Whole-exome Sequencing Identify Rare Variants in Novel Candidate Genes with Non-syndromic Patent Ductus Arteriosus

Ying Gao
Shidong Hospital
Ying Liu
Shidong Hospital
Jiaoyu Li
Xin Hua Hospital, School of Medicine, Shanghai Jiao Tong University
Yinghui Chen
Xin Hua Hospital, School of Medicine, Shanghai Jiao Tong University
Qi Zhang
Xin Hua Hospital, School of Medicine, Shanghai Jiao Tong University
Bingyao Zhang
Xin Hua Hospital, School of Medicine, Shanghai Jiao Tong University
Pengjun Zhao
Xin Hua Hospital, School of Medicine, Shanghai Jiao Tong University
Bo Chen (1205293877@qq.com)
Xin Hua Hospital, School of Medicine, Shanghai Jiao Tong University

Research Article

Keywords: Congenital heart defects, Patent Ductus Arteriosus, Whole-exome sequencing, Rare variants

Posted Date: January 15th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-142511/v1

Whole-exome Sequencing Identify Rare Variants in Novel Candidate Genes with Non-syndromic Patent Ductus Arteriosus

Ying Gao [1#], Ying Liu[1#], Jiaoyu Li[2#], YingHui Chen[2], Qi Zhang[2], Bingyao Zhang[2], Pengjun Zhao[2*] & Bo Chen[2*]

[1] Department of Pediatric, Shidong Hospital, Shanghai, China

[2] Department of Pediatric Cardiology, Xin Hua Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China

[*]Correspondence: Pengjun Zhao: pjunzhao@sina.com; Bo Chen: 1205293877@qq.com

[#]Ying Gao, Ying Liu and Jiaoyu Li contribute to this study equally.

Abstract

Background

Patent Ductus Arteriosus (PDA) is one of the most common congenital heart defects that can cause pulmonary hypertension, heart failure, and even death. Prior studies have suggested a role for genetics in determining spontaneous ductal closure, however the clinical characteristics and genetic cause underlying PDA remain unclear.

Results

Therefore, to further explore genetic etiology of PDA, we applied Whole-exome Sequencing (WES) in 39 unrelated isolated, non-syndromic PDA patients and 100 healthy controls. Through a series of bio-information filtering strategies, the candidate genes are prioritized by comprehensively considering factors such as gene functional enrichment, expression pattern and mutation burden during heart development. 18 rare damage variants of 6 total novel genes (SOX8, NES, CDH2, ANK3, EIF4G1, HIPK1) were identified for the first time and these pathogenic candidates are also highly expressed in the heart of human embryos.

Conclusions

WES is an efficient diagnostic tool for identifying PDA related genes. The finding of our study contributes new insights into the molecular basis of PDA and may inform further studies on genetic risk factors for this congenital birth defect.

Keywords: Congenital heart defects, Patent Ductus Arteriosus, Whole-exome sequencing, Rare variants

Background

The ductus arteriosus (DA) is a normal fetal structure connecting the pulmonary artery and descending aorta to maintain blood circulation in fetal period[1]. It becomes pathological if it remains patent after birth[1]. Failure of the ductus arteriosus to close after birth is termed patent ductus arteriosus (PDA) and is one of the most common heart defects. It is accounting for 15%-20% of the total number of congenital heart defect. Its incidence is about 1/2000 in term infant and 8/1000 in premature infant[2]. Persistent ductal shunting may lead to pulmonary overcirculation and induce systemic hypoperfusion, increasing the risk of pulmonary hypertension, infective endocarditis, heart failure and even death[3].

From the perspective of cardiac development, the DA shut down functionally in 15 hours after birth in healthy term infants[4]. This process occurs by abrupt contraction of the muscular wall of the PDA, which is associated with a balance of neurohumoral factors. The increase of contractile elements, such as PO2 and endothelin-1, and the decrease of relaxants, such as PGE2 levels and nitric oxide, are the main factors to cause the closure of ductus arteriosus[4]. Under the action of these hormones, neural-crest-derived cells migrate into the subendothelial space, transform to vascular smooth muscle cells (VSMCs). Then with the contraction of the medial membrane and the circular muscle in the ductus arteriosus, the lumen is shortened and finally closed[5]. The occurrence of PDA has both inherited and acquired causes. However, the etiology and pathogenesis are still not completely known yet.

The understanding of the genetic mechanism of PDA initially came from the syndrome type patent ductus arteriosus. Chromosomal abnormalities, including aneuploidy and microdeletion, are the most common causes of ductus arteriosus syndrome. Previous studies have confirmed several syndromes with

patent ductus arteriosus including Turner Syndrome (45, XO), Kartagener Syndrome, Klinefelter Syndrome (47, XXY), etc. [6-8]. Other than chromosomal rearrangements, a single gene mutation can also cause syndromic PDA, including Noonan Syndrome (PTPN11 mutation), Holt-Oram Syndrome (TBX5 mutation) and Char Syndrome (TFAP2B mutation)[9-11]. With the fruition for human genome sequencing, genetic factors assume a paramount part in the pathogenesis of PDA. However, little is known about the genetic mechanism of isolated non-syndromic patent ductus arteriosus. Previous studies have demonstrated that rare damaging mutations in MYH11, TFAP2B were detected in some isolated non-syndromic PDA patients[12]. Erdogan et al. conducted an array comparative genome screening in 105 patients with CHD and found a 1.92MB deletion in 1q21.1(CJA5) in an isolated PDA patient[13]. However, most of previous studies focus on the known pathogenic gene mutations of syndromic PDA. The molecular genetic mechanisms of non-syndromic PDA are still largely unknown. Using WES and bioinformatics methods to detect rare variants associated to PDA have never been reported yet.

In our study, to systematically examine the clinical characteristics and genetic cause of isolated, non-syndromic PDA, we recruited 39 unrelated, isolated, non-syndromic PDA patients and 100 healthy children to performed whole exome sequencing (WES). Through a series of bioinformatics filtering steps, we identified 18 rare damaging variants in 6 candidate genes (SOX8, NES, CDH2, ANK3, EIF4G1, HIPK1). Notably, we found that these candidate genes are highly expressed in human embryonic hearts. Therefore, we hope our discovery of the pathogenic genes could fill the underlying mechanism of PDA and promote further experimental analysis.

Methods

Patients and Consents

39 unrelated isolated, non-syndromic PDA patients (Han Chinese) and 100 healthy children were recruited in Xinhua hospital affiliated to Shanghai Jiao Tong University. All structural heart phenotypes were assessed by echocardiography or cardiac catheterization in both groups. And the case groups were further diagnosed by cardiac catheterization or surgery. Patients with multiple major cardiac defects were excluded. Similarly, patients with any pregnancy risk factors such as premature birth, infection were also excluded. The study protocol and the ethics were approved by the medical ethics committee of Xinhua Hospital (Approval No. XHEC-D-2020-001). All patients and their parents signed the informed consents. We have also certified that the study was strictly in accordance with the Declaration of Helsinki and International Ethical Guidelines for Health-related Research Involving Humans.

DNA extraction and Whole Exome Sequencing

The genomic DNA of all participants was extracted from blood samples by using the QIAamp DNA Blood Mini Kit (QIAGEN, Germany). DNA samples were stored at -80 °C until further use. Genomic DNA was eluted, purified, and amplified by ligation-mediated PCR and then subjected to DNA sequencing on the Illumina platform. Qualified DNA samples from the groups of PDA and controls were performed WES to detecting rare variations. The Clean data was obtained by removing adaptor sequences and low- quality reads.

SNP identified and Quality fliting

Under default settings, BWA-mem (v0.7.12-r1039) [14] was used to map clean data to 1000 Genomes Project (Version human_glk_v37). Duplicated reads were marked and removed by PICARD software. The resulting BAM files were then sorted and indexed by Base Quality Score Recalibration (BQSR)[15]. Then we used GATK HaplotypeCaller module to detect variants based on the American College of

Medical Genetics (ACMG) criteria guidelines. Next, the GVCF files of all samples were subjected to joint genotyping. Variant quality control and filtering were performed based on variant quality score recalibration (VQSR) by building GMM model[16]. We used ANNOVAR53 to annotate the variants for functional and population frequency information. All potentially damaging variants on the candidate genes were classified into five groups, including pathogenic, likely pathogenic, variant uncertain significance, likely benign and benign[17]. We fileted for rare damaging variants with the following criteria: (1) read quality> 20 bp, (2) minor allele frequency (MAF)<5% (3) variants frequency < 1% in 1000G database, ESP6500 database, Exac database and gnomAD database, (4) Removing small (< 10bp) non-shift indel mutation in Repeat region, (5) fliting out synonymous mutations and non-synonymous mutations that are not predicted to be deleterious by PolyPhen, SIFT, or MutationTaster tools, (6) fliting out the variants without annotation information in all exome database[17].

Variants Fliting based on Fisher Exact test and Burden analysis

Differences in baseline characteristics between cases and controls were assessed by a Fisher exact test for categorical variables with the "R" statistical package. For comparisons, ***P***-value < 0.05 was considered statistically significant. Subsequently, for capturing rare target genes in a limited range, we aggregated the SNP data based on the gene level and conducted gene-based Burden analysis to increase statistical power. The different variant sites located in the same gene were put together as a whole for disease association analysis. We filtered for candidate genes based on Burden analysis with the following criteria: (1) ***P***-value < 0.05 or FDR < 0.05, (2) hit by at least one variant in 3 cases (3) not found in any sample of control group. Then we prioritized genes based on Fisher exact test and Burden analysis.

Functional enrichment analysis and Network analysis

To further fliting the candidate gene associated with PDA, we performed functional enrichment analysis to identify the function of above candidate genes. Pathway analysis of the candidate gene profiling results was performed using the Gene ontology (GO version: 30.10.2017) and KEGG pathway (http://www.genome.jp/kegg/pathway.html) mapping within the web-based tool database for annotation, visualization and integrated discovery[18, 19]. The significant threshold was set to be an adjusted P-value < 0.05. In addition, we also prioritized those genes based on functional enrichment analysis. Furthermore, to detect relationship between our candidate genes and known disease-causing genes, we performed a protein-protein interaction (PPI) network analysis[20]. PPI gene network was generated by Cytoscape software based on STRING database.

Tissue collection and Expression detection

In addition to the genes prioritized above, we also prioritized genes according to the expression in human embryonic heart. Previous studies have divided eight embryonic weeks (56 days) into 23 internationally accepted Carnegie stages[21]. To further investigate the potential function of our candidate gene, we collected human embryonic heart in different Carnegie stages from S10 to S16 after medical termination of pregnancy from Xinhua hospital. RNA was extracted and purified by Experion automated gel electrophoresis system and e RNeasy MinElute Cleanup Kit. Then we used Affymetrix HTA 2.0 microarray to detect the expression patterns of our candidate genes.

Results

Table 1: Characteristics of 39 PDA Patients

Patients Characteristics	Numbers
Age	2.92±2.44
Male n (%)	15（38%）
Female n (%)	24（61%）
Male-to-Female ration （%）	62%
BMI (kg/m2)	16.58±4.34
PDA size (mm)	2.87±1.68
Birth weight（kg）	2.96±0.73
Gestational age（week）	39.04±1.46
Associated cardiac defect n (%)	
VSD n (%)	2（5%）
ASD n (%)	7（18%）
Others n (%)	2（5%）

All values are expressed as mean ± SD or n (%).

