Computational and Structural Biotechnology Journal 18 (2020) 381–392

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Identification of rare variants in novel candidate genes in pulmonary atresia patients by next generation sequencing

Xin Shi [a,1], Li Zhang [b,1], Kai Bai [a,1], Huilin Xie [a], Tieliu Shi [c], Ruilin Zhang [d], Qihua Fu [e], Sun Chen [a], Yanan Lu [a,*], Yu Yu [a,f,*], Kun Sun [a,*]

[a] Department of Pediatric Cardiovascular, Xinhua Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai 200092, China
[b] Key Laboratory of Advanced Theory and Application in Statistics and Data Science, East China Normal University, Ministry of Education, Shanghai, China
[c] The Center for Bioinformatics and Computational Biology, Shanghai Key Laboratory of Regulatory Biology, the Institute of Biomedical Sciences and School of Life Sciences, East China Normal University, Shanghai, China
[d] School of Life Sciences, Fudan University, Shanghai 200438, China
[e] Medical Laboratory, Shanghai Children's Medical Center, School of Medicine, Shanghai Jiao Tong University, Shanghai 200127, China
[f] Institute for Developmental and Regenerative Cardiovascular Medicine, Xinhua Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai 200092, China

ARTICLE INFO

Article history:
Received 18 July 2019
Received in revised form 10 January 2020
Accepted 29 January 2020
Available online 12 February 2020

Keywords:
Congenital heart defect
Pulmonary atresia
Whole-exome sequencing
Rare variants
Gene mutations

ABSTRACT

Pulmonary atresia (PA) is a rare congenital heart defect (CHD) with complex manifestations and a high mortality rate. Since the genetic determinants in the pathogenesis of PA remain elusive, a thorough identification of the genetic factors through whole exome sequencing (WES) will provide novel insights into underlying mechanisms of PA. We performed WES data from PA/VSD (n = 60), PA/IVS (n = 20), TOF/PA (n = 20) and 100 healthy controls. Rare variants and novel genes were identified using variant-based association and gene-based burden analysis. Then we explored the expression pattern of our candidate genes in endothelium cell lines, pulmonary artery tissues, and embryonic hearts. 56 rare damage variants of 7 novel candidate genes (*DNAH10, DST, FAT1, HMCN1, HNRNPC, TEP1,* and *TYK2*) were certified to have function in PA pathogenesis for the first time. In our research, the genetic pattern among PA/VSD, PA/IVS and TOF/PA were different to some degree. Taken together, our findings contribute new insights into the molecular basis of this rare congenital birth defect.

1. Introduction

Pulmonary atresia (PA) is a rare complex congenital heart defect (CHD), which occurs in 0.01% of live births and 1.3–3.4% of all heart malformations [1,2]. PA is a complex heart lesion with heterogeneity in embryogenesis, surgical strategy, prognosis, as well as anatomic variance. PA can be mainly classified into two categories according to whether there is a ventricular septal defect (VSD) or not: pulmonary atresia with ventricular septal defect (PA/VSD), pulmonary atresia with intact ventricular septum (PA/IVS) [3–5].

PA/VSD belongs to conotruncal defect (CTD) which is characterized as a group of congenital cardiac outflow tract anomalies [6]. PA/VSD shares many similar structural and pathological features with tetralogy of Fallot (TOF), and is considered as the most extreme form of TOF patients [7,8]. PA/IVS is a rare heart lesion which is quite another matter [9,10]. There are many differences in anatomical abnormalities between PA/VSD and PA/IVS. In PA/IVS, all components of the right ventricle (RV) can be affected since little blood flows into or out of RV, which leads to the undeveloped and hypoplastic inlet RV. Without early diagnosis and proper treatment, the survival rate of PA/IVS is extremely poor [11].

With the improvements in surgical performance and cure rate, the focus of attention on pulmonary atresia has shifted from

Abbreviations: PA, Pulmonary atresia; CHD, congenital heart defect; WES, whole exome sequencing; PA/VSD, Pulmonary atresia with ventricular septal defect; PA/IVS, Pulmonary atresia with intact ventricular septum; CTD, Conotruncal defect; TOF, tetralogy of Fallot; RV, right ventricle; ACMG, American College of Medical Genetics; FDR, False discovery rates; PPI, protein–protein interactions; STRING, Search Tool for the Retrieval of Interacting Genes; GEO, Gene Expression Omnibus; RT-qPCR, Reverse Transcription Quantitative PCR; HPAECs, Human Pulmonary Artery Endothelial Cells; MAF, minor allele frequency; ExAC, Exome Aggregation Consortium; gnomAD, Genome Aggregation Database; SNP, single nucleotide polymorphism; GSEA, gene set enrichment analysis; LOF, loss-of-function.

* Corresponding authors at: Department of Pediatric Cardiology, Xinhua Hospital, 1665 Kongjiang Road, Shanghai 200092, China.
E-mail addresses: luscmc@189.cn (Y. Lu), yuyu@xinhuamed.com.cn (Y. Yu), sunkun@xinhuamed.com.cn (K. Sun).
[1] Co-first authors.

https://doi.org/10.1016/j.csbj.2020.01.011

immediate outcomes to underlying mechanisms. A proportion of PA/VSD is reported to be associated with rare copy number variants, including deletions in 5q14.1dup (DHFR), 10p13dup (CUBN) and 17p13.2del (CAMTA2), and most of PA/VSD patients have 22q11 deletion syndrome [12,13]. Other than chromosomal rearrangements, previous literatures have implicated rare damaging mutations in GAJ5, GDF1, MTHFR, MYH6, and HEY gene family that were detected in some PA/VSD patients [14–18]. Previous studies have assumed that gene expression pattern in PA/VSD and PA/IVS should be different. But most of previous researches focused on PA/VSD, little is known about the prevalence and spectrum of genetic alterations in PA/IVS. He et al. conducted a genome-wide scanning in 54 PA/IVS patients and 20 PA/VSD patients and found that no rare CNVs were detected in patients with PA/IVS, which might be totally different from PA/VSD in terms of genetics as well as anatomy [19]. The exact differences of genetic pattern between PA/VSD and PA/IVS remain unknown, which requires further study to emphasize the marked heterogeneity and etiological complexity of these rare and complex CHDs.

Genetic factors play an important role in PA although no major pathogenic gene has been identified to date. Previously reported pathogenic genes cover only a small part of the molecular mechanism underlying PA pathogenesis, it still lacks enough evidence to evaluate the contribution of gene pattern in different subtypes of PA [20–22]. Therefore, we applied whole exome sequencing technology to identify the genes pattern in 100 PA cases and 100 healthy controls. Then we detected the expression pattern of candidate genes in HPAECs, human pulmonary artery, and human embryonic hearts. At last, we considered 7 genes (*DNAH10, DST, FAT1, HMCN1, HNRNPC, TEP1,* and *TYK2*) as the novel candidate genes that most likely underlie PA pathogenesis. Our findings demonstrate that rare damaging variants could lead to the occurrence of PA. Moreover, we identified novel candidate genes which have never been reported as related to the development of pulmonary artery. Our study is also an effort to distinguish the genetic pattern between PA/VSD, PA/IVS and TOF/PA. These results could fill the void of the underlying mechanism of PA, and assist in the identification of candidate genes for PA/VSD, PA/IVS and TOF/PA.

2. Materials and methods

2.1. Study population

Our cohort included sporadic cases with PA/VSD (n = 60), TOF/PA (n = 20), PA/IVS (n = 20) and 100 healthy controls diagnosed by echocardiogram, cardiac catheterization, or surgery. Patients with an identified chromosomal or syndromic disorder or a situs anomaly were excluded. The present study was approved by the Ethics Committee of the Xinhua Hospital Affiliated to the Shanghai Jiao Tong University School of Medicine and was conducted according to the principles expressed in the Declaration of Helsinki. Participants and/or their legal guardians involved in this study gave a written informed consent prior to inclusion in the study.

2.2. DNA extraction and whole exome sequencing (WES)

Genomic DNA samples of the PA cases and healthy controls were obtained with written informed consent. DNA extraction from blood samples was carried out using the QIAamp™ DNA and Blood Mini kit (Qiagen) according to the manufacturer's instructions. WES was performed using the Agilent Sure Select Target Enrichment kit (V6 58 Mb; Agilent Technologies) for sequence capture and the Illumina HiSeq2500 for sequencing (Illumina) to a target depth of 100×.

2.3. Reads mapping, variants calling and annotation

The whole exome sequencing reads of 300 bp (150 bp at each end) were mapped to UCSC human reference genome (GRCh37/hg19 assembly) using BWA-mem mode [23] with default options followed by removal of PCR duplicates, low-quality reads (BaseQ < 20). The resulting SAM files were then sorted and indexed by samtools [24]. The variant calling was performed by the Genome Analysis Toolkit (GATK) best-practices workflow [25]. The variants were annotated by ANNOVAR (20601685) with the databases of RefSeq gene, SIFT [26], PolyPhen [27], MutationTaster [28], 1000G [29], and ExAC and gnomAD [30]. As illustrated in Fig. 1, the rare damaging variants were finally screened using the American College of Medical Genetics (ACMG) criteria guidelines [31].

