow tract obstruction.

Table 1 Comparison with existing databases

	CHDbase	CHD-RF-KB (up to 10 January 2020)	ClinVar (up to 10 January 2020)	HGMD (up to 10 January 2020)
Data source	Manually curated	Manually curated	User submitted	Manually curated
Target disease	Syndromic CHD, nonsyndromic CHD	Nonsyndromic CHD	A wide range of diseases	A wide range of diseases
No. of publications	1114	305	163	445
Evidence type	Genetic association, SNV/Indel, Expression, CNV, Linkage, and Other	Genetic association, SNV/Indel, CNV, and Methylation	Genetic association, SNV/Indel, and CNV	Genetic association and SNV/Indel
No. of genes	1145 (1124 CHD susceptibility genes and 21 negative genes)	303	53	874
No. of variations	2585 SNVs/Indels, 1006 CNVs	1194 SNVs/Indels, 567 CNVs	760 SNVs/Indels, 14 CNVs	2152 SNVs/Indels
Collected information	Publication, population, study design, result, and sample information	Publication, result, and sample information	Publication and result information	Publication and result information
Annotation	Gene annotation and variation annotation	Variation annotation	Variation annotation	None
Data analysis	Expression profile, functional enrichment, and CHD classification	Functional enrichment	None	None

Note: CHD, congenital heart disease; SNV, single nucleotide variant; CNV, copy number variation; Indels, insertion-deletion variants; HGMD, the Human Gene Mutation Database.

Furthermore, we demonstrated the utility of CHDbase through pattern analyses of genetic and phenotypic diversity in CHD. The expression and functional features of CHD-related genes were systemically analyzed to describe the holistic picture of CHD genetic factors, which should facilitate the study of CHD pathogenesis and the identification of novel susceptibility genes. Based on the most comprehensive genotype–phenotype correlations integrated into CHDbase, we further identified CHD groups with similar genetic origins. Although the classification is based on the genetic background, it recapitulates the traditional classification in that similar malformations could be detected in the same cluster, providing new insight into the genetic backgrounds underlying the pathogenesis and recurrence patterns of CHD. Notably, for this approach of classification, the integrity and reliability of genotype–phenotype correlations are crucial. Currently, although hundreds of genetic loci involved in CHD have been reported and integrated, the list of genes associated with CHD is still not complete and conclusive. Thus, the classification in this study is limited to the 27 most frequent CHD types. The molecular classification of CHD could thus be improved along with more accurate gene–CHD associations identified in the future.

Taken together, CHDbase can not only improve research on the etiology and pathogenesis of CHD but also aid in clinical practice for CHD, including diagnosis, genetic counseling, and treatment. The database will routinely be updated to keep pace with the research progress of CHD and built into a "one-stop" knowledgebase for exploring CHD susceptibility factors.

Data availability

CHDbase is publicly available at http://chddb.fwgenetics.org.

Competing interests

The authors have declared no competing interests.

CRediT authorship contribution statement

Wei-Zhen Zhou: Conceptualization, Data curation, Formal analysis, Methodology, Visualization, Writing - original draft, Writing - review & editing, Project administration, Funding acquisition. **Wenke Li:** Software, Methodology, Visualization. **Huayan Shen:** Data curation, Supervision, Writing - original draft. **Ruby W. Wang:** Formal analysis, Methodology, Writing - original draft. **Wen Chen:** Conceptualization, Data curation, Supervision. **Yujing Zhang:** Data curation, Visualization. **Qingyi Zeng:** Data curation. **Hao Wang:** Data curation. **Meng Yuan:** Data curation. **Ziyi Zeng:** Data curation. **Jinhui Cui:** Data curation. **Chuan-Yun Li:** Resources, Writing - review & editing. **Fred Y. Ye:** Supervision, Methodology, Writing - original draft. **Zhou Zhou:** Conceptualization, Project administration, Funding acquisition. All authors have read and approved the final manuscript.

Acknowledgments

This work was supported by the Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 31801103) and the Chinese Academy of Medical Sciences Initiative for Innovative Medicine (Grant No. 2016-I2M-1-016). We acknowledge Dr. Adam Yongxin Ye at Harvard Medical School, Boston Children's Hospital for his support and contribution to the project, and Dr. Ge Gao at Peking University for assistance with database comparison.

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2022.08.001.