Population

39 unrelated isolated, non-syndromic PDA patients (Han Chinese) and 100 healthy children were recruited in Xinhua hospital with ages ranging from 2 months to 13 years. Among these patients, 28%

were accompanied by common cardiac defects, including atrial septal defect (ASD, n = 7), ventricular septal defect (VSD, n = 2) and others (n=2) (Table 1). All our samples are full-term, and no one was accompanied by other major cardiac structural abnormalities or developmental syndrome. Whole-exome sequencing, performed in all samples at an average depth of coverage of approximately 105 times per base, identified 411344 single-nucleotide variants (SNVs) and 23 101 indels across the genome. Through a series of filtering strategies mentioned in Figure 1, 11910 rare damaging variants were screen with a threshold of minor allele frequency (MAF) at 0.5%. As illustrate in figure 2, we found more rare damaging variants in PDA group than control group, which was observed in splice site, nonsense mutation and missense mutation. Consistently, the mutation type of C>T and G>A accounted for the majority of base mutations compared with other types (Figure 2). Based on these mutations, we next adopted a bioinformatic filtering strategy to identify candidate genes associated to PDA.

Figure 1

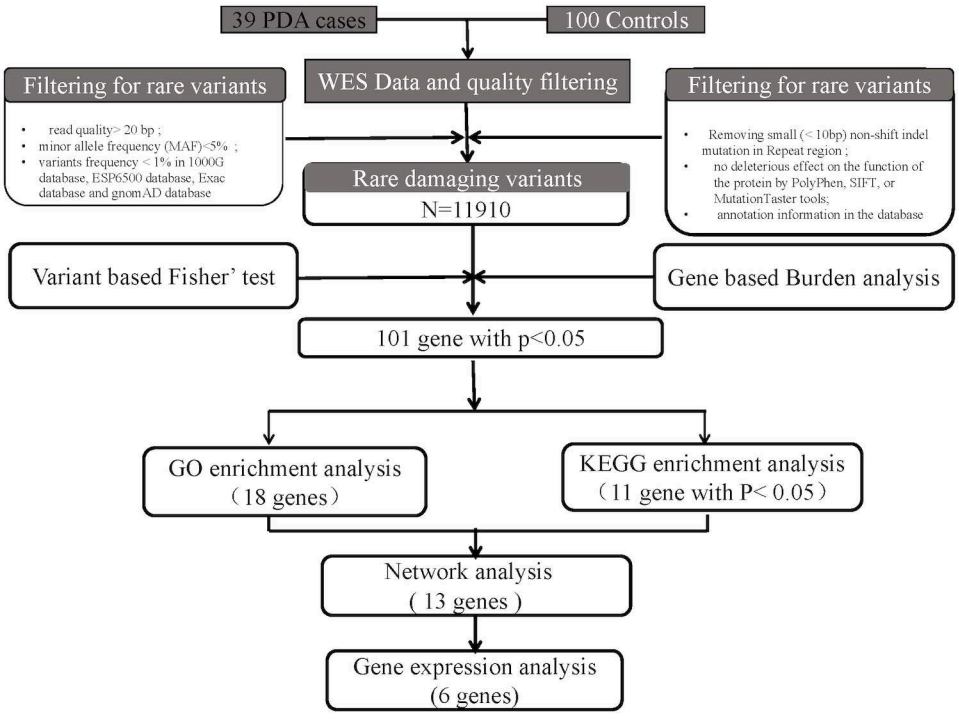
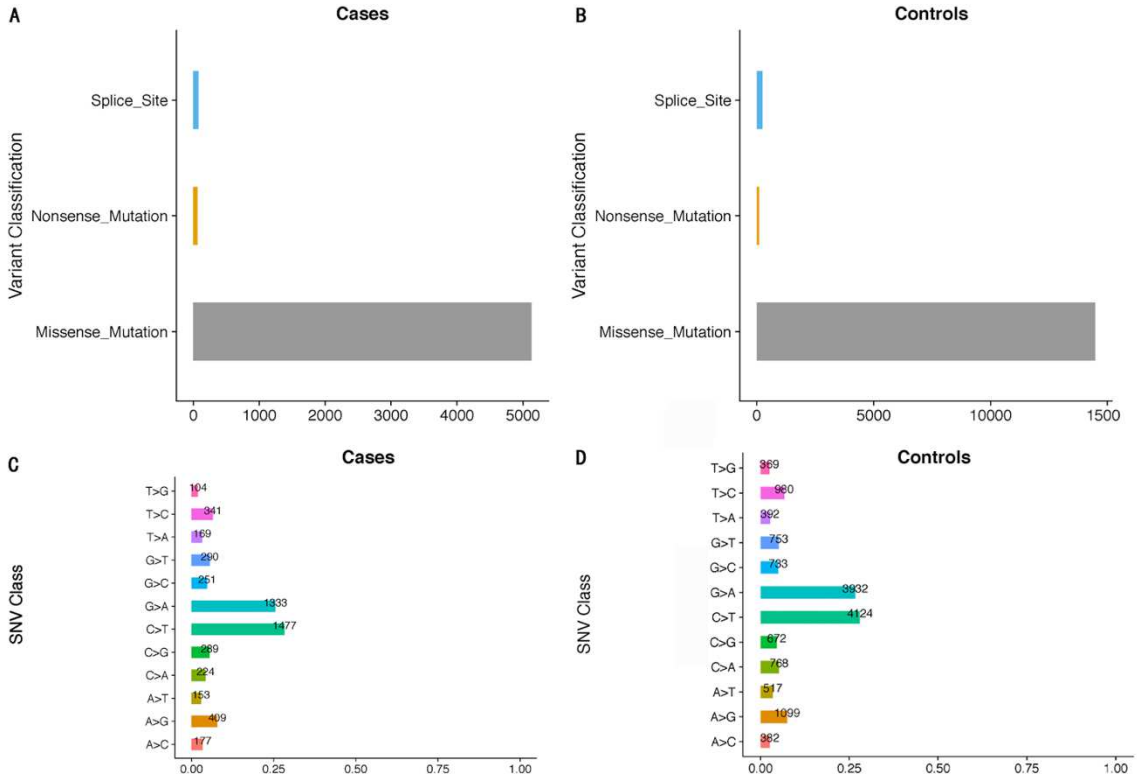


Figure 2



Variants identified based on Fisher exact test

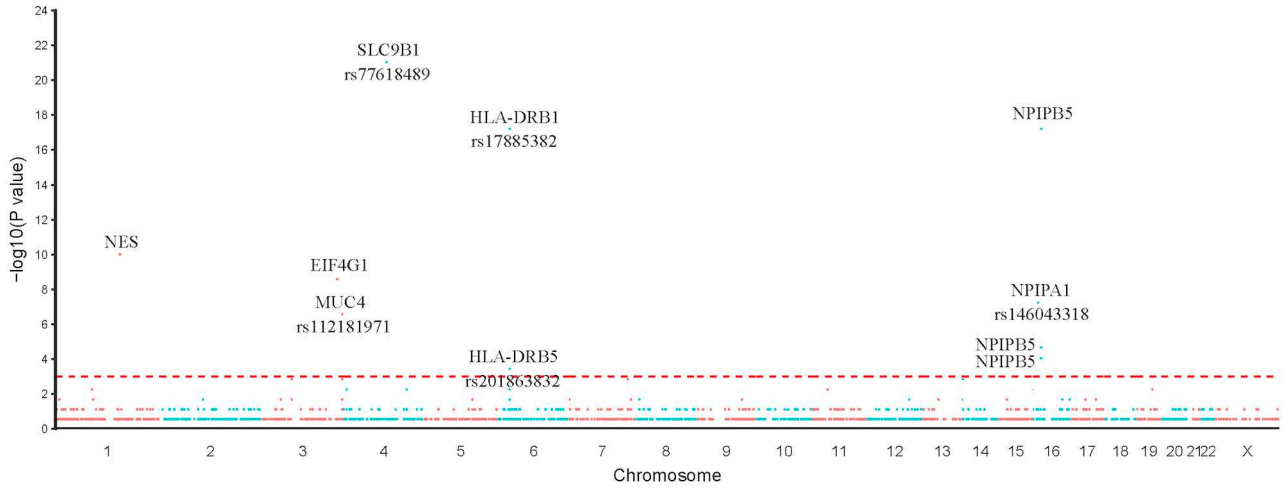
To investigate the genetic cause of PDA, we next used Fisher exact test to the p-value of allele frequency for each SNP between case and control group. Then we genotype these snps and identified 44 candidate snps located within genes, based on an empirical false discovery rate (FDR) of 0.05 or P-valve 0.05 (Table 2). And we prioritized the variants based on Fisher exact test and showed the top ten snps with statistically significant in Figure 3. Notably, we found that snp rs103826685 and snp rs32552095 located in gene SLC9B1 and HLA-DRB1 were significantly enriched($P<0.0001$). Comparing approach with the single-point analysis, carried out with Fisher exact test, we obtain that SNP rs103826685 is the most significantly associated.

Table 2: SNP Fliting Based on Fisher Exact Test

Chromosome	Gene	Position	Case mutation	Case normal	Control mutation	Control normal	p-value	Variant type
1	LRRC8C	90179703	4	35	0	100	0.006	T>G
1	NES	156640657	16	23	0	100	0.000	A>C
11	LRRC4C	40136434	4	35	0	100	0.006	C>T
14	SLC7A8	23612372	5	34	0	100	0.001	T>G
16	SOX8	1034733	4	35	0	100	0.006	A>C
16	NPIPA1	15045634	12	27	0	100	0.000	T>C
16	NPIPB5	22545658	8	31	0	100	0.000	A>C
16	NPIPB5	22546505	25	14	0	100	0.000	G>T
16	NPIPB5	22546506	7	32	0	100	0.000	A>C
19	MAP3K10	40719910	4	35	0	100	0.006	C>G
3	ZNF717	75786264	5	34	0	100	0.001	G>T
3	EIF4G1	184033621	14	25	0	100	0.000	G>C
3	MUC4	195506722	5	34	0	100	0.001	G>A
3	MUC4	195506723	5	34	0	100	0.001	T>G
3	MUC4	195514174	11	28	0	100	0.000	G>T
4	USP17L20	9217567	4	35	0	100	0.006	C>A
4	USP17L17	9246041	4	35	0	100	0.006	C>A
4	SLC9B1	103826685	29	10	0	100	0.000	T>A

4	LRBA	151770608	4	35	0	100	0.006	A>C
6	VARS	31746821	4	35	0	100	0.006	G>A
6	HLA-DRB5	32487344	6	33	0	100	0.000	T>C
6	HLA-DRB1	32552095	25	14	0	100	0.000	C>T
7	TCAF2	143400090	5	34	0	100	0.001	G>A