2.4. PA-associated genes and variants

The Fisher's exact test was conducted to test the association between variants and the disease status. The SNP status was encoded as 0 or 1, where 0 indicated that no SNP alleles were found, and 1 indicated that at least 1 SNP allele was detected. The SNP status and sample class labels where 1 indicated PA patients and 0 indicated control samples were used to obtain the 2×2 contingency table for Fisher's exact test, and P-value < 1e−4 was considered statistically significant. Furthermore, we aggregated the SNP data at the gene level. The samples were considered as mutated samples at gene level if they carried at least one rare mutation of the given gene. Similarly, the gene-burden-based association was determined by Fisher's exact test. The gene level P-values were adjusted with the false discovery rates (FDR) method, where FDR < 0.05 or P-value < 0.005 was considered statistically significant.

2.5. Protein–protein interaction (PPI) analysis

The protein–protein interactions were extracted from the Search Tool for the Retrieval of Interacting Genes (STRING) database (http://string-db.org) [32,33], which could critically assess and integrate protein–protein interactions (PPI), including both direct (physical) and indirect (functional) associations. The gene set consisting of known PA genes and the 20 statistically significant PA-associated genes were mapped to the PPI network together. With the removal of the proteins without connected nodes, the resulting PPI subnetwork constructed by the known and novel PA genes were visualized by Cytoscape [34].

2.6. Gene expression analysis of the microarray datasets

The gene expression datasets (GSE67492 [35] and GSE30428 [36]) of cardiomyocytes were collected from Gene Expression Omnibus (GEO) database, which normalized the gene expression data. In addition, our previously published time course expression data of human embryonic heart were measured using Affymetrix HTA 2.0 microarray platform [37]. The median of expression levels across the samples was used to represent the expression level for each gene. The rankings of the PA-associated genes in the expression profiles were visualized by R/Bioconductor fgsea package.

2.7. RNA extraction and RT-qPCR assay

Total RNA was extracted from Human Pulmonary Artery Endothelial Cells (HPAECs), pulmonary artery and embryonic heart from humans, respectively. HPAECs were obtained from the Chinese Academy of Sciences (Shanghai, China). The human tissue samples consisted of pulmonary artery from 3 healthy controls

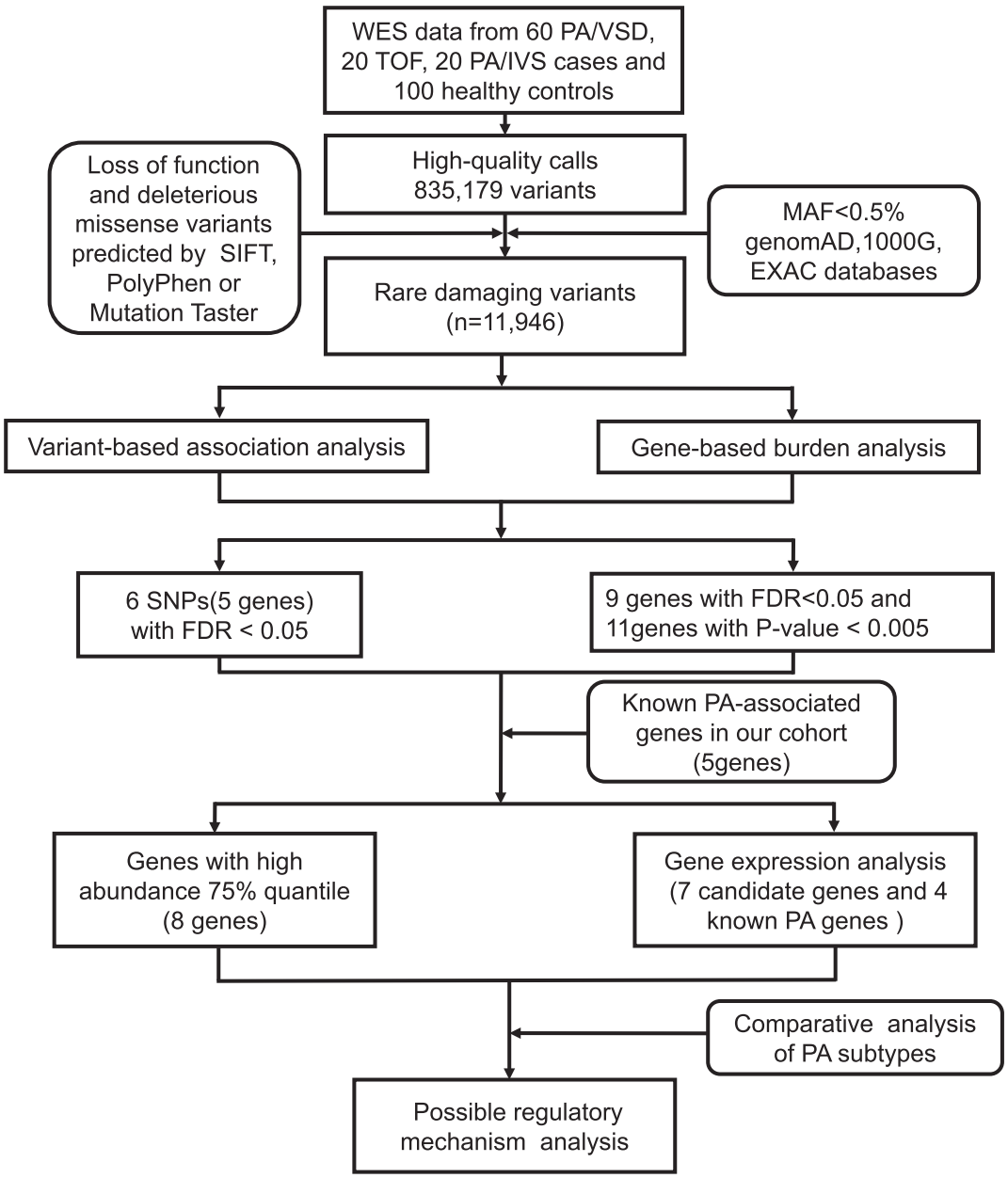

Fig. 1. The schematic diagram of systematic data analysis of the PA-associated genes/variants. An analytical strategy workflow of the different steps taken during whole-exome sequencing analysis with gene expression detection is shown. After variant calling and annotation, variants were then filtered by variant-based association analysis, gene-based burden analysis. Candidate genes were collected by gene expression analysis and network analysis. WES, whole-exome sequencing; PA/VSD, pulmonary atresia with ventricular septal defect; PA/IVS, pulmonary atresia with intact ventricular septum; TOF, tetralogy of Fallot; SNP, single nucleotide polymorphism; MAF, minor allele frequency; FDR, false discovery rate.

and Carnegie stages 11 through 15 human embryonic heart harvested from Shanghai Xinhua Hospital. The integrity and purity of the RNA was detected by the Experion automated gel electrophoresis system (Bio-Rad) and the NanoDrop 2000c spectrophotometer (ThermoFisher Scientific) [38]. RNA extraction and the RT-qPCR procedure were described previously. The RT-qPCR primers are listed in Supplementary Table S1.

2.8. PA-subtype associated genes

The genes that were only present in one of PA/VSD, TOF/PA, or PA/IVS and absent in controls were considered as potential PA-subtype associated genes. For each gene, the samples with variants should be more than 2. To detect potential relationships among these subtype candidate genes, we mapped all these subtype genes and known PA genes to the STRING network and visualized the network using Cytoscape. Moreover, we compared the number of total variants, nonsynonymous SNVs, and LOF variants among these three subtypes.

3. Results

3.1. Overview of pathogenic gene discovery based on WES data analysis

The present cohort of 100 unrelated patients with PA were collected from Shanghai Xinhua Hospital, which consisted of 60 PA/VSD patients, 20 PA/IVS patients, and 20 TOF/PA patients (Table 1). In addition, 100 healthy people were collected as normal controls. The whole exome sequencing was conducted on all 200 samples.

Table 1
Summary of demographic and clinical information of 100 study patients.