ORCID

ORCID 0000-0002-3085-1581 (Wei-Zhen Zhou)
ORCID 0000-0003-0984-9468 (Wenke Li)
ORCID 0000-0003-3749-4991 (Huayan Shen)
ORCID 0000-0001-6544-3660 (Ruby W. Wang)
ORCID 0000-0001-6658-1926 (Wen Chen)
ORCID 0000-0002-1147-8953 (Yujing Zhang)
ORCID 0000-0002-2607-306X (Qingyi Zeng)
ORCID 0000-0002-9883-3435 (Hao Wang)
ORCID 0000-0002-1158-0595 (Meng Yuan)
ORCID 0000-0002-3049-5636 (Ziyi Zeng)
ORCID 0000-0002-5018-8641 (Jinhui Cui)
ORCID 0000-0002-8008-8430 (Chuan-Yun Li)
ORCID 0000-0001-9426-934X (Fred Y. Ye)
ORCID 0000-0002-7674-9857 (Zhou Zhou)

References

[1] Bernier PL, Stefanescu A, Samoukovic G, Tchervenkov CI. The challenge of congenital heart disease worldwide: epidemiologic and demographic facts. Semin Thorac Cardiovasc Surg Pediatr Card Surg Annu 2010;13:26–34.
[2] van der Linde D, Konings EE, Slager MA, Witsenburg M, Helbing WA, Takkenberg JJM, et al. Birth prevalence of congenital heart disease worldwide: a systematic review and meta-analysis. J Am Coll Cardiol 2011;58:2241–7.
[3] Oyen N, Poulsen G, Boyd HA, Wohlfahrt J, Jensen PK, Melbye M. Recurrence of congenital heart defects in families. Circulation 2009;120:295–301.
[4] Nogee JM, Jay PY. The heritable basis of congenital heart disease: past, present, and future. Circ Cardiovasc Genet 2016;9:315–7.
[5] Pierpont ME, Brueckner M, Chung WK, Garg V, Lacro RV, McGuire AL, et al. Genetic basis for congenital heart disease: revisited: a scientific statement from the American heart association. Circulation 2018;138:e653–711.
[6] Jin SC, Homsy J, Zaidi S, Lu Q, Morton S, DePalma SR, et al. Contribution of rare inherited and *de novo* variants in 2871 congenital heart disease probands. Nat Genet 2017;49:1593–601.
[7] Botto LD, Lin AE, Riehle-Colarusso T, Malik S, Correa A. National Birth Defects Prevention Study. Seeking causes: classifying and evaluating congenital heart defects in etiologic studies. Birth Defects Res A Clin Mol Teratol 2007;79:714–27.
[8] Houyel L, Khoshnood B, Anderson RH, Lelong N, Thieulin AC, Goffinet F, et al. Population-based evaluation of a suggested anatomic and clinical classification of congenital heart defects based on the International Paediatric and Congenital Cardiac Code. Orphanet J Rare Dis 2011;6:64.
[9] Ellesoe SG, Workman CT, Bouvagnet P, Loffredo CA, McBride KL, Hinton RB, et al. Familial co-occurrence of congenital heart defects follows distinct patterns. Eur Heart J 2018;39:1015–22.
[10] Yang L, Liu X, Chen Y, Shen B. An update on the CHDGKB for the systematic understanding of risk factors associated with non-syndromic congenital heart disease. Comput Struct Biotechnol J 2021;19:5741–51.
[11] Franklin RCG, Beland MJ, Colan SD, Walters HL, Aiello VD, Anderson RH, et al. Nomenclature for congenital and paediatric cardiac disease: the International Paediatric and Congenital Cardiac Code (IPCCC) and the Eleventh Iteration of the International Classification of Diseases (ICD-11). Cardiol Young 2017;27:1872–938.
[12] McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The ensembl variant effect predictor. Genome Biol 2016;17:122.
[13] Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PEM. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. Hum Mutat 2008;29:6–13.
[14] Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 2019;47:D607–13.
[15] Fahed AC, Gelb BD, Seidman JG, Seidman CE. Genetics of congenital heart disease: the glass half empty. Circ Res 2013;112:707–20.
[16] Andersen TA, de K, Linde, Lind, Troelsen, Larsen LA. Of mice and men: molecular genetics of congenital heart disease. Cell Mol Life Sci 2014;71:1327–52.
[17] Nees SN, Chung WK. Genetic basis of human congenital heart disease. Cold Spring Harb Perspect Biol 2020;12 a036749.
[18] Williams K, Carson J, Lo C. Genetics of congenital heart disease. Biomolecules 2019;9:879.
[19] Sayers EW, Beck J, Brister JR, Bolton EE, Canese K, Comeau DC, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2020;48:D9–D.
[20] Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, et al. GeneCards Version 3: the human gene integrator. Database (Oxford) 2010;2010:baq020.
[21] Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM^{R}), an online catalog of human genes and genetic disorders. Nucleic Acids Res 2015;43:D789–98.
[22] Bult CJ, Blake JA, Smith CL, Kadin JA, Richardson JE, Mouse Genome Database Group. Mouse Genome Database (MGD) 2019. Nucleic Acids Res 2019;47:D801–6.
[23] Bradford Y, Conlin T, Dunn N, Fashena D, Frazer K, Howe DG, et al. ZFIN: enhancements and updates to the Zebrafish Model Organism Database. Nucleic Acids Res 2011;39:D822–9.
[24] Kohler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, Gourdine JP, et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Res 2019;47:D1018–27.
[25] Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 2009;106:9362–7.
[26] Pinero J, Ramirez-Anguita JM, Sauch-Pitarch J, Ronzano F, Centeno E, Sanz F, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res 2020;48:D845–55.
[27] GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet 2013;45:580–5.
[28] Cardoso-Moreira M, Halbert J, Valloton D, Velten B, Chen C, Shao Y, et al. Gene expression across mammalian organ development. Nature 2019;571:505–9.
[29] Asp M, Giacomello S, Larsson L, Wu C, Furth D, Qian X, et al. A spatiotemporal organ-wide gene expression and cell atlas of the developing human heart. Cell 2019;179:1647–60.
[30] Lee TY, Hsu JBK, Chang WC, Wang TY, Hsu PC, Huang HD. A comprehensive resource for integrating and displaying protein post-translational modifications. BMC Res Notes 2009;2:111.
[31] Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, et al. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res 2011;39:D691–7.
[32] Karp PD, Ouzounis CA, Moore-Kochlacs C, Goldovsky L, Kaipa P, Ahren D, et al. Expansion of the BioCyc collection of