Figure 3



Genes identified based on Burden analysis

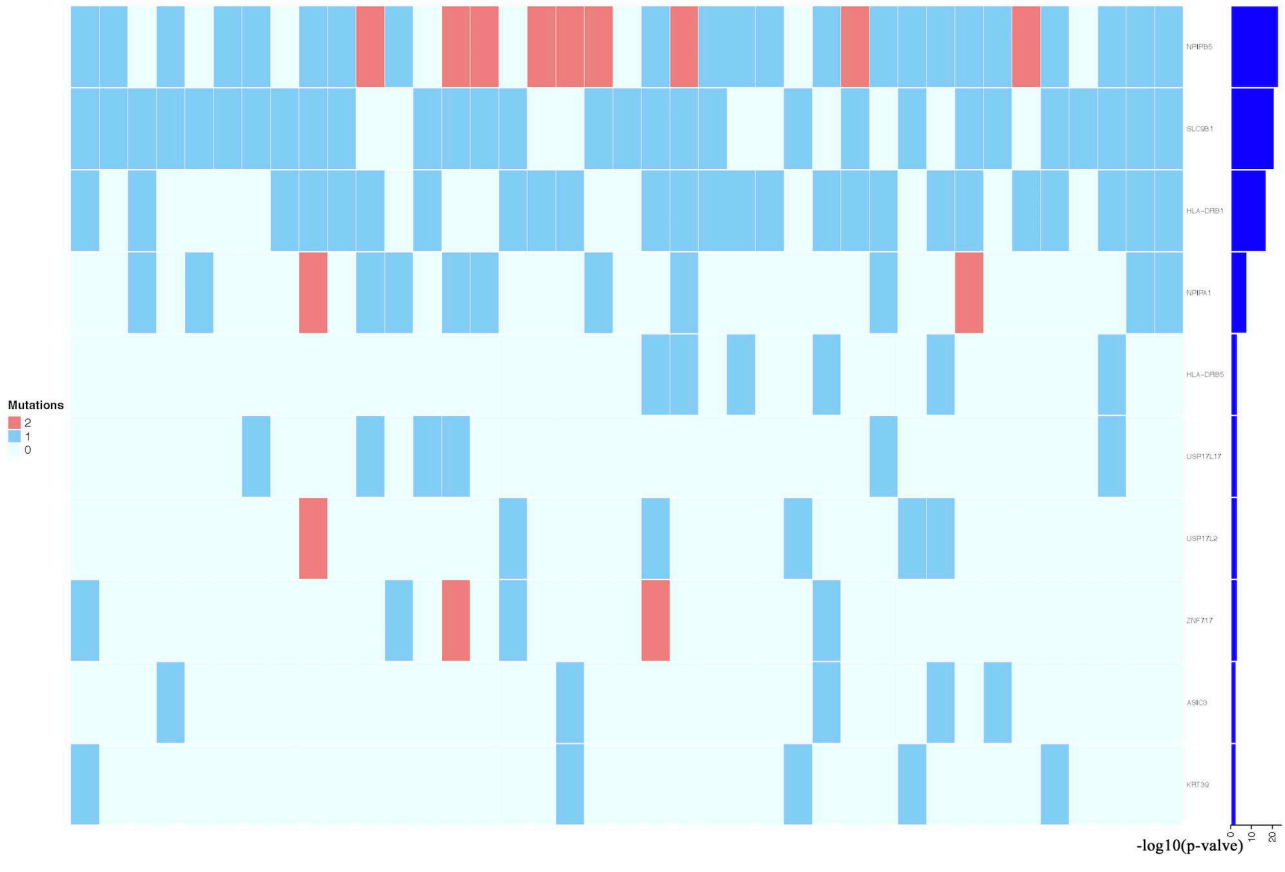
To further increase statistical power, we aggregated the SNP data at the gene level and performed burden analysis. Given thresholds of 0.05 for P-value, we identified 57 genes with potential pathogenicity as PDA-associated (Table 3). Subsequently, we prioritized these genes based on Burden analysis and showed the top ten genes with statistically significant in Heatmap (Figure 4). Among these genes, we found that NPIPB5, SLC9B1, and HLA-DRB1 were considered as the top three with high confidence. Notably, SLC9B1, and HLA-DRB1 was also the most significant based on Fisher exact test.

Table 3: Gene fliting based on Burden analysis

Gene	Case mutation	Case normal	Control mutation	Control normal	p-value
ASIC3	5	34	0	100	0.001
CFAP45	4	35	0	100	0.006
CYP21A2	4	35	0	100	0.006
EVI5	4	35	0	100	0.006
HIPK1	4	35	0	100	0.006
HLA-DRB1	25	14	0	100	0.000
HLA-DRB5	6	33	0	100	0.000
KRT39	5	34	0	100	0.001
LRRC4C	4	35	0	100	0.006
MAP3K10	4	35	0	100	0.006
NPIPA1	13	26	0	100	0.000
NPIPB5	31	8	0	100	0.000
POTEE	5	34	0	100	0.001
SLC9B1	29	10	0	100	0.000
SLX4	4	35	0	100	0.006
SOX8	4	35	0	100	0.006
TBC1D3F	4	35	0	100	0.006
TCAF2	5	34	0	100	0.001

USP17L11	5	34	0	100	0.001
USP17L17	6	33	0	100	0.000
USP17L18	5	34	0	100	0.001
USP17L2	6	33	0	100	0.000
USP17L20	5	34	0	100	0.001
VARS	4	35	0	100	0.006
ZNF717	6	33	0	100	0.000

Figure 4

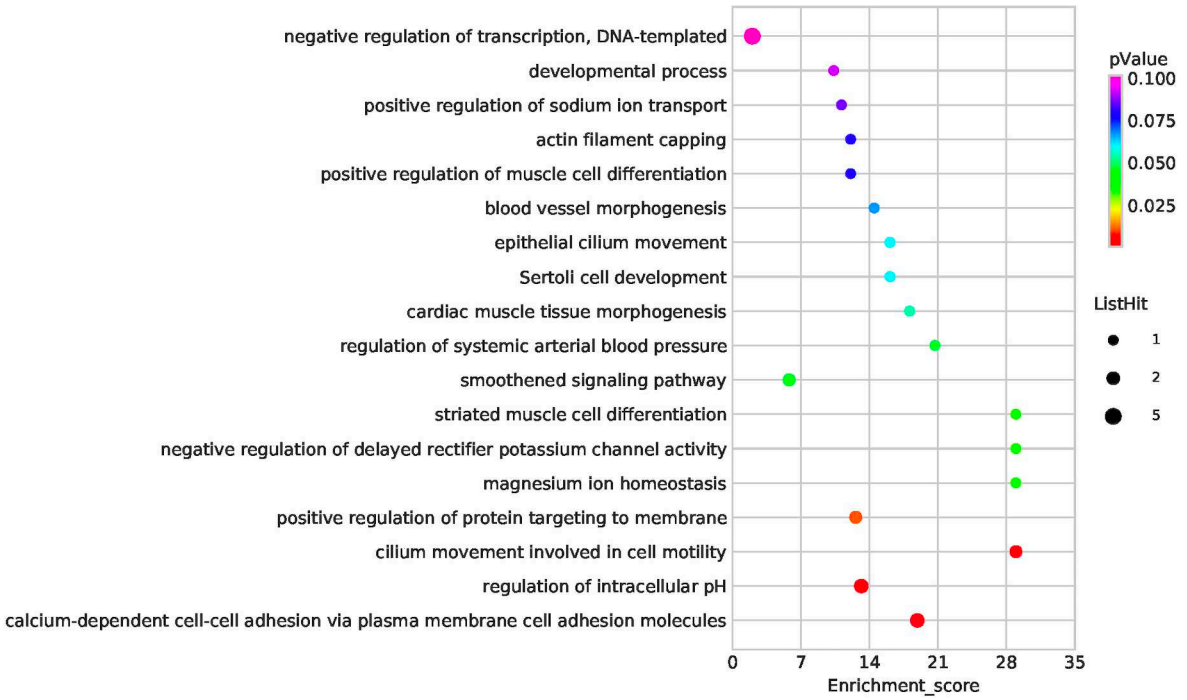


Function analysis

The Gene Ontology (GO) terms ($p < 0.05$) are described as a network of biological processes, which are organized in a way of overlapping in space and clustered according to their relationships[18]. To

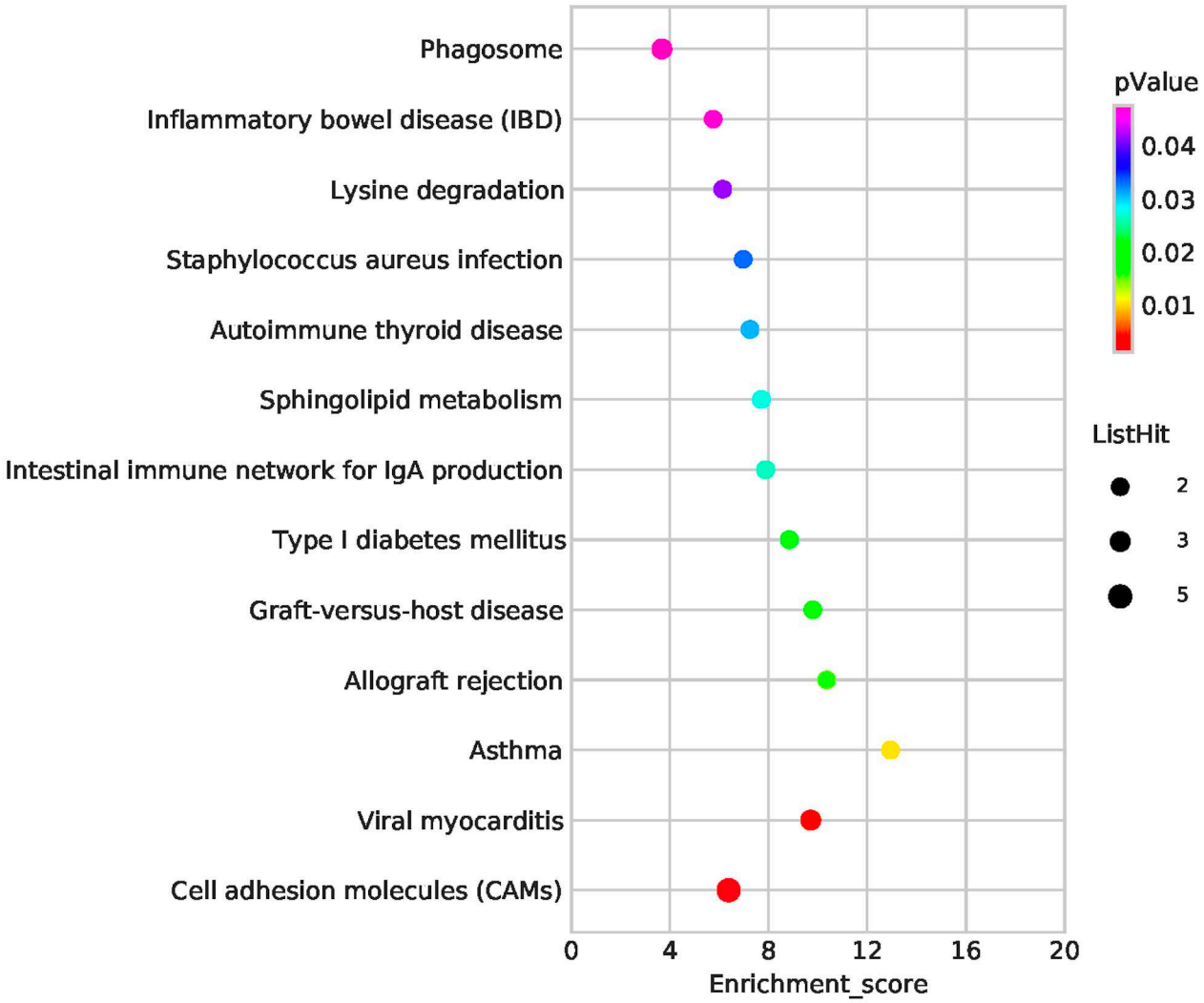
further test the significance of these genes, we conducted extensive genetic functional enrichment analysis. We next analyzed which Gene Ontology (GO) terms and KEGG pathways were enriched in these 101 candidate genes after Fisher exact test and Burden analysis. Functional enrichment analysis of differentially expressed genes revealed that Gene Ontology (GO) terms associated with thiol-dependent ubiquitinyl hydrolase activity (TermID: GO:0036459), peptide antigen binding (TermID: GO:0042605) and ubiquitin-dependent protein catabolic process (TermID: GO:0006511) were highly enriched in the upregulated gene set. A particular focus was placed on terms representing prostaglandin, apoptosis, and heart development. (Figure 5). Moreover, KEGG analysis of the direct gene targets in PDA patients revealed enrichment in pathways related to Cell adhesion molecules (CAMs) (TermID: path: hsa04514, pvalue:0.001), Viral myocarditis (TermID: path: hsa05416, pvalue:0.0035) and Asthma (TermID: path: hsa05310, pvalue:0.01 (Figure 6). Based on the results of functional enrichment analysis, we screened some pathway genes related to cardiovascular development.