Patient characteristics	PA/VSD	PA/IVS	TOF
Mean age at diagnosis (years)	4.4 ± 3.55	0.84 ± 0.43	1.66 ± 2.6
Female, n, (%)	21(35)	10(50)	8(40)
Male, *n*, (%)	39(65)	10(50)	12(60)
Total, n	60	20	20
Female-to-male ratio	1:1.9	1:1	1:1.5
Associated cardiac lesion, n, (%)			
Left superior vena cava (LSVC)	1(5)	0	2(10)
Right aortic arch (RAA)	1(1.7)	2(10)	3(15)
Total anomalous pulmonary vein connection (TAPVC)	0	1(5)	0
Aortic stenosis (AS)	0	1(5)	1(5)

As illustrated in Fig. 1, the analysis of whole exome sequencing (WES) data from the 100 PA cases and 100 healthy controls successfully called 835,179 variants with high confidence. Following ACMG criteria guidelines, we screened a total of 11,946 rare damaging variants with a threshold of minor allele frequency (MAF) at 0.5% in databases of gnomAD, 1000G, and ExAC [30], and the deleteriousness was predicted by bioinformatics tools (see Materials and methods). The downstream analyses including variant-based association analysis, gene-based burden analysis, known PA genes/variants, and comparative analysis of PA subtypes, were conducted based those variants.

3.2. The overall comparison of the rare damaging variants between the PA and control groups

Based on the 11,946 rare damaging variants, we found more variants in PA group than that in control (Wilcoxon rank-sum test, $P < 2.2e{-}16$), consistently, which was also observed in missense variant, nonsense variant, frameshift insertion or deletion, and splice variant (Fig. 2A). Moreover, the SNP accounted for most the variant types in both case and control groups (Fig. 2B). C < T mutation occupied the most of base mutation than other types (Fig. 2C). The ratios between transitions and transversions in PA and control groups were about 2.00 and 1.98, which was close to 2 of theoretical ratio.

3.3. Identification of PA-associated variants by Fisher's exact test

To identify the PA-associated variants, we conducted Fisher's exact test on the rare damaging variants as described above. We observed six variants that were more frequently detected in PA group than in control group (FDR < 0.05, Fig. 2D). The six variants were located within *SLC9B1*, *MUC6*, *HNRNPC*, *OR11H12*, and *ATXN3*. Particularly, two variants fell from the same gene of *HNRNPC*, which showed a predominant significance (*P*-value < 1e−15). Moreover, the variant in *SLC9B1* also exhibited a significance level similar to the variants in *HNRNPC*, suggesting that these two genes may play key roles in PA pathogenesis.

3.4. Identification of PA-associated genes by gene-based burden test

To identify PA-associated genes, we first aggregated the variants at the gene level, and then performed the Fisher's exact test for each gene. Given loose thresholds of 0.05 for *P*-value and 3 PA cases with more than one variant, we identified 178 genes with potential pathogenicity (Fig. 3A), of which, *ATXN3*, *TTN*, *HNRNPC*, *SLC9B1*, *OBSCN*, *MUC4*, *FAT1*, *OR11H12*, and *DST* were considered as the ones with high confidence (Fig. 3B, FDR <0.05). In accordance with the PA-associated variants, *HNRNPC*, *SLC9B1*, *OR11H12*, and *ATXN3* were also identified by the gene-based burden test. In addition, we also identified another eleven genes with relatively low confidence (*P*-value <0.005). Totally, 20 genes were identified as PA-associated.

As complex genetic disorders may have multiple pathogenic genes, these genes were generally co-occurred in the same sample, or exclusively occurred in different samples. Specifically, *HNRNPC* was exclusive of *OR11H12* and *SLC9B1* (Fig. 3C, $P < 0.1$), respectively. In contrast, other 8 gene pairs, such as *SDK1-MUC4*, *DNAH1-DNAH10*, *DNAH1-FAT1*, and *SDK1-PCDHGA8*, were observed to be co-occurred in PA samples ($P < 0.1$). These results suggested that gene pairs with exclusivity or co-occurrence might be involved in the same biological process or cooperated with each other to cause the disease.

To further investigate the roles of the PA-associated genes by gene-based burden test, we mapped the PA-associated genes and 45 manually curated known PA genes to the PPI network (Supplementary Table S2). We observed that *HNRNPC*, *TEP1*, *TTN*, *NEB*, *OBSCN*, *SYNE2*, and *PKD1* were directly or indirectly connected with the subnetwork of the known PA genes (Fig. 3D). In addition, two gene pairs, *DNAH1-DNAH10* and *FRAS1-HMCN1*, related to each other in the PPI network. Particularly, *DNAH1-DNAH10* was also observed to be co-occurred in PA samples, which suggested that interaction between *DNAH1* and *DNAH10* may play key roles in the occurrence of PA.

3.5. Comprehensive analysis of the expression levels of the PA-associated genes

To further investigate the potential function of the PA-associated genes/variants in heart or cardiomyocyte, we collected two publicly available gene expression datasets (GSE67492 and GSE30428) of heart right ventricular from Gene Expression Omnibus (GEO). We observed that 20 PA-associated genes were significantly enriched in the genes with high expression in these two datasets ($P < 0.05$, Fig. 4A and 4B) based on the gene set enrichment analysis (GSEA). *TTN*, *DST*, *HNRNPC*, *SYNE2*, and *OBSCN* were highly expressed at top 15% in human heart tissues of both datasets. Moreover, we also investigated the expression levels of these genes in human embryonic hearts, which was published by our previous study [37], and found that *TTN*, *HNRNPC*, *FAT1*, *FRAS1*, and *DST* expressed higher than 75% quantile of the total genes (Fig. 4C). The integrative analysis of the gene expression and PPI network revealed that *TTN*, *HNRNPC*, *SYNE2*, and *OBSCN* may be the critical genes for PA occurrence. Notably, *HNRNPC* was the most significant PA-associated gene/variant.

To further filter the novel candidate genes of PA, we used RT-qPCR assay to detect the mRNA levels of the 21 candidate genes and 5 known PA genes in HPAECs, human pulmonary artery tissues harvested from 3 controls. Then, we further filtered these genes according to whether they were highly expressed in different tissues. By assessing gene expression in different samples, we finally obtained 7 candidate genes and 56 rare damage-associated non-synonymous variants in our WES data (Table 2). Finally, we got 7 candidate genes (*DNAH10, DST, FAT1, HMCN1, HNRNPC, TEP1,* and *TYK2*) and 4 known PA genes (*DOCK6, FANCD2, FGD5,* and *NOTCH1*) both expressed in cell line and human tissue (Fig. 4D).

3.6. Mutational spectrum of known PA-associated genes in our cohort

In addition to the novel PA-associated genes/variants, the presence and mutational spectrum of known PA-associated genes in our cohort was also investigated. With a threshold of *P*-value at 0.05, *NOTCH1*, *DOCK6*, *FANCD2*, *FGD5*, and *FLT4* were identified as significant known PA-associated genes (Fig. 5A). The mutational spectrum of these genes showed that the missense variants were mostly located within the functional domains of the corresponding proteins (Fig. 5B). Particularly, all the four missense variants of