pathway/genome databases to 160 genomes. Nucleic Acids Res 2005;33:6083–9.
[33] Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, et al. KEGG for linking genomes to life and the environment. Nucleic Acids Res 2008;36:D480–4.
[34] Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, et al. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 2003;13:2129–41.
[35] Turei D, Korcsmaros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nat Methods 2016;13:966–7.
[36] Oughtred R, Stark C, Breitkreutz B-J, Rust J, Boucher L, Chang C, et al. The BioGRID interaction database: 2019 update. Nucleic Acids Res 2019;47:D529–41.
[37] Luck K, Kim DK, Lambourne L, Spirohn K, Begg BE, Bian W, et al. A reference map of the human binary protein interactome. Nature 2020;580:402–8.
[38] Cotto KC, Wagner AH, Feng YY, Kiwala S, Coffman AC, Spies G, et al. DGIdb 3.0: a redesign and expansion of the drug-gene interaction database. Nucleic Acids Res 2018;46:D1068–73.
[39] Avram S, Bologa CG, Holmes J, Bocci G, Wilson TB, Nguyen DT, et al. DrugCentral 2021 supports drug discovery and repositioning. Nucleic Acids Res 2021;49:D1160–9.
[40] Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, et al. Pharmacogenomics knowledge for personalized medicine. Clin Pharmacol Ther 2012;92:414–7.
[41] Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 2020;581:434–43.
[42] Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics 2009;25:1091–3.
[43] Gurvitz M, Ionescu-Ittu R, Guo L, Eisenberg MJ, Abrahamowicz M, Pilote L, et al. Prevalence of cancer in adults with congenital heart disease compared with the general population. Am J Cardiol 2016;118:1742–50.
[44] Mandalenakis Z, Karazisi C, Skoglund K, Rosengren A, Lappas G, Eriksson P, et al. Risk of cancer among children and young adults with congenital heart disease compared with healthy controls. JAMA Netw Open 2019;2:e196762.
[45] Zhang Q, Liu W, Zhang HM, Xie GY, Miao YR, Xia M, et al. hTFtarget: a comprehensive database for regulations of human transcription factors and their targets. Genomics Proteomics Bioinformatics 2020;18:120–8.
[46] Li Q, Wang K. InterVar: clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines. Am J Hum Genet 2017;100:267–80.
[47] Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. 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;17:405–24.
[48] Kiran VS, Choudhary S, Shaik A, Gadabanahalli K, Raj V, Bhat V. The spectrum of cardiac anomalies associated with heterotaxy: a single-center study of a large series based on computed tomography. Pediatr Cardiol 2020;41:1414–24.
[49] Gustapane S, Leombroni M, Khalil A, Giacci F, Marrone L, Bascietto F, et al. Systematic review and meta-analysis of persistent left superior vena cava on prenatal ultrasound: associated anomalies, diagnostic accuracy and postnatal outcome. Ultrasound Obstet Gynecol 2016;48:701–8.
[50] Landrum MJ, Chitipiralla S, Brown GR, Chen C, Gu B, Hart J, et al. ClinVar: improvements to accessing data. Nucleic Acids Res 2020;48:D835–44.
[51] Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, et al. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet 2017;136:665–77.