Figure 5



193

194 **Figure 6**



Network analysis

To further test the role of gene after functional enrichment analysis, we mapped the 29 our candidate genes based on analysis above and 240 known pathogenic genes to the PPI network (Table S1). The known genes from previous literature were divided into two different gene groups, which are related to cardiac and vascular development and PDA. The result showed that NES and CDH2 had the most direct and obvious relation to known pathogenic genes, both in known CHD related genes and PDA related gene. Moreover, CDH2 and NES have the highest weight and located the center of PPI network (Figure

7, Figure 8). Therefore, based on the degree of correlation, we screened some candidate gene for final verification.

Figure 7

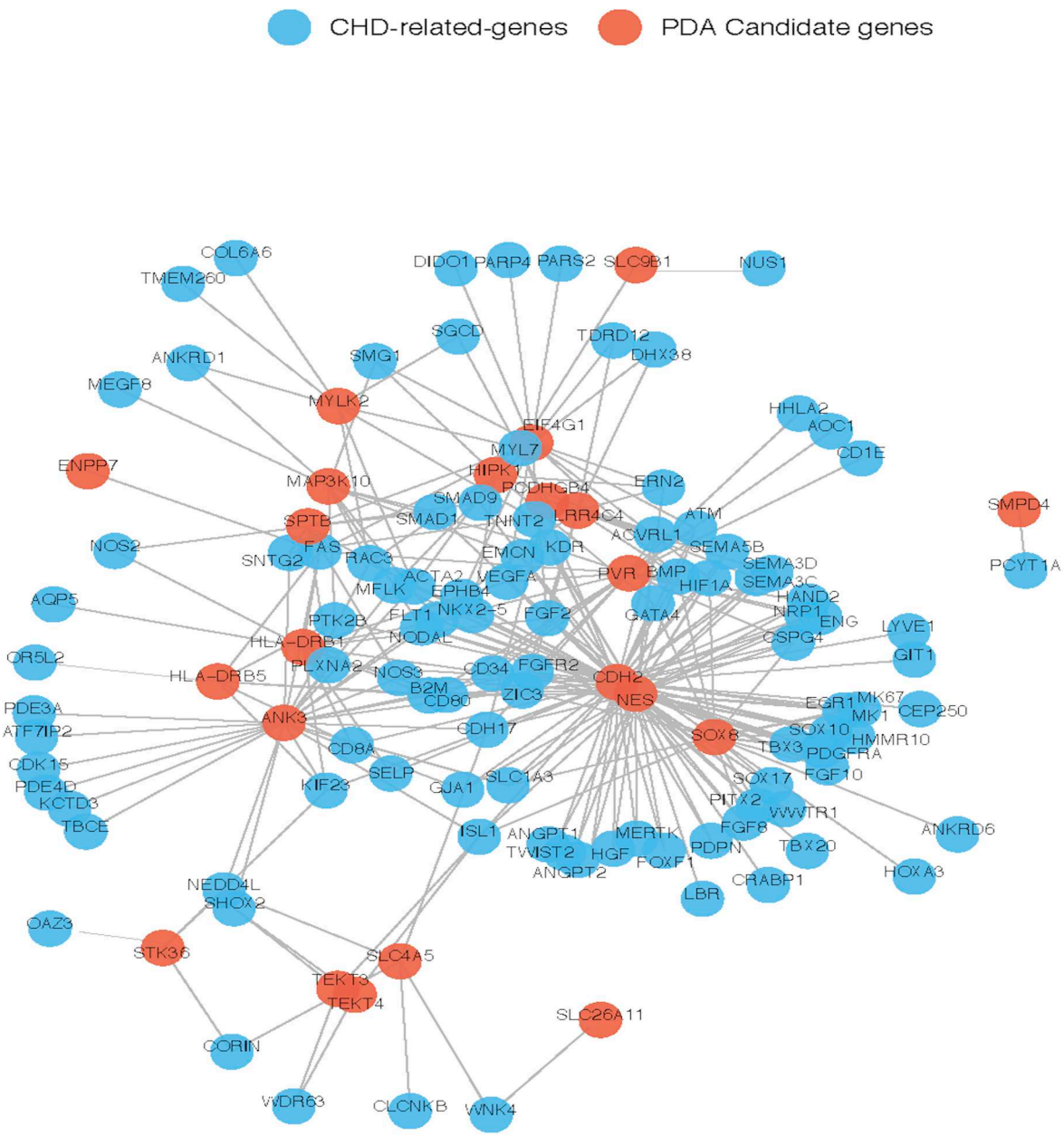


Figure 8

208

209 **Detection of candidate genes expression in human embryonic heart**

210 To further investigate the potential function of our candidate gene expression in human heart, 11 genes

211 were tested expression in human embryonic heart at different Carnegie stages. Then we prioritized those

212 candidate gene according to the amount of expression and identified final 6 pathogenic genes (SOX8, NES,

213 CDH2, ANK3, EIF4G1 and HIPK1) (Figure 9). Among them, we found that CDH2 expressed the most

214 highly in the embryonic heart (Figure 10).

Figure 9

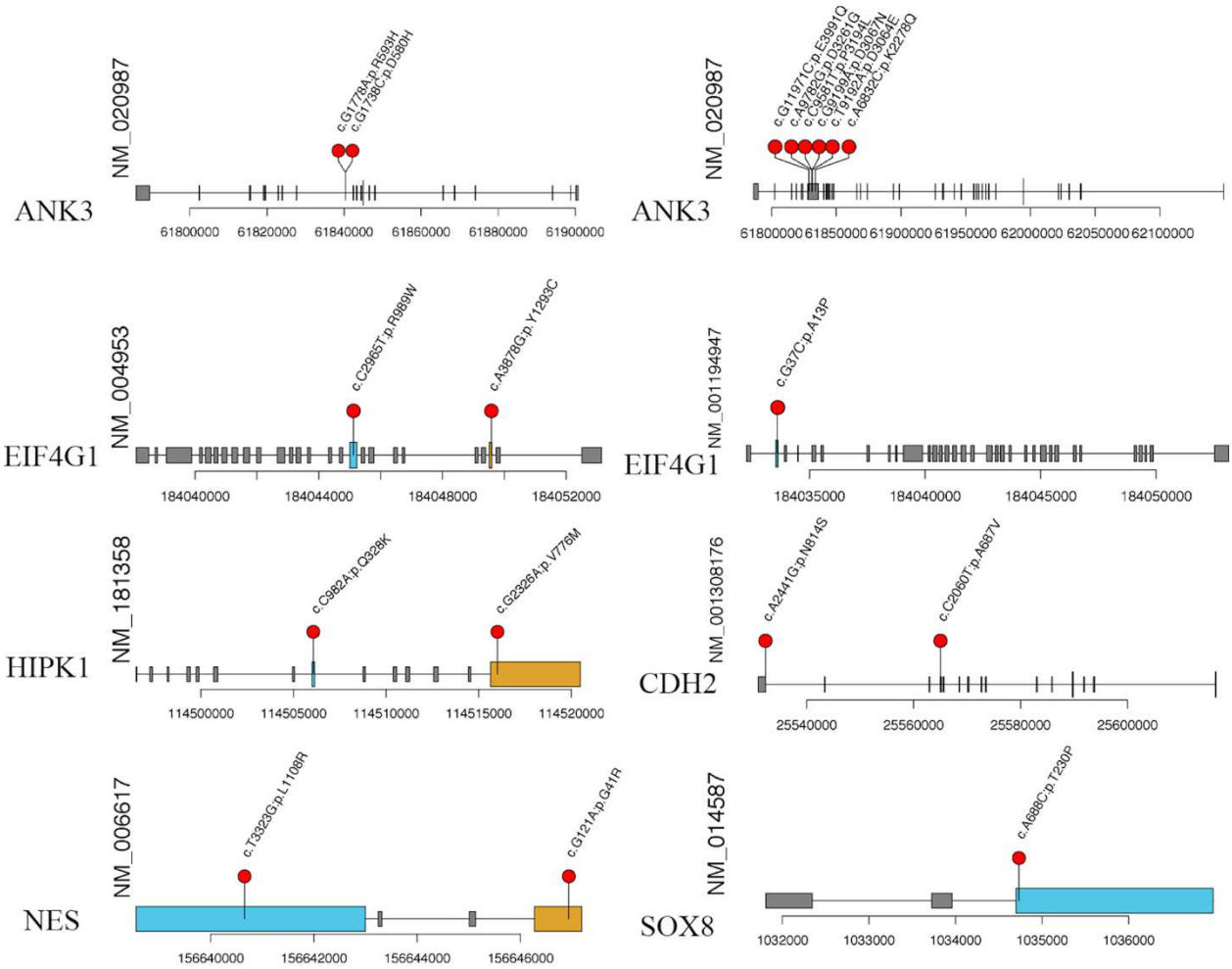
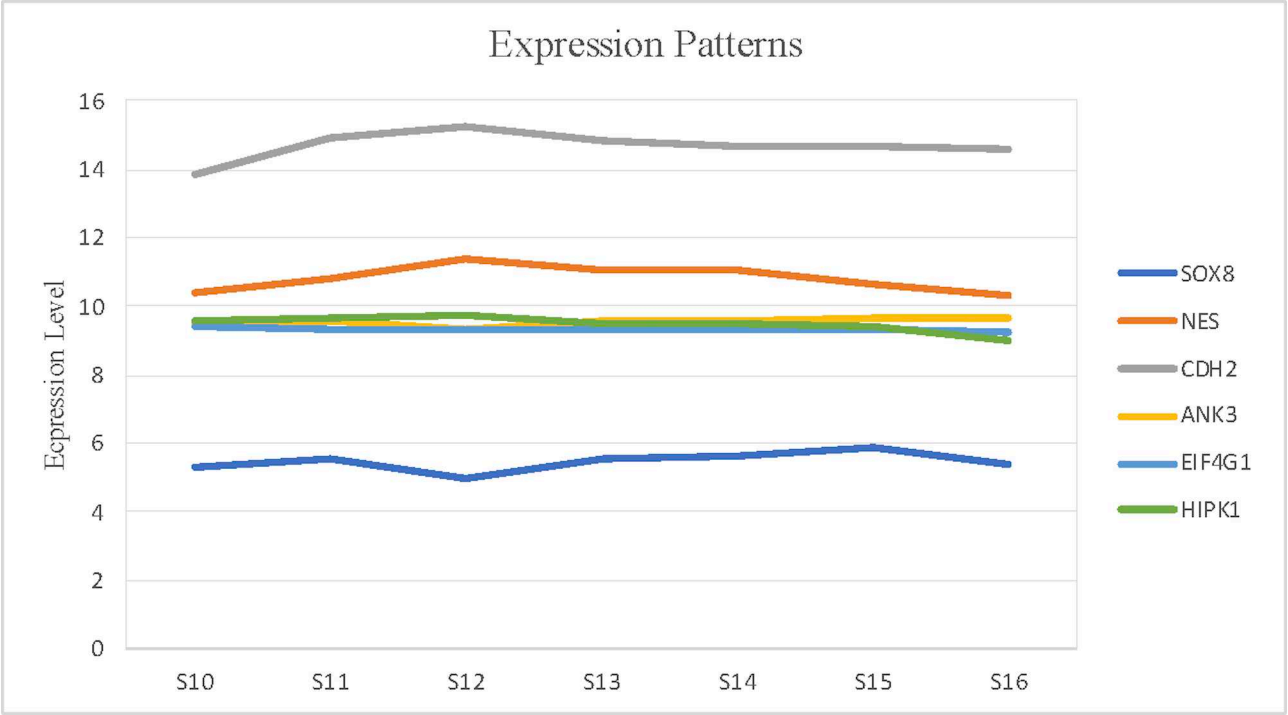


Figure 10



Discussion

As one of the most common congenital heart defects, the underlying molecular genetic mechanisms of PDA are still largely unknown. In this study, we explored the clinical characteristics and performed the WES to identify rare variants and candidate gene in 39 PDA patients and 100 healthy controls. Through a series of bio-information filtering strategies, we prioritized the candidate genes by comprehensively considering factors such Fisher exact test, mutation burden, gene network and expression level. We finally identified 18 rare damaging variants in 6 totally novel candidate genes (SOX8, NES, CDH2, ANK3, EIF4G1 and HIPK1) associated with PDA. In addition, CDH2 expressed the most highly in human embryonic heart and seems to be the most important candidate gene in our study. We hope to perform further study with larger sample size soon.