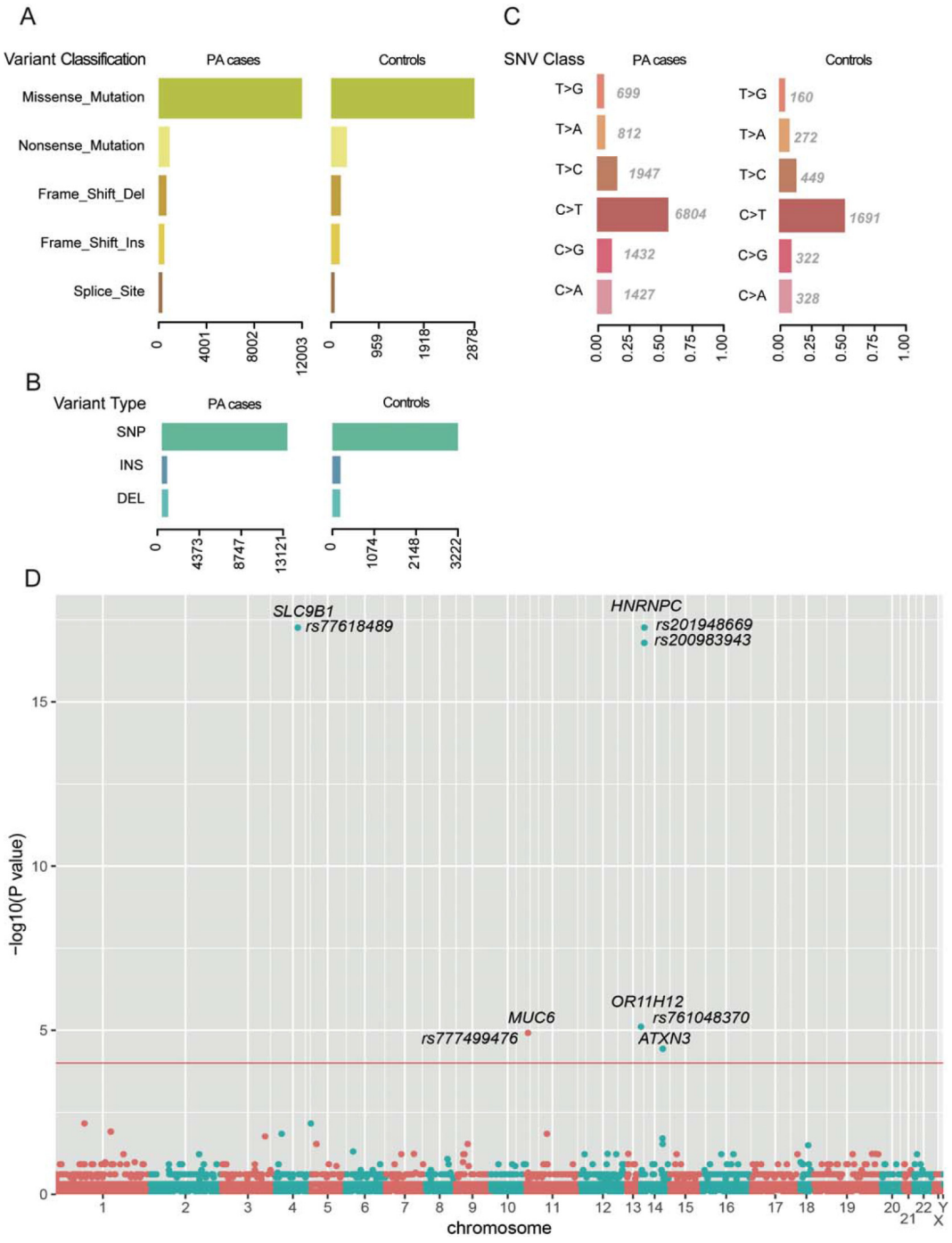

Fig. 2. The comparisons of the rare damaging variants between the PA and control groups. The number of variants in each variant classification, variant type, and SNV class are presented in (A), (B), and (C), respectively. (D) PA-associated variants by Fisher's exact test. Each node represents a variant, and the y-axis represent the statistical significance level. The threshold of the *P*-value is 0.0001, which is represented by the red line. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

FANCD2 were located within the *FANCD2* domain, suggesting that *FANCD2* domain was an important functional domain for this protein, and may be closely associated with PA (Fig. 5B). Moreover, we also observed a hotspot of *FGD5* at the splice site of the 2527th nucleotide of the coding sequence (Fig. 5B). In addition, the four missense variants in *FLT4* were located within the functional domains, while the nonsense variant at the 82nd amino acid disrupted all the functional domains. The mutational spectrum of the known PA genes improved our understanding of the functional roles of these variants in PA.

3.7. Comparative analysis of the PA subtypes

As we described above, the exact differences of genetic pattern between PA/VSD, TOF/PA, and PA/IVS remains unknown. We next performed comparative analysis and identified 11, 16, and 13 genes that only presented in PA/VSD, TOF/PA, and PA/IVS, respectively ($n \geq 2$, Fig. 6A). Most of these genes were directly connected with the known PA genes (Fig. 6B). Moreover, the total number of rare damaging variants was observed to be higher in TOF/PA than in PA/VSD and PA/IVS (Fig. 6C, $P < 0.1$). More specifically, TOF/PA

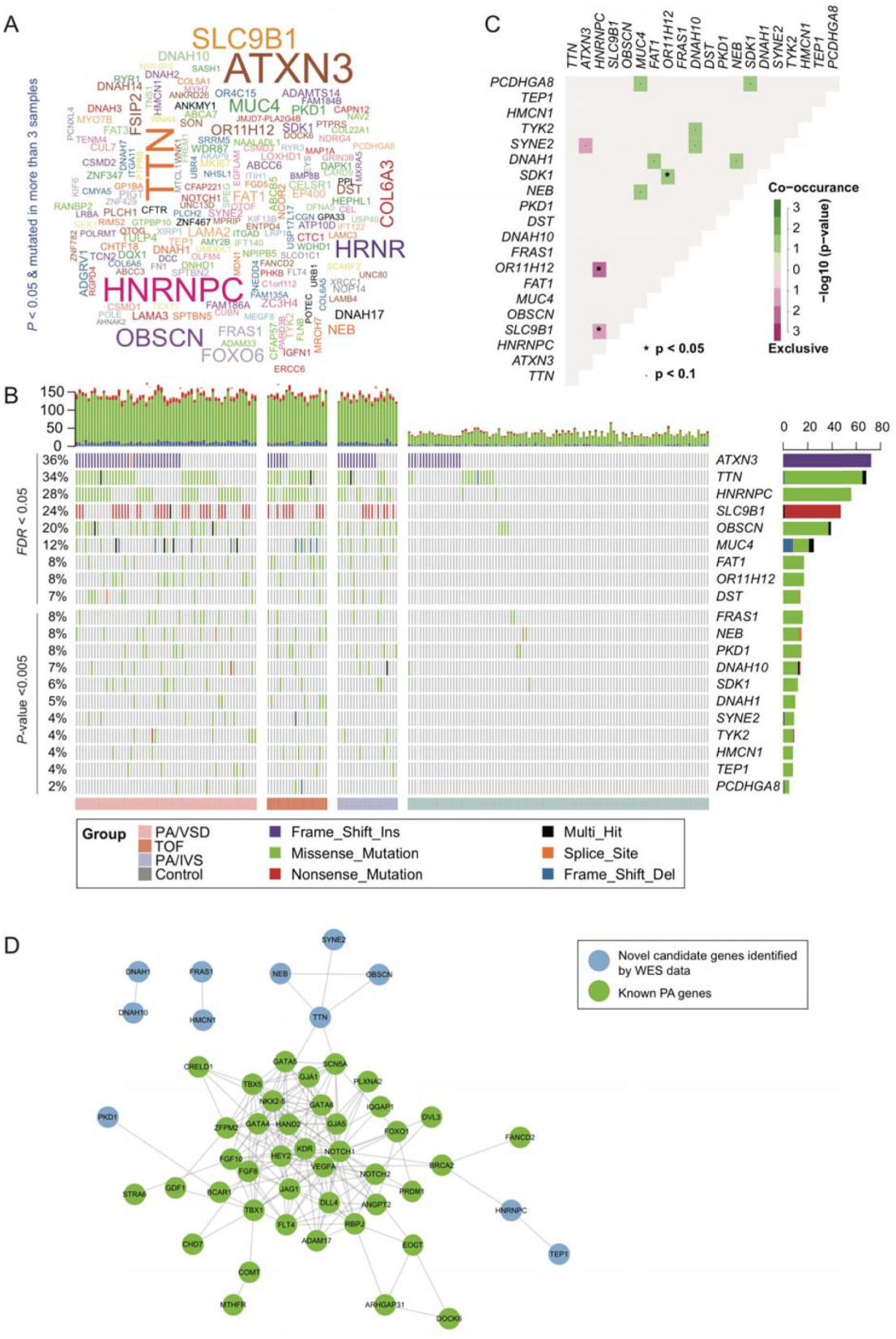

Fig. 3. The PA-associated genes identified by gene-based burden analysis. (A) The word-cloud of the PA-associated genes ($P < 0.05$ & mutated in more than 3 samples). (B) The landscape of the PA-associated genes in cases and controls. The bars on the top and right of the panel represent the number of variant types for each sample and each gene, respectively. The grey cells indicate that no variants were detected in the corresponding samples. (C) The corrplot of 20 genes were identified by gene-based burden analysis, where $^{*}P < 0.05$ and $^{*}P < 0.1$. (D) The protein–protein interaction (PPI) network of known (green nodes) and novel (blue nodes) PA-associated genes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

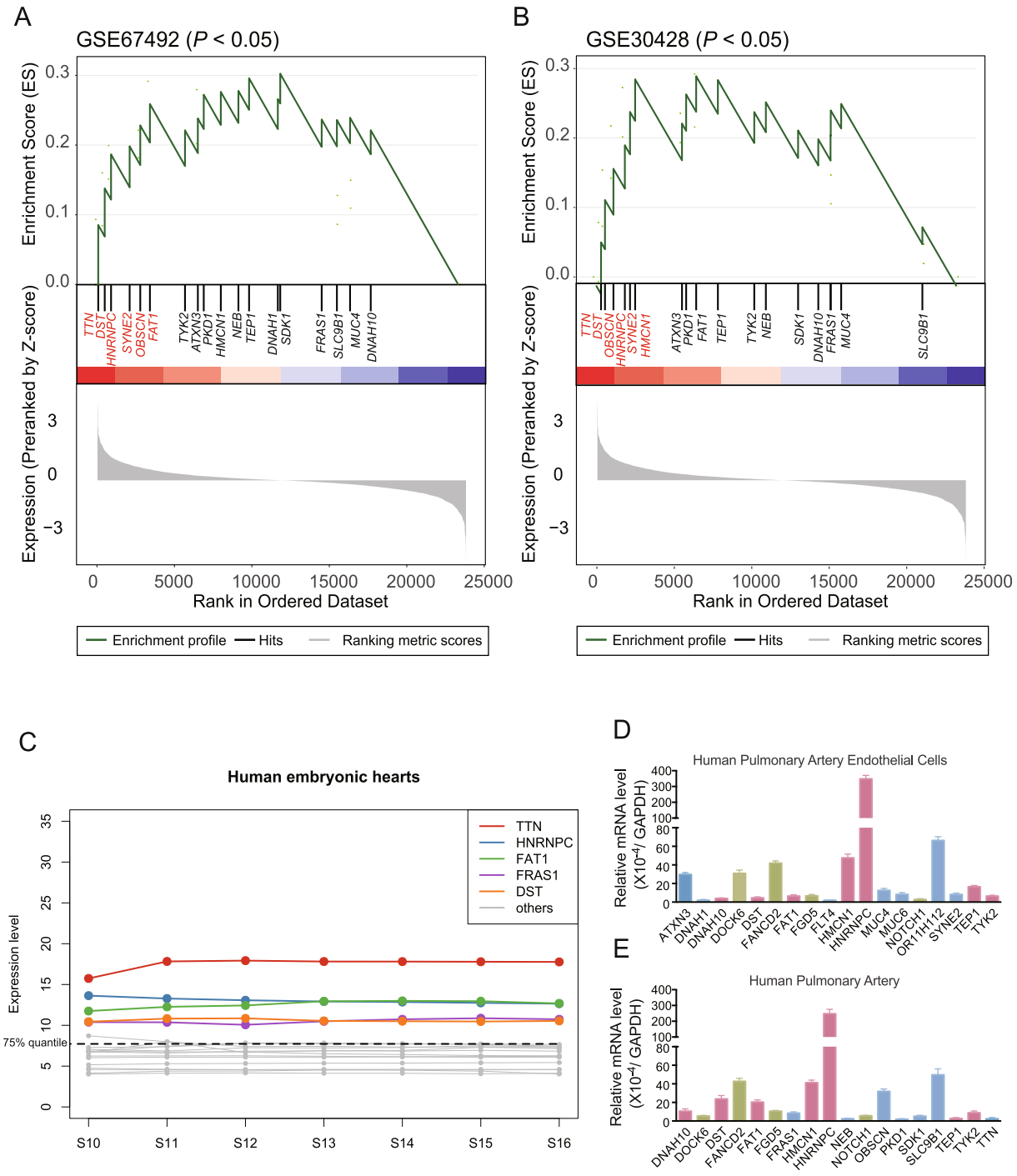

Fig. 4. The expression levels of the PA-associated genes in heart. The expression levels of novel PA-associated genes were investigated in cardiomyocytes of GSE67492 (A), GSE30428 (B), and human embryonic hearts (C). RT-qPCR was conducted to detect the expression levels of these novel PA-associated genes in HPAECs (D) and human pulmonary artery (E).

has more nonsynonymous SNVs than PA/IVS and PA/VSD, while PA/VSD has more loss-of-function (LOF) variants than PA/IVS and TOF/PA (Fig. 6C, $P < 0.1$).

4. Discussion

As a class of rare congenital heart disorders, the genetic etiology and underlying mechanism of PA remains unclear. To clarify the underlying genomic risks of PA and the differences in genetic background among PA/VSD, TOF/PA, and PA/IVS, we adopted WES to identify rare damaging variants and novel candidate genes in 100 PA cases and 100 healthy controls. Further network analysis and gene expression analysis demonstrated *DNAH10, DST, FAT1, HMCN1, HNRNPC, TEP1,* and *TYK2,* as the totally novel genes associated with PA pathogenesis. Since the genetic pattern between PA/VSD, PA/IVS and TOF/PA remain elusive, we had already tried to find out the differences in gene mutations and gene levels among these three types. However, since the number of patients differed widely (PA/VSD, n = 60; TOF/PA, n = 20; PA/IVS, n = 20), our candidate genes from variant-based association analysis and gene burden analysis in case-control were not figured out in any subtypes. We adopted comparative analysis and identified 11, 16,

Table 2
Rare variant of 7 candidate genes associated with PA.

Chromosome	Position	Base change	Gene	AAChange	TxChange	ExAC	1000G	gnomAD	SIFT	Polyphen2	Mutation Taster
Chr12	124270362	A > G	DNAH10	p.M373V	c.A1117G	–	–	–	D	D	D
Chr12	124311277	A > G	DNAH10	p.N1290S	c.A3869G	6.63E−05	–	–	D	P	D
Chr12	124317860	A > G	DNAH10	p.E1464G	c.A4391G	–	–	–	D	D	D
Chr12	124333343	C > T	DNAH10	p.R1888X	c.C5662T	2.88E−05	–	3.23E−05	–	–	A
Chr12	124341713	C > -	DNAH10	p.C2065fs	c.6195delC	9.47E−06	–	–	–	–	–
Chr12	124349245	C > T	DNAH10	p.R2220W	c.C6658T	1.91E−05	–	6.46E−05	D	D	D
Chr12	124399076	C > T	DNAH10	p.T3400M	c.C10199T	0.0004	0.01	0.0005	D	D	D
Chr12	124409730	T > C	DNAH10	p.V3849A	c.T11546C	–	–	–	D	D	D
Chr12	124409745	G > A	DNAH10	p.G3854E	c.G11561A	0.0005	0.01	0.0006	D	D	D
Chr6	56357793	T > C	DST	p.N4711S	c.A14132G	1.07E−05	–	3.23E−05	T	D	–
Chr6	56362244	T > C	DST	p.H4600R	c.A13799G	2.85E−05	–	–	T	D	–
Chr6	56376149	T > C	DST	p.N4090S	c.A12269G	–	–	–	T	D	–
Chr6	56391232	C > T	DST	p.R3900H	c.G11699A	1.90E−05	–	3.23E−05	D	D	–
Chr6	56417016	G > A	DST	p.T3406M	c.C10217T	2.87E−05	–	3.23E−05	T	D	–
Chr6	56434698	A > G	DST	p.S2493P	c.T7477C	–	–	–	T	D	–
Chr6	56475225	C > A	DST	p.K1728N	c.G5184T	–	–	–	D	D	–
Chr6	56494178	T > C	DST	p.K1416E	c.A4246G	–	–	–	D	D	–
Chr6	56494186	T > A	DST	p.E1413V	c.A4238T	–	–	–	D	D	–
Chr6	56499727	A > G	DST	p.Y1038H	c.T3112C	–	–	–	D	D	–
Chr6	56504561	G > A	DST	p.T844I	c.C2531T	0.0004	–	0.0004	T	D	–
Chr6	56504843	G > A	DST	p.R802C	c.C2404T	4.74E−05	–	6.46E−05	D	D	–
Chr6	56569155	C > T	DST	p.V234M	c.G700A	–	–	3.23E−05	D	D	–
Chr6	56819282	A > C	DST	p.F35C	c.T104G	6.76E−05	–	6.46E−05	D	D	–
Chr4	187517784	C > T	FAT1	p.V4304M	c.G12910A	9.72E−06	–	–	D	D	D
Chr4	187525089	T > A	FAT1	p.I3531F	c.A10591T	–	–	–	D	P	D
Chr4	187527348	G > A	FAT1	p.T3409M	c.C10226T	0.0001	0.0017	3.24E−05	T	D	D
Chr4	187535411	T > C	FAT1	p.N3055D	c.A9163G	0.0001	–	9.68E−05	D	D	D
Chr4	187539275	A > G	FAT1	p.L2822P	c.T8465C	0.0007	0.01	0.0003	D	P	D
Chr4	187539671	T > C	FAT1	p.Y2690C	c.A8069G	4.74E−05	–	6.46E−05	T	D	D
Chr4	187540271	G > A	FAT1	p.P2490L	c.C7469T	–	–	–	D	D	D
Chr4	187542491	G > C	FAT1	p.T1750R	c.C5249G	–	–	–	D	D	D
Chr4	187549895	G > T	FAT1	p.T1449K	c.C4346A	–	–	–	D	D	D
Chr4	187584696	C > T	FAT1	p.D1113N	c.G3337A	0.0002	0.01	0.0002	D	D	D
Chr1	186008985	G > C	HMCN1	p.V2052L	c.G6154C	–	–	–	T	D	D
Chr1	186038870	C > T	HMCN1	p.A2652V	c.C7955T	0.0001	–	6.46E−05	T	D	D
Chr1	186056393	G > A	HMCN1	p.G3031S	c.G9091A	4.72E−05	0.0017	9.71E−05	D	D	D
Chr1	186060046	T > G	HMCN1	p.I3295S	c.T9884G	9.51E−06	–	9.72E−05	D	P	D
Chr1	186064443	A > G	HMCN1	p.M3455V	c.A10363G	1.89E−05	–	3.23E−05	D	P	D
Chr1	186106797	C > T	HMCN1	p.P4584S	c.C13750T	2.83E−05	–	3.23E−05	D	D	D
Chr1	186147652	G > T	HMCN1	p.D5350Y	c.G16048T	–	–	–	D	D	D
Chr1	186157125	C > T	HMCN1	p.R5509W	c.C16525T	6.60E−05	–	–	D	D	D
Chr14	21679599	T > C	HNRNPC	p.D268G	c.A803G	0.0025	–	0.0022	T	D	D
Chr14	21679600	C > A	HNRNPC	p.D268Y	c.G802T	0.0025	–	0.0022	T	D	D
Chr14	20836703	A > G	TEP1	p.C2593R	c.T7777C	–	–	–	D	D	D
Chr14	20846278	G > A	TEP1	p.R1876W	c.C5626T	4.71E−05	–	9.70E−05	D	D	N
Chr14	20850789	T > C	TEP1	p.Y1378C	c.A4133G	0.0002	0.0017	9.71E−05	D	D	N
Chr14	20852074	A > T	TEP1	p.S1180T	c.T3538A	2.24E−05	–	–	T	D	D
Chr14	20854615	C > T	TEP1	p.R951H	c.G2852A	2.94E−05	–	3.23E−05	T	D	N
Chr14	20854642	C > T	TEP1	p.R942H	c.G2825A	9.59E−06	–	3.23E−05	D	D	D
Chr19	10468778	G > A	TYK2	p.R738W	c.C2212T	7.42E−05	–	–	D	D	D
Chr19	10469928	G > A	TYK2	p.R700W	c.C2098T	3.98E−05	–	3.23E−05	D	D	D
Chr19	10472598	C > T	TYK2	p.V603M	c.G1807A	0.0003	–	0.0002	D	D	N
Chr19	10473102	G > A	TYK2	p.R503X	c.C1507T	1.22E−05	–	–	–	–	A
Chr19	10476375	C > T	TYK2	p.V277M	c.G829A	5.98E−05	–	3.23E−05	D	D	N
Chr19	10477205	C > T	TYK2	p.E173K	c.G517A	5.65E−05	–	–	D	D	D
Chr19	10489024	T > G	TYK2	p.Q20P	c.A59C	5.70E−05	–	3.24E−05	T	P	D