N-cadherin, encoded by CDH2, is a family of cadherins mediated cell-cell adhesion in multiple tissues. Its structure consists of a single transmembrane domain, a cytoplasmic domain and five conserved extracellular cadherin domains (ECI-V)[22]. In our study, we found two variants (rs25565020,

rs25532304) in CDH2 in 4 patients with PDA. In addition, CDH2 had highest weight and located the centre of PPI network, both in known CHD related genes and PDA related gene. Further investigation showed CDH2 had high expressions in human embryonic hearts. Previous studies in mouse have note the importance of CDH2 in the proper development of the heart, brain, and skeletal structures[23]. Interestingly, genetic analyses in zebrafish showed that mutation in in the EC-I or EC-IV domains of cdh2 resulted the same heart defect phenotypes[24]. Moreover, Mayosi BM et al. used Whole exome sequencing to detect novel rare variant in patients with arrhythmogenic cardiomyopathy and proved that this mutation changes the conservative amino acids of the cadherin 2 protein[25]. Since the relationship between CDH2 and PDA is unclear, additional studies are needed to determine how genetic perturbations of CDH2 contribute to PDA.

In our study, 16 patients (42%) were detected to have same variant(rs156646936) in NES. And in network analysis, we observed a strong correlation between NES and known pathogenic genes. NES belongs to a member of the human tissue kallikrein family of secreted serine proteases[26]. Several studies have confirmed that it plays an important role in carcinogenesis, such as breast, prostate, testicular cancers and leukemia[27]. Further experimental evidence suggests that its function as a tumor suppressor gene may be achieved by hypermethylation of the CpG island of the NES[28]. And no evidence shows mutations of NES gene in PDA. ANK3 belongs to a member of the Ankyrin family that is expressed in several different isoforms in many tissues. And it plays key roles in activities such as cell motility, activation, proliferation, contact, and the maintenance of specialized membrane domains. In our study, 8 patients (10%) were detected to have variants in ANK3. Previous studies have shown that ANK3 variants are associated with schizophrenia, autism, epilepsy and intellectual disability[29, 30]. Studies from knockout mouse models have revealed that loss of

function of ANK3 leads to defects in cardiac calcium handling and arrhythmias[31]. Although the role of NES and ANK3 in the pathogenesis of PDA was supported by bioinformatic analyses, our study was limited by the lack of experimental evidence to validate the deleteriousness of the variants.

EIF4G1 encoded protein which is a component of the multi-subunit protein complex EIF4F. IF4G plays a crucial role in translation initiation, serving as a scaffolding protein that binds several initiation factors (the cap-binding protein eIF4E, the RNA helicase eIF4A, and eIF3)[32]. In our study, 15 patients were detected to have 3 types variants and the same variant(rs184033621) were detected in 14 cases. EIF4G1 modulates the proliferation, apoptosis, angiogenesis of most tumour types by limiting step during the initiation phase of protein synthesis and interacting with Ubiquitin-specific protease 10 (USP10)[33]. Moreover, phosphorylation of EIF4G1 specifically activates PKC-Ras-ERK signaling pathway, which is involved in the control of growth and proliferation[34]. Disease associated with EIF4G1 include Parkinson Disease, non-small cell lung carcinoma, prostate[33]. Although the relationship between EIF4G1 cardiovascular development still unknow, EIF4G1 might be a potentially pathogenic to PDA.

HIPK1 belongs to the Ser/Thr family of protein kinases and HIPK subfamily. Among its related pathways are Regulation of TP53 Activity and Cardiac conduction. Homeodomain interacting protein kinases, HIPK1 and HIPK2, play a key role in embryonic development by regulating TGF-β-dependent angiogenesis[35, 36]. HIPK1 loss-of-function conditional knockout mice exhibit defects in primitive/definitive hematopoiesis, vasculogenesis, angiogenesis and neural tube closure[36]. In addition, HIPK1 can interact with homeobox proteins and other transcription factors to regulate a variety of biological processes, such as signal transduction, apoptosis, embryonic development, retinal vascular

dysfunction[35]. In our study, only two variants (rs114516009, rs114506069) were detected in 4 PDA individuals, novel variants never been reported before. Further investigation showed HIPK1 had high expressions in human embryonic hearts. Additional experiments are needed to determine how genetic mechanism of HIPK1 contribute to PDA.

SOX8 belongs to a member of the SOX (SRY-related HMG-box) family of transcription factors involved in the regulation of embryonic development and in the determination of the cell fate[37]. In our data, the same rare variant (rs1034733) was detected in 3 PDA patients. The expression of SOX8 is essential in the developing heart correlates with heart septation and with the differentiation of the connective tissue of the valve leaflets[38]. Moreover, previous studies revealed that overexpression of Sox8 might associated with hypoxia-induced cell injury by activating the PI3K/AKT/mTOR pathway and MAPK[39]. Interestingly, the closure of DA after birth is closely related to blood oxygenation level and hypoxia can lead to the increase of endogenous PGE2 release, and directly lead to the opening of the ductus arteriosus[1]. Thus far, SOX8 may be a novel candidate gene in the pathogenesis.

Conclusions

Our study did have some limitation. Lack of parental samples and small sample size limited our ability to find the genetic background of PDA. Thus, more fundamental researches are needed to determine our candidate genes contributed to PDA. In conclusion, through a series of bioinformatics filtering steps, we identified 18 rare damaging variants in 6 total novel candidate genes (SOX8, NES, CDH2, ANK3, EIF4G1, HIPK1) associated with PDA. The discovery of these genes opens up a new field for the genetic research of PDA and provides a new idea for understanding the pathogenesis of PDA.

Abbreviations

PDA: Patent Ductus Arteriosus

WES: Whole-exome Sequencing

VSMCs: Vascular smooth muscle cells

DA: Ductus arteriosus

PPI: Protein-protein interaction

GO: Gene Ontology

ASD: Atrial septal defect

VSD: ventricular septal defect

References

1. Benitz WE, Committee on F, Newborn AAoP. Patent Ductus Arteriosus in Preterm Infants. Pediatrics. 2016; 137.

2. Hoffman JI, Kaplan S. The incidence of congenital heart disease. J Am Coll Cardiol. 2002; 39:1890-900.

3. Mitra S, Florez ID, Tamayo ME, Mbuagbaw L, Vanniyasingam T, Veroniki AA et al. Association of Placebo, Indomethacin, Ibuprofen, and Acetaminophen With Closure of Hemodynamically Significant Patent Ductus Arteriosus in Preterm Infants: A Systematic Review and Meta-analysis. JAMA. 2018; 319:1221-38.

4. Crockett SL, Berger CD, Shelton EL, Reese J. Molecular and mechanical factors contributing to ductus arteriosus patency and closure. Congenit Heart Dis. 2019; 14:15-20.

5. Li N, Subrahmanyan L, Smith E, Yu X, Zaidi S, Choi M et al. Mutations in the Histone Modifier PRDM6 Are Associated with Isolated Nonsyndromic Patent Ductus Arteriosus. Am J Hum Genet. 2016; 98:1082-91.

6. Gravholt CH, Viuff MH, Brun S, Stochholm K, Andersen NH. Turner syndrome: mechanisms and management. Nat Rev Endocrinol. 2019; 15:601-14.

7. Yang D, Liu BC, Luo J, Huang TX, Liu CT. Kartagener syndrome. QJM. 2019; 112:297-8.

8. Groth KA, Skakkebaek A, Host C, Gravholt CH, Bojesen A. Clinical review: Klinefelter syndrome--a clinical update. J Clin Endocrinol Metab. 2013; 98:20-30.

9. Satoda M, Zhao F, Diaz GA, Burn J, Goodship J, Davidson HR et al. Mutations in TFAP2B cause Char syndrome, a familial form of patent ductus arteriosus. Nat Genet. 2000; 25:42-6.

10. Vanlerberghe C, Jourdain AS, Ghoumid J, Frenois F, Mezel A, Vaksmann G et al. Holt-Oram syndrome: clinical and molecular description of 78 patients with TBX5 variants. Eur J Hum Genet. 2019; 27:360-8.

11. Pannone L, Bocchinfuso G, Flex E, Rossi C, Baldassarre G, Lissewski C et al. Structural, Functional, and Clinical Characterization of a Novel PTPN11 Mutation Cluster Underlying Noonan Syndrome. Hum Mutat. 2017; 38:451-9.

12. Harakalova M, van der Smagt J, de Kovel CG, Van't Slot R, Poot M, Nijman IJ et al. Incomplete segregation of MYH11 variants with thoracic aortic aneurysms and dissections and patent ductus arteriosus. Eur J Hum Genet. 2013; 21:487-93.

335 13. Erdogan F, Larsen LA, Zhang L, Tumer Z, Tommerup N, Chen W et al. High frequency of
336 submicroscopic genomic aberrations detected by tiling path array comparative genome
337 hybridisation in patients with isolated congenital heart disease. J Med Genet. 2008;
338 45:704-9.
339 14. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.
340 Bioinformatics. 2009; 25:1754-60.
341 15. Okonechnikov K, Conesa A, Garcia-Alcalde F. Qualimap 2: advanced multi-sample
342 quality control for high-throughput sequencing data. Bioinformatics. 2016; 32:292-4.
343 16. Gezsi A, Bolgar B, Marx P, Sarkozy P, Szalai C, Antal P. VariantMetaCaller: automated
344 fusion of variant calling pipelines for quantitative, precision-based filtering. BMC
345 Genomics. 2015; 16:875.
346 17. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J et al. Standards and guidelines
347 for the interpretation of sequence variants: a joint consensus recommendation of the
348 American College of Medical Genetics and Genomics and the Association for Molecular
349 Pathology. Genet Med. 2015; 17:405-24.
350 18. Gene Ontology C. Gene Ontology Consortium: going forward. Nucleic Acids Res. 2015;
351 43:D1049-56.
352 19. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on
353 genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017; 45:D353-D61.
354 20. Brohee S, Faust K, Lima-Mendez G, Vanderstocken G, van Helden J. Network Analysis
355 Tools: from biological networks to clusters and pathways. Nat Protoc. 2008; 3:1616-29.

21. O'Rahilly R. Human embryo. Nature. 1987; 329:385.

22. Alimperti S, Andreadis ST. CDH2 and CDH11 act as regulators of stem cell fate decisions. Stem Cell Res. 2015; 14:270-82.