and 13 specific genes that only presented in PA/VSD, PA/IVS, or TOF/PA. In PPI network, we found that most of these genes were directly connected with the known PA genes. We hope to perform further study with larger sample size soon.

HNRNPC belongs to the subfamily of heterogeneous nuclear ribonucleoproteins (hnRNPs). Two rare variants were detected in 31 PA/VSD patients, 13 TOF/PA patients, 9 PA/IVS patients with a high percentage of 53% in our cohort. Moreover, we observed that *HNRNPC* was directly connected with the subnetwork of the known PA genes. Further investigation found that *HNRNPC* had high expressions in HPAECs, human pulmonary artery tissues, and human embryonic hearts, the same as observed in two publicly available gene expression datasets. Heterogeneous Telomere bound hnRNPs including hnRNP A1, A2-B1, D and E, and telomerase bound hnRNPs including hnRNPA1 C1/C2 and D [39]. *HNRNPA1* was highly expressed during cardiac development and two rare mutations of *HNRNPA1* were detected in human CHDs. *HNRNPA1*$^{-/-}$ mice displayed the VSD, PTA, DORV phenotypes, and such modification led to the dysregulation of cardiac transcription factors, including Isl1, Nkx2.5, and Tbx1, and of signaling pathways, including BMP, FGF, and Notch pathways [40,41]. Since then, *HNRNPC* was considered as the most significant and novel candidate genes for patients of PA. *TEP1*, which is a component of the ribonucleoprotein complex responsible for telomerase activity which catalyzes the addition of new telomeres on the chromosome ends, was observed to have strong correlation with

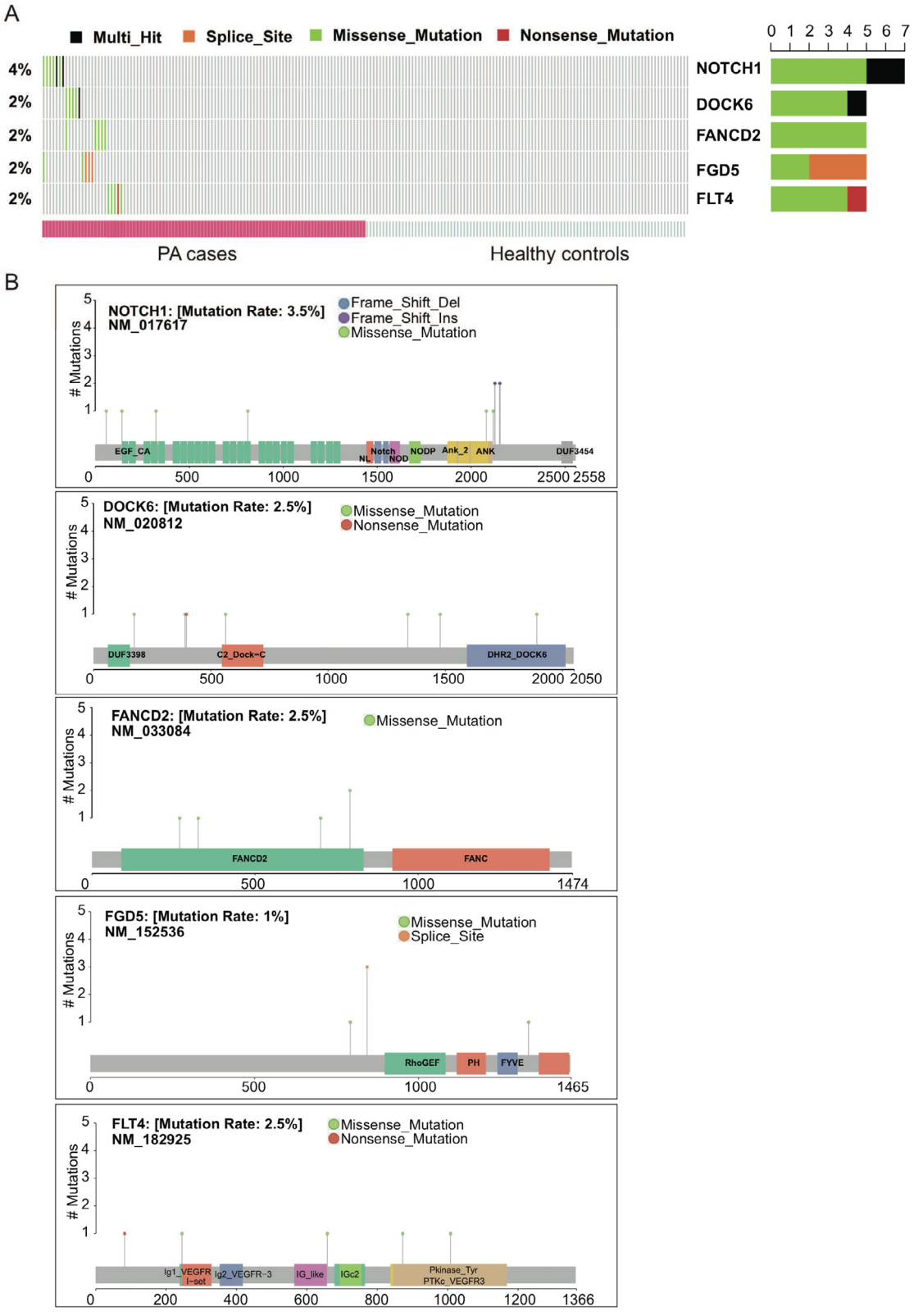

Fig. 5. Mutational spectrum of five known PA genes. (A) The distribution of the five known PA genes, *NOTCH1*, *DOCK6*, *FANCD2*, *FGD5*, and *FLT4*, in the PA cases. (B) The specific amino acid sites of variants on the encoded proteins or protein domains.

HNRNPC, which directly connected with known PA genes in PPI network. *TEP1* was detected to have 6 rare damaging variants from 6 PA/VSD, 1 PA/IVS and 1 TOF/PA patients. Although the expression of *TEP1* was relatively low, we still conjectured that TEP1 and *HNRNPC* might have common effects on PA pathogenesis. Since the relationship between *HNRNPC*, *TEP1* and congenital heart defect remained unclear, their functions deserved to be further studied.

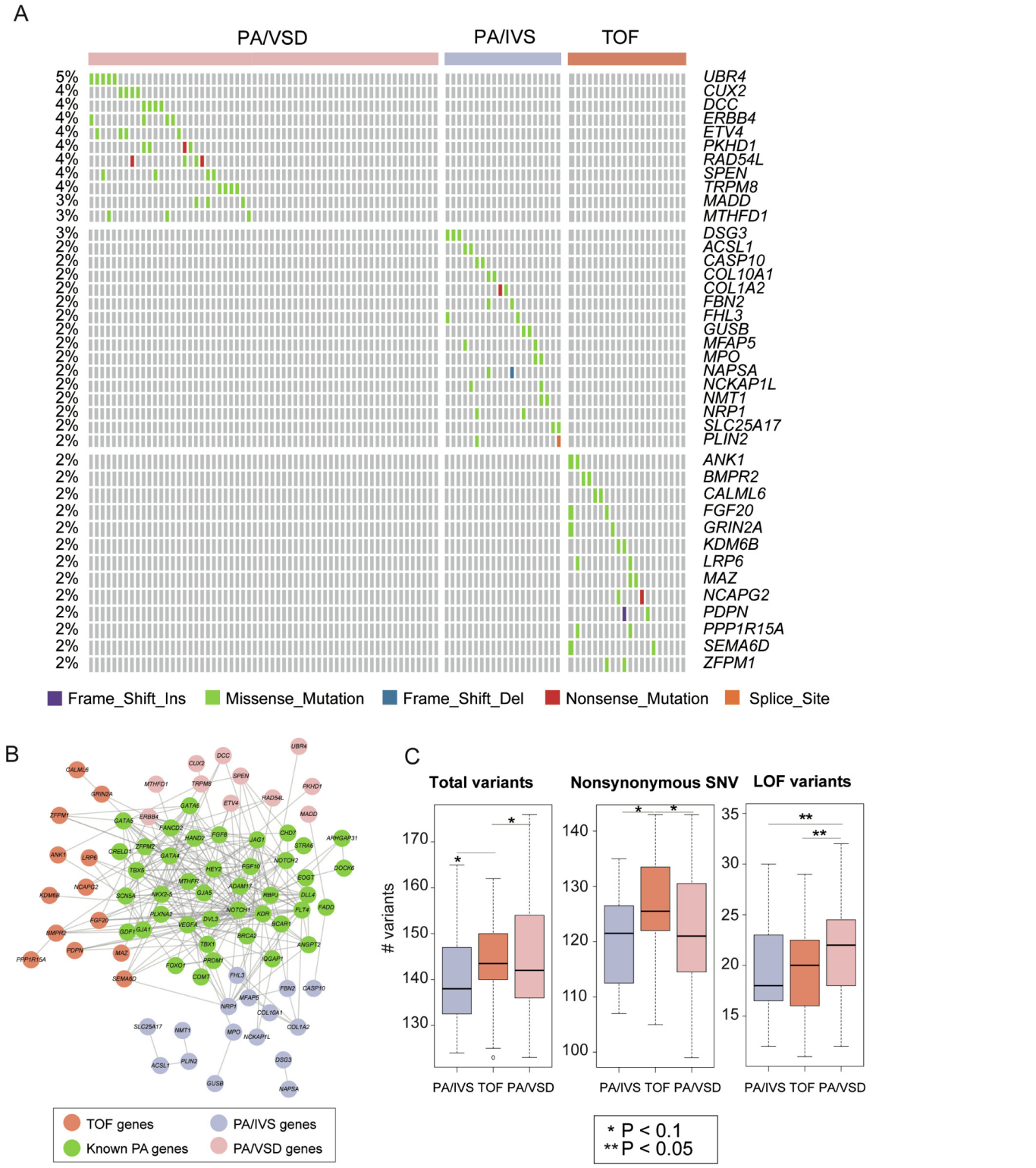

Fig. 6. PA-subtype associated genes/variants and variant types. (A) The PA-subtype-specific genes that only present in one of the three subtypes, and absent in controls. (B) The PPI subnetwork constructed by mapping the known PA genes and the PA-subtype associated genes. (C) The comparisons between the three PA-subtypes about the number of total variants, nonsynonymous SNVs, and LOF variants.

DNAH1 and *DNAH10* belong to a component of the inner dynein arms, which could be found in cilia and flagella and consist of the outer and inner dynein arms attached to the peripheral microtubule doublets. Our previous study showed that *DANH10* played an important role in the pathogenesis of defects in heterotaxy syndrome patients with congenital heart defects [42]. In our cohort, 5 PA/VSD patients and 5 TOF/PA patients were detected to have the rare variants in *DANH1*, while 9 PA/VSD, 1 TOF/PA and 3PA/IVS were detected in *DANH10*. We found that *DNAH1-DNAH10* was observed to be co-occurred in corrplot and they had a directly connection in network. We concluded that the gene pairs with co-occurrence might be relevant to the same biological process

associated with PA pathogenesis. *FAT1* encodes a tumor suppressor essential for cellular polarization, directed cell migration and modulating cell–cell contact [43]. *FAT1* had 10 rare damaging variants appearing in 7 PA/VSD, 3PA/IVS, and 7 TOF/PA after filtering the WES data, which was co-occurrence with *DNAH1*. We found that *FAT1* was highly expressed during human heart embryogenesis, yet whether *FAT1* play a causal or modifier role in PA is still elusive. Lacking literature supports, the underlying mechanism of this interaction of *DNAH1- DNAH10* and *DNAH1-FAT* need further study.

DST belongs to a member of the plakin protein family of adhesion junction plaque proteins, which has multiple cytoskeleton-binding domains. It exhibited a high expression in heart and pulmonary artery tissues in GSEA analysis [44,45]. 13 rare damaging variants of *DST* were discovered in 9 PA/VSD, 2TOF/PA, 3PA/IVS patients. In accordance to previous studies, *DST* plays an important role in cell–cell adhesion structures, including desmosomes, gap junctions, and adherens junctions of cardiac muscle [46]. 3 PA/VSD, 3 TOF/PA, and 2 PA/VSD patients had 8 rare damaging variants in *HMCN1,* which encodes a large extracellular member of the immunoglobulin superfamily. *HMCN1* showed the similar significance mRNA expression level as *HNRNPC.* Previous study demonstrated that *HMCN1* could affect the formation of transient cell contacts that are required for tissue organization, migration, invasion, as well as the formation of stable cell–cell and cell–ECM contacts by TGF-β signaling pathway [47]. *TYK2* encodes a member of the tyrosine kinase, which associates with the cytoplasmic domain of type I and type II cytokine receptors and promulgates cytokine signals by phosphorylating receptor subunits [48,49]. *TYK2* got 7 rare damaging variants in 7 PA/VSD, 2 PA/IVS patients. Thus far, the functions of *DST, HMCN1* and *TYK2* in cardiovascular development remain unknown, and they might be associated with PA pathogenesis in novel manners.

Our study did have some limitations. First, small sample size made it difficult to conform our candidate genes in validation cohort. Then the lack of parental samples limited our ability to study the genetic backgrounds of these patients. In addition, the functions of our candidate genes need to be further verified with transgenic animal research. In summary, an effective analytical bioinformatic approach allowed us to identify rare damage variants in novel genes, which would play a vital role in PA pathology. Our candidate genes open new fields of investigation into PA pathology and provide novel insights into pulmonary artery development.

5. Conclusions

This analysis supports 7 novel candidate genes (*DNAH10, DST, FAT1, HMCN1, HNRNPC, TEP1,* and *TYK2*) which have not previously been reported in either humans or animals, contributing to the etiology of PA. Different subtypes of PA, namely PA/VSD, PA/IVS and TOF/PA had different genetic pattern to a certain extent. Overall, the data presented here builds confidence in further exploring the underlying mechanism of PA, which will encourage further verification with fundamental research.

Ethics approval and consent to participate

The present study was approved by the Ethics Committee of the Xinhua Hospital Affiliated to the Shanghai Jiao Tong University School of Medicine, Shanghai, China. This study was conducted in accordance with the principles expressed in the Declaration of Helsinki. Participants and/or their legal guardians involved in this study gave a written informed consent prior to inclusion in the study.

Declaration of competing interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

We would like to acknowledge funding from the National Natural Science Foundation of China (81974021, 81974012, 81720108003, 81670285, and 81900280), Shanghai Sailing Program (19YF1431600), and the Experts Recruitment Program of Xinhua Hospital (Re-013).

Authors' contributions

KS, YY, and YNL conceived and designed the project and are responsible for the overall content. LZ, and TLS analyzed and interpreted the WES data. HLX, RLZ, QHF, and SC collected the samples and clinical information. XS, LZ, and KB prepared the manuscript. KS, YY, and YNL contributed to revising the manuscript. All authors contributed to and discussed the results and critically reviewed the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank all the patients who contributed samples for this research.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2020.01.011.