23. Radice GL, Rayburn H, Matsunami H, Knudsen KA, Takeichi M, Hynes RO. Developmental defects in mouse embryos lacking N-cadherin. Dev Biol. 1997; 181:64-78.

24. Masai I, Lele Z, Yamaguchi M, Komori A, Nakata A, Nishiwaki Y et al. N-cadherin mediates retinal lamination, maintenance of forebrain compartments and patterning of retinal neurites. Development. 2003; 130:2479-94.

25. Mayosi BM, Fish M, Shaboodien G, Mastantuono E, Kraus S, Wieland T et al. Identification of Cadherin 2 (CDH2) Mutations in Arrhythmogenic Right Ventricular Cardiomyopathy. Circ Cardiovasc Genet. 2017; 10.

26. Luo L, Herbrick JA, Scherer SW, Beatty B, Squire J, Diamandis EP. Structural characterization and mapping of the normal epithelial cell-specific 1 gene. Biochem Biophys Res Commun. 1998; 247:580-6.

27. Luo LY, Rajpert-De Meyts ER, Jung K, Diamandis EP. Expression of the normal epithelial cell-specific 1 (NES1; KLK10) candidate tumour suppressor gene in normal and malignant testicular tissue. Br J Cancer. 2001; 85:220-4.

28. Li B, Goyal J, Dhar S, Dimri G, Evron E, Sukumar S et al. CpG methylation as a basis for breast tumor-specific loss of NES1/kallikrein 10 expression. Cancer Res. 2001; 61:8014-21.

29. Leussis MP, Berry-Scott EM, Saito M, Jhuang H, de Haan G, Alkan O et al. The ANK3 bipolar disorder gene regulates psychiatric-related behaviors that are modulated by lithium and stress. Biol Psychiatry. 2013; 73:683-90.

30. Wirgenes KV, Tesli M, Inderhaug E, Athanasiu L, Agartz I, Melle I et al. ANK3 gene expression in bipolar disorder and schizophrenia. Br J Psychiatry. 2014; 205:244-5.

31. Mohler PJ, Splawski I, Napolitano C, Bottelli G, Sharpe L, Timothy K et al. A cardiac arrhythmia syndrome caused by loss of ankyrin-B function. Proc Natl Acad Sci U S A. 2004; 101:9137-42.

32. Haimov O, Sehrawat U, Tamarkin-Ben Harush A, Bahat A, Uzonyi A, Will A et al. Dynamic Interaction of Eukaryotic Initiation Factor 4G1 (eIF4G1) with eIF4E and eIF1 Underlies Scanning-Dependent and -Independent Translation. Mol Cell Biol. 2018; 38.

33. Cao Y, Wei M, Li B, Liu Y, Lu Y, Tang Z et al. Functional role of eukaryotic translation initiation factor 4 gamma 1 (EIF4G1) in NSCLC. Oncotarget. 2016; 7:24242-51.

34. Dobrikov M, Dobrikova E, Shveygert M, Gromeier M. Phosphorylation of eukaryotic translation initiation factor 4G1 (eIF4G1) by protein kinase C{alpha} regulates eIF4G1 binding to Mnk1. Mol Cell Biol. 2011; 31:2947-59.

35. Aikawa Y, Nguyen LA, Isono K, Takakura N, Tagata Y, Schmitz ML et al. Roles of HIPK1 and HIPK2 in AML1- and p300-dependent transcription, hematopoiesis and blood vessel formation. EMBO J. 2006; 25:3955-65.

396 36. Shang Y, Doan CN, Arnold TD, Lee S, Tang AA, Reichardt LF et al. Transcriptional
397 corepressors HIPK1 and HIPK2 control angiogenesis via TGF-beta-TAK1-dependent
398 mechanism. PLoS Biol. 2013; 11:e1001527.
399 37. Haseeb A, Lefebvre V. The SOXE transcription factors-SOX8, SOX9 and SOX10-share a
400 bi-partite transactivation mechanism. Nucleic Acids Res. 2019; 47:6917-31.
401 38. Montero JA, Giron B, Arrechedera H, Cheng YC, Scotting P, Chimal-Monroy J et al.
402 Expression of Sox8, Sox9 and Sox10 in the developing valves and autonomic nerves of
403 the embryonic heart. Mech Dev. 2002; 118:199-202.
404 39. Gong LC, Xu HM, Guo GL, Zhang T, Shi JW, Chang C. Long Non-Coding RNA H19
405 Protects H9c2 Cells against Hypoxia-Induced Injury by Targeting MicroRNA-139. Cell
406 Physiol Biochem. 2017; 44:857-69.

407 Figure legends

408 **Figure 1: Bioinformatics filtering strategy workflow for the candidate genes.** Through a series of
409 filtering methods, we finally identified 6 candidate genes. The potentially damaging variants in candidate
410 genes were subjected to validation via human embryonic heart expression analysis.

411 **Figure 2: The comparisons of the rare damaging variants between the PDA and control groups.**
412 The number of variants in each variant classification and SNV class between cases and controls are
413 presented in (A), (B), (C) and (D), respectively.

414 **Figure 3: Single SNP allele frequency and genotype frequency p-values were obtained using the**
415 **fisher exact test.** X-axis represents the position of each snp (represented in circles) on human

chromosome, Y-axis is the –log P-value of Fisher Exact test. Top ten variants in our study were represented in the figure.

Figure 4: Heatmap representing the top 10 genes identified in Burden analysis. Heatmap that shows the mutational burden (P-value< 0.05) of the top ten gene based on gene-based burden analysis in PDA patients. The heatmap was generated by using R package, the mutation values were normalized per gene over all PDA samples. Each box in the heatmap represent a single variant in a case, with the dark red indicating high gene mutation ration in gene-based Burden analysis.

Figure 5: Bubble plot of the GO analysis. Bubble plot summarizing enrichment for the most significant biological process GO terms associated to differentially expressed genes. The bubble size indicates the frequency of the GO term, while the color indicates the P-value.

Figure 6: **Bubble plot of the KEGG pathway analysis.** The representative enriched pathways shown by KEGG analysis. The bubble size indicates the frequency of the KEGG term, while the color indicates the P-value.

Figure 7: Interaction between our candidate genes and known CHD-related genes. PPI network was generated by Cytoscape software and our candidate pathogenic genes and the known CHD-related genes were uploaded in STRING database. Each node represents one gene, and each edge represents the protein-protein interaction collected from BioGRID.

Figure 8: Interaction between our candidate genes and known PDA-related genes. PPI network was generated by Cytoscape software and Our candidate pathogenic genes and the known CHD-related genes were uploaded in STRING database. Each node represents one gene, and each edge represents the protein-protein interaction collected from BioGRID.

Figure 9: The specific amino acid sites of variants of our candidate gene. The red balls represent the location of rare variant on the encoded proteins or protein domains.

Figure 10: Expression of candidate genes in human embryonic heart. The expression patterns of candidate genes in human embryonic heart at different stages of S10 to S16 were analyzed by microarray. X-axis represents the different stages of human embryonic heart, while the Y-axis indicates the level of gene expression.

Declarations

Ethics approval and consent to participate

The studies involving human participants were reviewed and approved by the Medical Ethics Committee of Xinhua Hospital (Approval No. XHEC-D-2020-001). Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Consent for publication

Not applicable.

Availability of data and materials

The datasets supporting the conclusions of this article are available in the NCBI SRA repository and have been compiled into the following repository for ease of access:

https://www.ncbi.nlm.nih.gov/sra/?term=SRP288538. The accession number is SRP288538.

Competing interests

The authors declare that they have no competing interests.

Author information

Affiliations

Department of Pediatric, Shidong Hospital, Shanghai, China

Ying Gao, Ying Liu

Department of Pediatric Cardiology, Xin Hua Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China

Jiaoyu Li, YingHui Chen, Qi Zhang, Bingyao Zhang, Pengjun Zhao and Bo Chen

Corresponding author

Correspondence to Bo Chen.

Funding

This study received financial supports from National Natural Science Foundation of China (82070386), the Project of Shanghai Municipal Health Commission (Grant No. 201940393).

Authors' contributions

PZ and BC contributed to design of the study and performed the statistical analysis. HY, QZ, and BY collected the blood samples from all subjects. BC and YG wrote the first draft of the manuscript. YG, YL and JY contributed to this study equally. PZ and BC revised the manuscript. All authors contributed to manuscript revision, read and approved the submitted version.

Acknowledgements

Not applicable.

Supplementary information

Additional file 1: Table S1.

Summary of the known genes from previous literature were divided into two different gene groups, which are related to cardiac and vascular development and PDA.

Figures

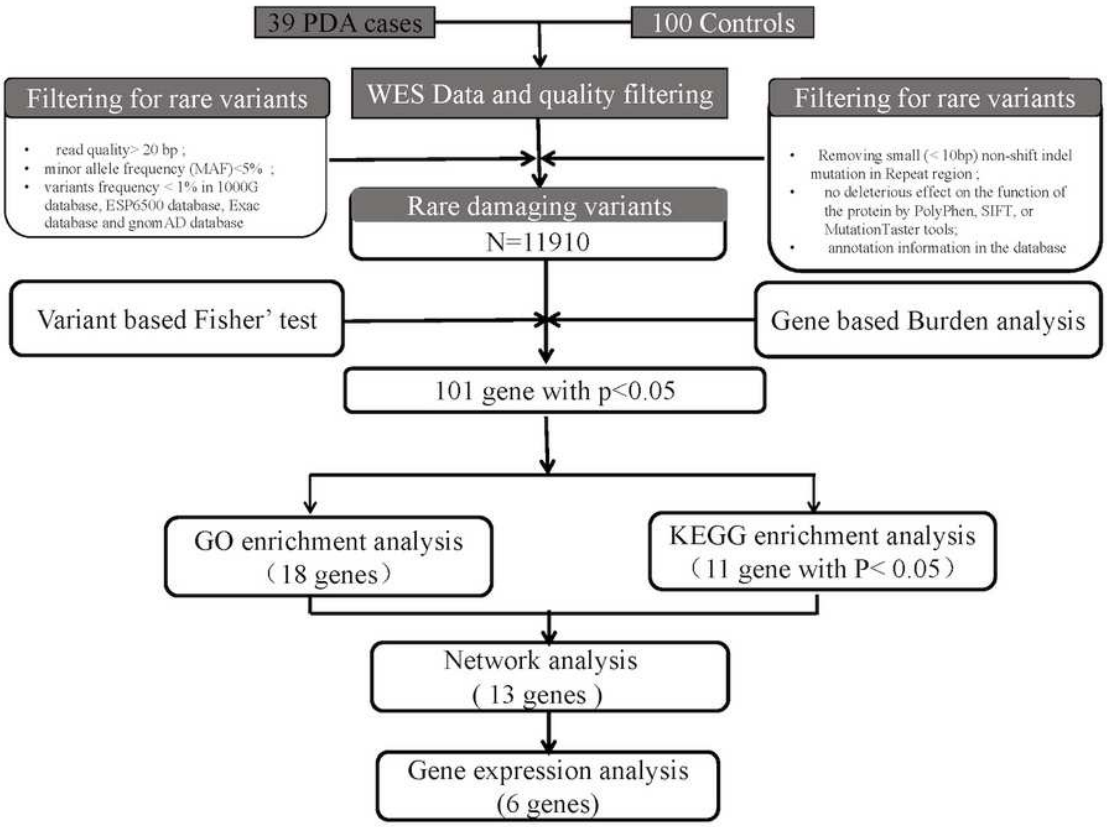


Figure 1

Bioinformatics filtering strategy workflow for the candidate genes. Through a series of filtering methods, we finally identified 6 candidate genes. The potentially damaging variants in candidate genes were subjected to validation via human embryonic heart expression analysis.

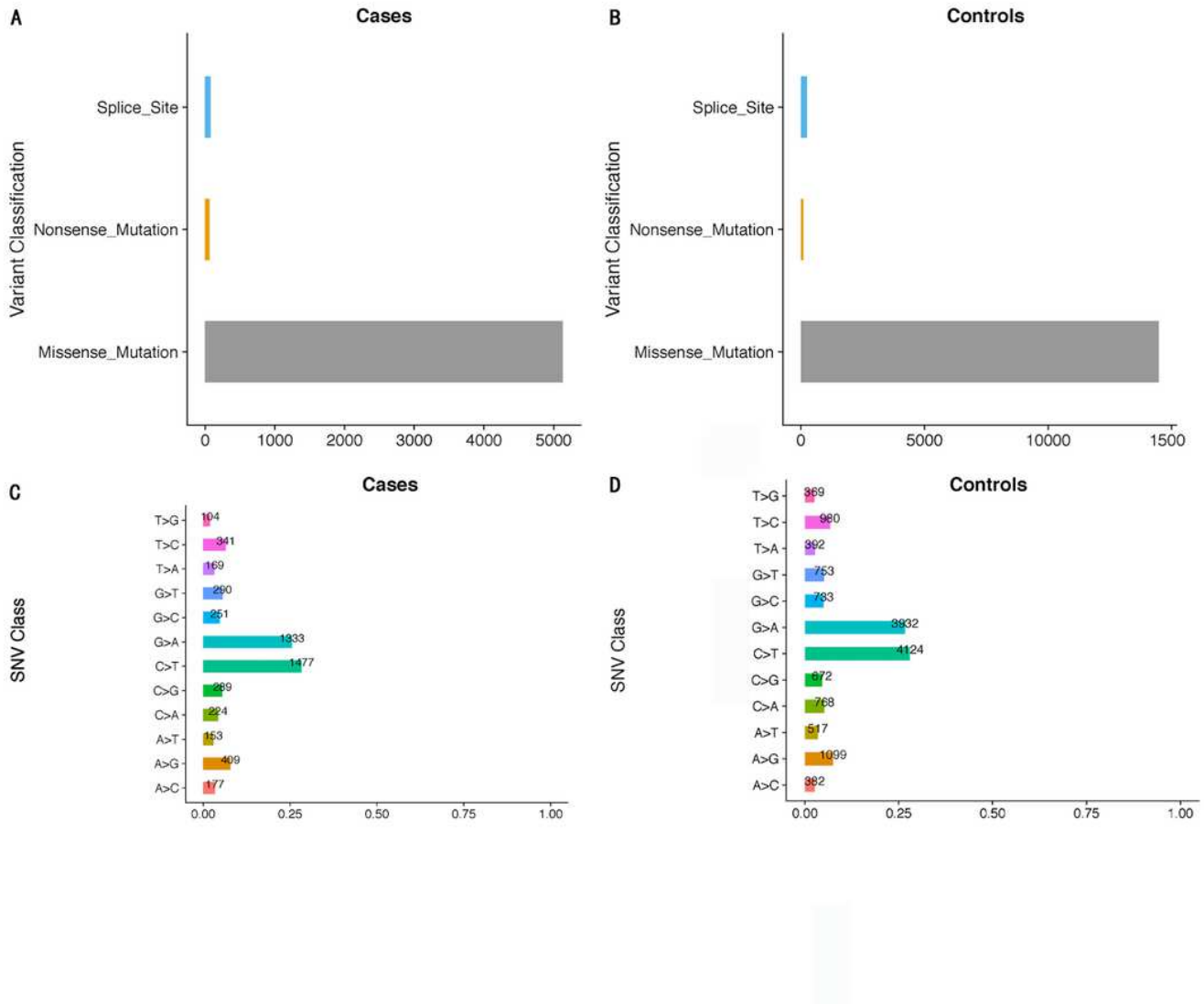


Figure 2

The comparisons of the rare damaging variants between the PDA and control groups. The number of variants in each variant classification and SNV class between cases and controls are presented in (A), (B), (C) and (D), respectively.

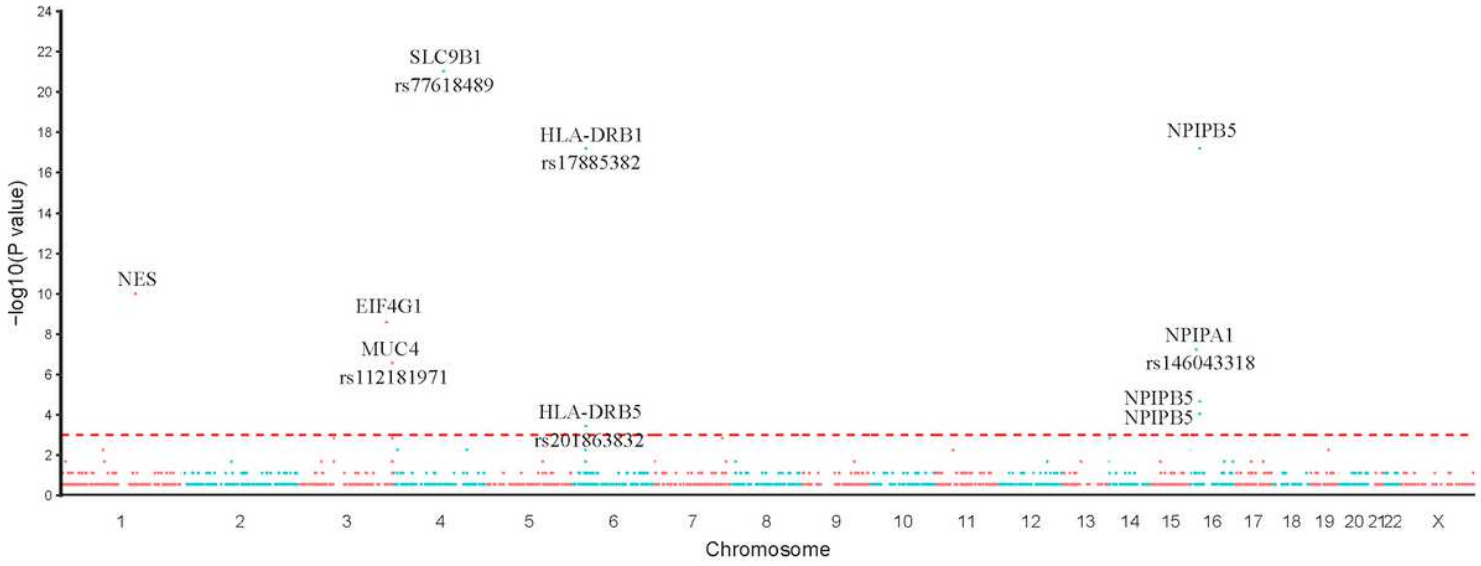


Figure 3

Single SNP allele frequency and genotype frequency p-values were obtained using the fisher exact test. X-axis represents the position of each snp (represented in circles) on human chromosome, Y-axis is the –log P-value of Fisher Exact test. Top ten variants in our study were represented in the figure.

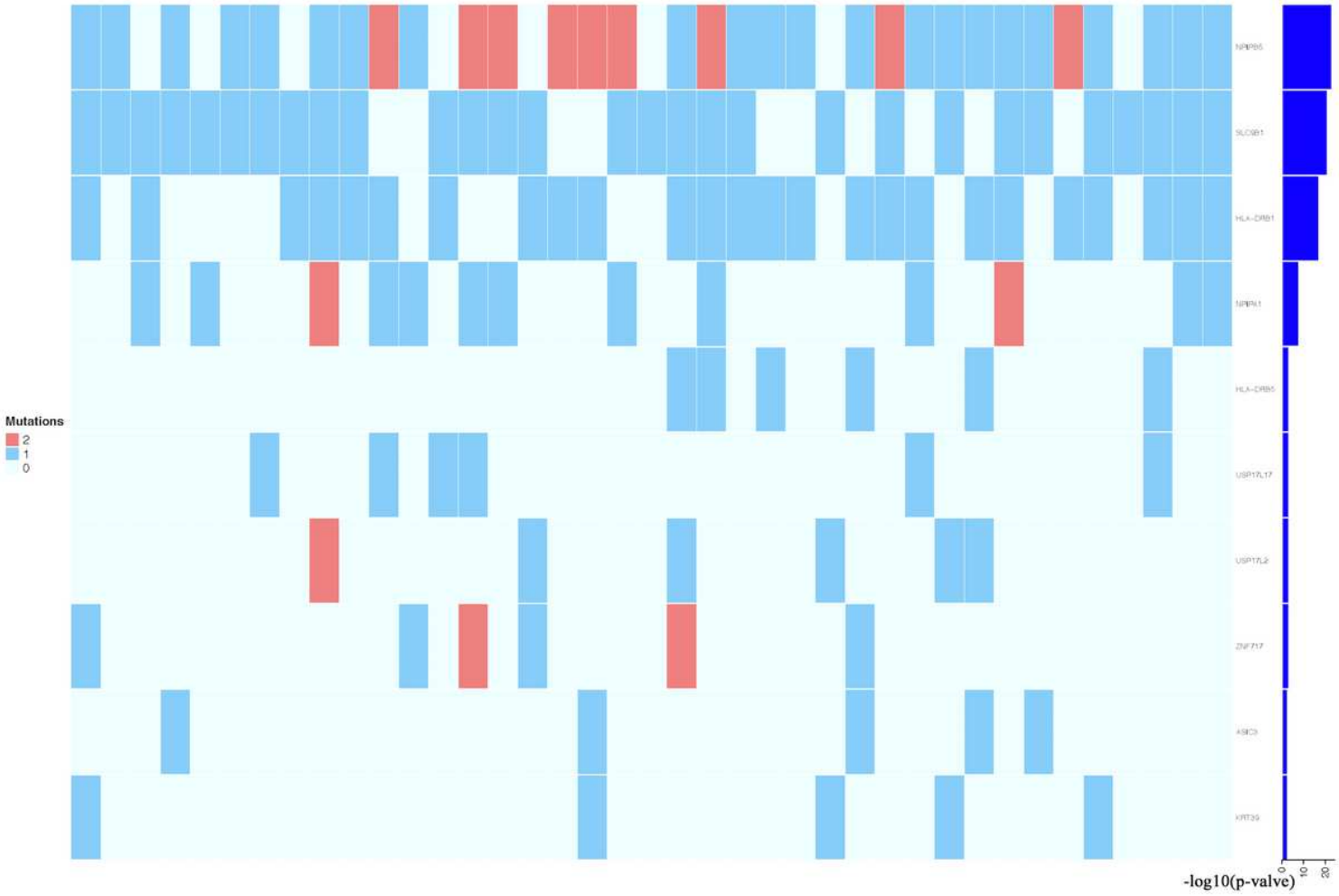


Figure 4

Heatmap representing the top 10 genes identified in Burden analysis. Heatmap that shows the mutational burden (P-value< 0.05) of the top ten gene based on gene-based burden analysis in PDA patients. The heatmap was generated by using R package, the mutation values were normalized per gene over all PDA samples. Each box in the heatmap represent a single variant in a case, with the dark red indicating high gene mutation ration in gene-based Burden analysis.