References

[1] Mackie AS, Gauvreau K, Perry SB, del Nido PJ, Geva T. Echocardiographic predictors of aortopulmonary collaterals in infants with tetralogy of fallot and pulmonary atresia. J Am Coll Cardiol 2003;41(5):852–7.
[2] Honjo O, Al-Radi OO, MacDonald C, et al. The functional intraoperative pulmonary blood flow study is a more sensitive predictor than preoperative anatomy for right ventricular pressure and physiologic tolerance of ventricular septal defect closure after complete unifocalization in patients with pulmonary atresia, ventricular septal defect, and major aortopulmonary collaterals. Circulation 2009;120(11 Suppl):S46–52.
[3] van der Linde D, Konings EE, Slager MA, et al. Birth prevalence of congenital heart disease worldwide: a systematic review and meta-analysis. J Am Coll Cardiol 2011;58(21):2241–7.
[4] Gao M, He X, Zheng J. Advances in molecular genetics for pulmonary atresia. Cardiol Young 2017;27(2):207–16.
[5] Digilio MC, Marino B, Grazioli S, Agostino D, Giannotti A, Dallapiccola B. Comparison of occurrence of genetic syndromes in ventricular septal defect with pulmonic stenosis (classic tetralogy of Fallot) versus ventricular septal defect with pulmonic atresia. Am J Cardiol 1996;77(15):1375–6.
[6] Xu YJ, Wang J, Xu R, et al. Detecting 22q11.2 deletion in Chinese children with conotruncal heart defects and single nucleotide polymorphisms in the haploid TBX1 locus. BMC Med Genet 2011;12:169.
[7] Pediatric Cardiac Genomics C, Gelb B, Brueckner M, et al. The Congenital Heart Disease Genetic Network Study: rationale, design, and early results. Circ Res 2013;112(4):698–706.
[8] Reddy VM, McElhinney DB, Amin Z, et al. Early and intermediate outcomes after repair of pulmonary atresia with ventricular septal defect and major aortopulmonary collateral arteries: experience with 85 patients. Circulation 2000;101(15):1826–32.
[9] Chubb H, Pesonen E, Sivasubramanian S, et al. Long-term outcome following catheter valvotomy for pulmonary atresia with intact ventricular septum. J Am Coll Cardiol 2012;59(16):1468–76.
[10] Tulzer G, Arzt W, Franklin RC, Loughna PV, Mair R, Gardiner HM. Fetal pulmonary valvuloplasty for critical pulmonary stenosis or atresia with intact septum. Lancet 2002;360(9345):1567–8.
[11] Daubeney PE, Delany DJ, Anderson RH, et al. Pulmonary atresia with intact ventricular septum: range of morphology in a population-based study. J Am Coll Cardiol 2002;39(10):1670–9.

[12] Greenway SC, Pereira AC, Lin JC, et al. De novo copy number variants identify new genes and loci in isolated sporadic tetralogy of Fallot. Nat Genet 2009;41(8):931–5.
[13] Xie L, Chen JL, Zhang WZ, et al. Rare de novo copy number variants in patients with congenital pulmonary atresia. PLoS ONE 2014;9(5):e96471.
[14] Granados-Riveron JT, Ghosh TK, Pope M, et al. Alpha-cardiac myosin heavy chain (MYH6) mutations affecting myofibril formation are associated with congenital heart defects. Hum Mol Genet 2010;19(20):4007–16.
[15] Silversides CK, Lionel AC, Costain G, et al. Rare copy number variations in adults with tetralogy of Fallot implicate novel risk gene pathways. PLoS Genet 2012;8(8):e1002843.
[16] Reamon-Buettner SM, Borlak J. HEY2 mutations in malformed hearts. Hum Mutat 2006;27(1):118.
[17] Soemedi R, Wilson IJ, Bentham J, et al. Contribution of global rare copy-number variants to the risk of sporadic congenital heart disease. Am J Hum Genet 2012;91(3):489–501.
[18] Cario H, Smith DE, Blom H, et al. Dihydrofolate reductase deficiency due to a homozygous DHFR mutation causes megaloblastic anemia and cerebral folate deficiency leading to severe neurologic disease. Am J Hum Genet 2011;88(2):226–31.
[19] He X, Zhang X, Jing H, et al. Rare copy number variations might not be involved in the molecular pathogenesis of PA-IVS in an unselected Chinese cohort. Pediatr Cardiol 2019;40(4):762–7.
[20] Grunert M, Dorn C, Cui H, et al. Comparative DNA methylation and gene expression analysis identifies novel genes for structural congenital heart diseases. Cardiovasc Res 2016;112(1):464–77.
[21] Wang J, Lu Y, Chen H, Yin M, Yu T, Fu Q. Investigation of somatic NKX2-5, GATA4 and HAND1 mutations in patients with tetralogy of Fallot. Pathology 2011;43(4):322–6.
[22] Beauchesne LM, Warnes CA, Connolly HM, et al. Prevalence and clinical manifestations of 22q11.2 microdeletion in adults with selected conotruncal anomalies. J Am Coll Cardiol 2005;45(4):595–8.
[23] Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010;26(5):589–95.
[24] Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25(16):2078–9.
[25] McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20(9):1297–303.
[26] Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 2003;31(13):3812–4.
[27] Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Human Genet 2013;7. Chapter 7:Unit7 20.
[28] Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods 2010;7(8):575–6.
[29] Genomes Project C, Auton A, Brooks LD, et al. A global reference for human genetic variation. Nature 2015;526(7571):68–74.
[30] Karczewski KJ, Weisburd B, Thomas B, et al. The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Res 2017;45(D1):D840–5.
[31] Richards S, Aziz N, Bale S, 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(5):405–24.
[32] Szklarczyk D, Franceschini A, Wyder S, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res 2015;43(Database issue):D447–52.
[33] Brohee S, Faust K, Lima-Mendez G, Vanderstocken G, van Helden J. Network Analysis Tools: from biological networks to clusters and pathways. Nat Protoc 2008;3(10):1616–29.
[34] Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13(11):2498–504.
[35] Hemnes AR, Brittain EL, Trammell AW, et al. Evidence for right ventricular lipotoxicity in heritable pulmonary arterial hypertension. Am J Respir Crit Care Med 2014;189(3):325–34.
[36] Kreymborg K, Uchida S, Gellert P, et al. Identification of right heart-enriched genes in a murine model of chronic outflow tract obstruction. J Mol Cell Cardiol 2010;49(4):598–605.
[37] Shi X, Huang T, Wang J, et al. Next-generation sequencing identifies novel genes with rare variants in total anomalous pulmonary venous connection. EBioMedicine 2018;38:217–27.
[38] Nolan T, Hands RE, Bustin SA. Quantification of mRNA using real-time RT-PCR. Nat Protoc 2006;1(3):1559–82.
[39] Ford LP, Wright WE, Shay JW. A model for heterogeneous nuclear ribonucleoproteins in telomere and telomerase regulation. Oncogene 2002;21(4):580–3.
[40] Zhang L, Chen Q, An W, et al. Novel pathological Role of hnRNPA1 (heterogeneous nuclear ribonucleoprotein A1) in vascular smooth muscle cell function and neointima hyperplasia. Arterioscler Thromb Vasc Biol 2017;37(11):2182–94.
[41] Yu Z, Tang PL, Wang J, et al. Mutations in Hnrnpa1 cause congenital heart defects. JCI Insight 2018;3(2).
[42] Liu C, Cao R, Xu Y, et al. Rare copy number variants analysis identifies novel candidate genes in heterotaxy syndrome patients with congenital heart defects. Genome Med 2018;10(1):40.
[43] Lahrouchi N, George A, Ratbi I, et al. Homozygous frameshift mutations in FAT1 cause a syndrome characterized by colobomatous-microphthalmia, ptosis, nephropathy and syndactyly. Nat Commun 2019;10(1):1180.
[44] Horie M, Yoshioka N, Takebayashi H. BPAG1 in muscles: structure and function in skeletal, cardiac and smooth muscle. Semin Cell Dev Biol 2017;69:26–33.
[45] Rahimi N. Defenders and challengers of endothelial barrier function. Front Immunol 2017;8:1847.
[46] Steiner-Champliaud MF, Schneider Y, Favre B, et al. BPAG1 isoform-b: complex distribution pattern in striated and heart muscle and association with plectin and alpha-actinin. Exp Cell Res 2010;316(3):297–313.
[47] Chowdhury A, Herzog C, Hasselbach L, et al. Expression of fibulin-6 in failing hearts and its role for cardiac fibroblast migration. Cardiovasc Res 2014;103(4):509–20.
[48] Trenti A, Zulato E, Pasqualini L, Indraccolo S, Bolego C, Trevisi L. Therapeutic concentrations of digitoxin inhibit endothelial focal adhesion kinase and angiogenesis induced by different growth factors. Br J Pharmacol 2017;174(18):3094–106.
[49] Cherian AV, Fukuda R, Augustine SM, Maischein HM, Stainier DY. N-cadherin relocalization during cardiac trabeculation. Proc Natl Acad Sci U S A 2016;113(27):7569–74.