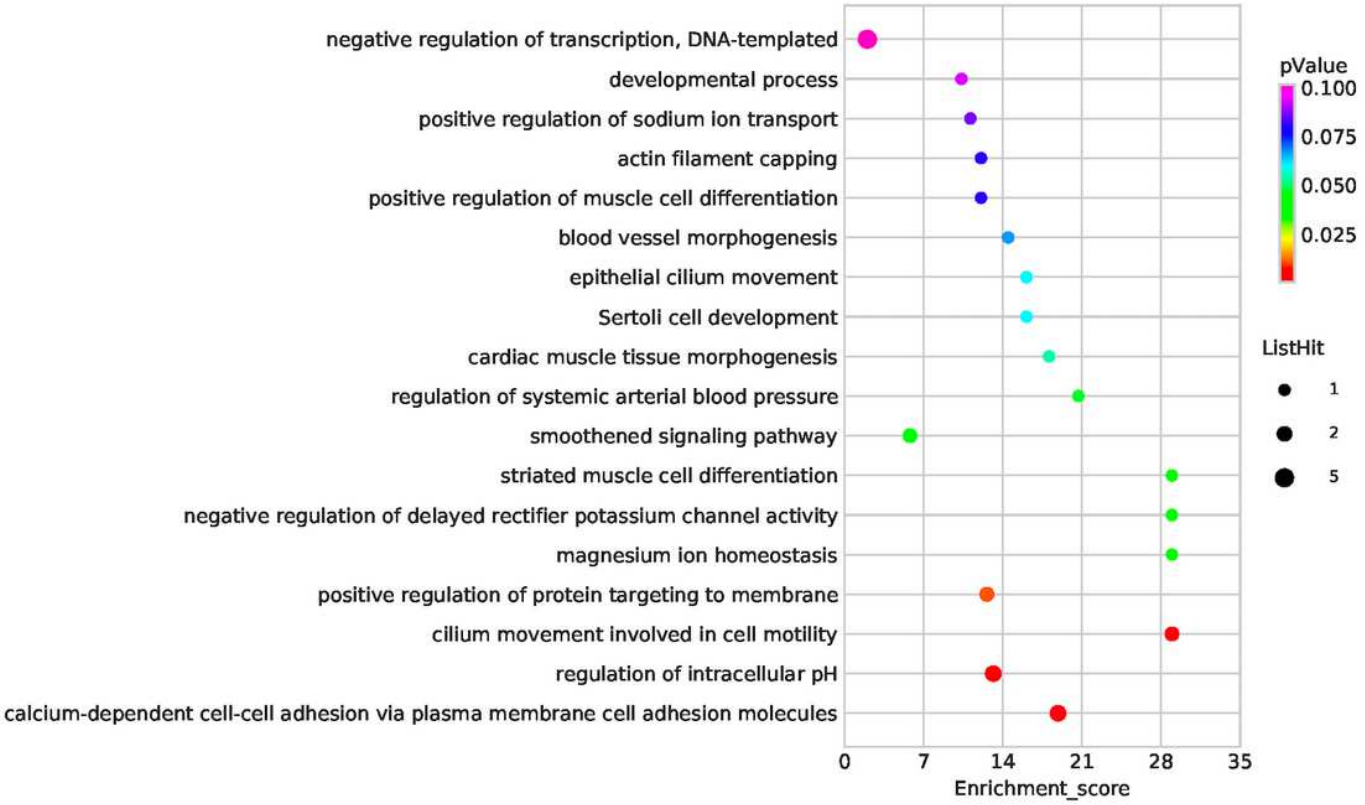


Figure 5

Bubble plot of the GO analysis. Bubble plot summarizing enrichment for the most significant biological process GO terms associated to differentially expressed genes. The bubble size indicates the frequency of the GO term, while the color indicates the P-value.

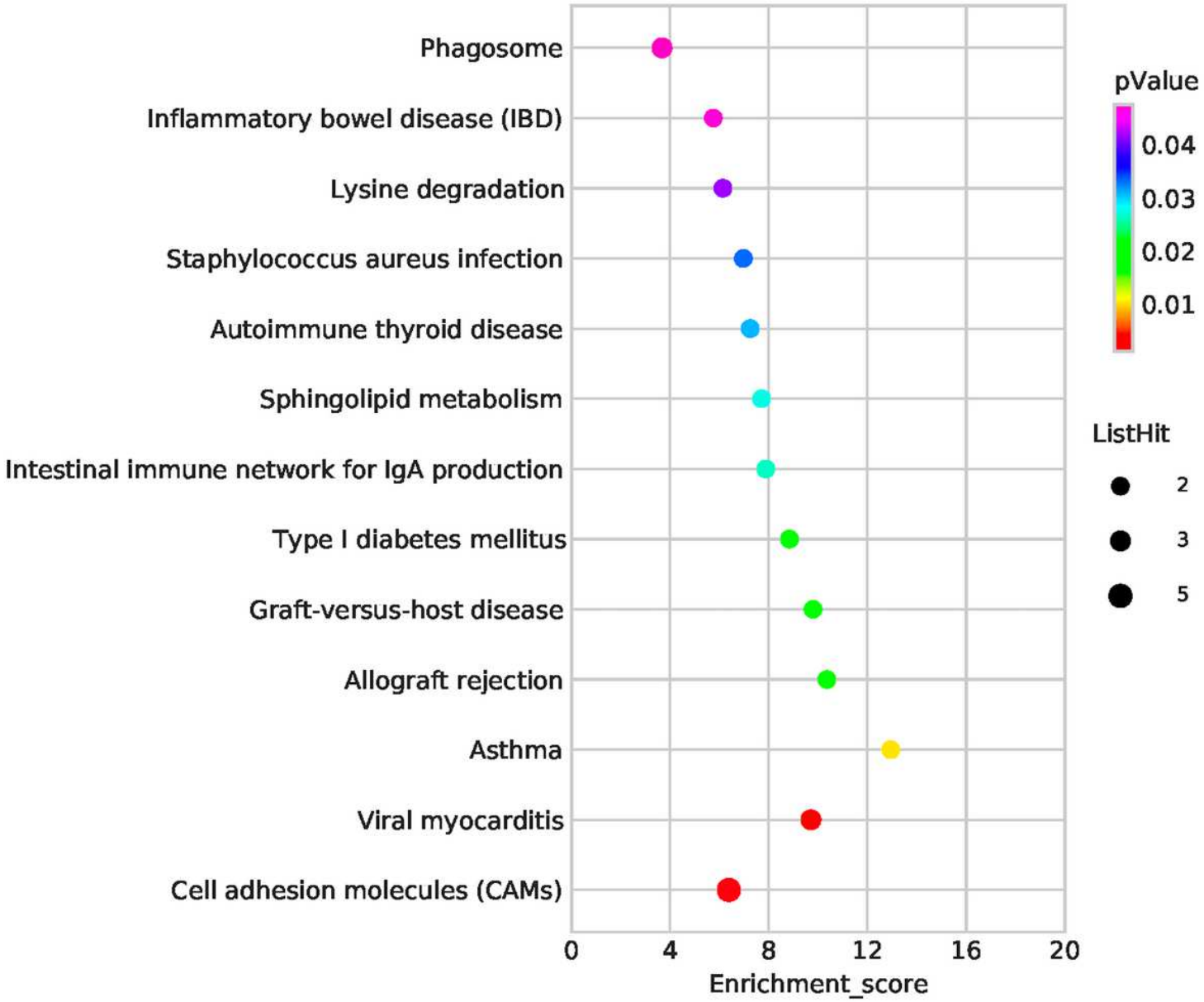


Figure 6

Bubble plot of the KEGG pathway analysis. The representative enriched pathways shown by KEGG analysis. The bubble size indicates the frequency of the KEGG term, while the color indicates the P-value.

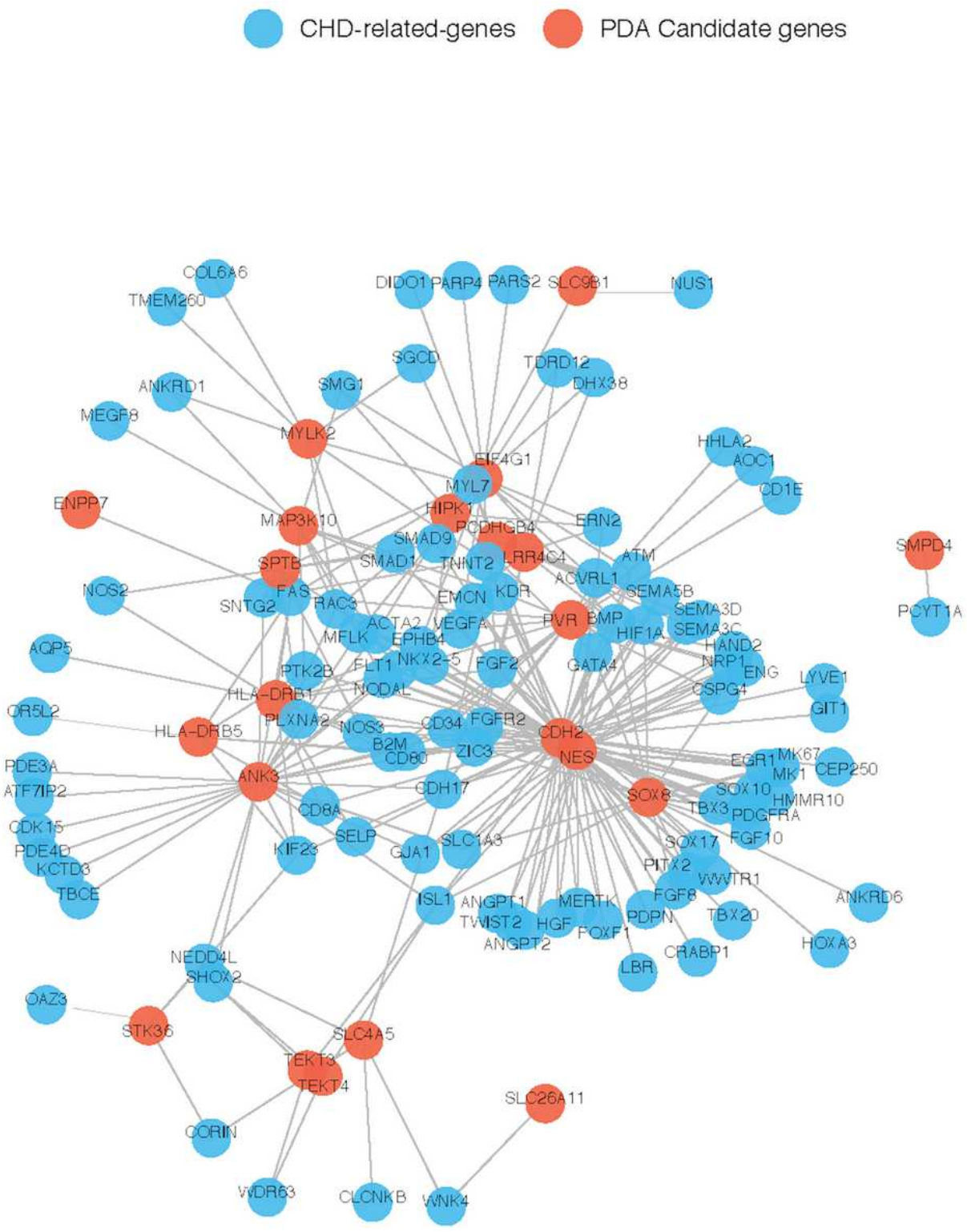


Figure 7

Interaction between our candidate genes and known CHD-related genes. PPI network was generated by Cytoscape software and our candidate pathogenic genes and the known CHD-related genes were uploaded in STRING database. Each node represents one gene, and each edge represents the protein-protein interaction collected from BioGRID.

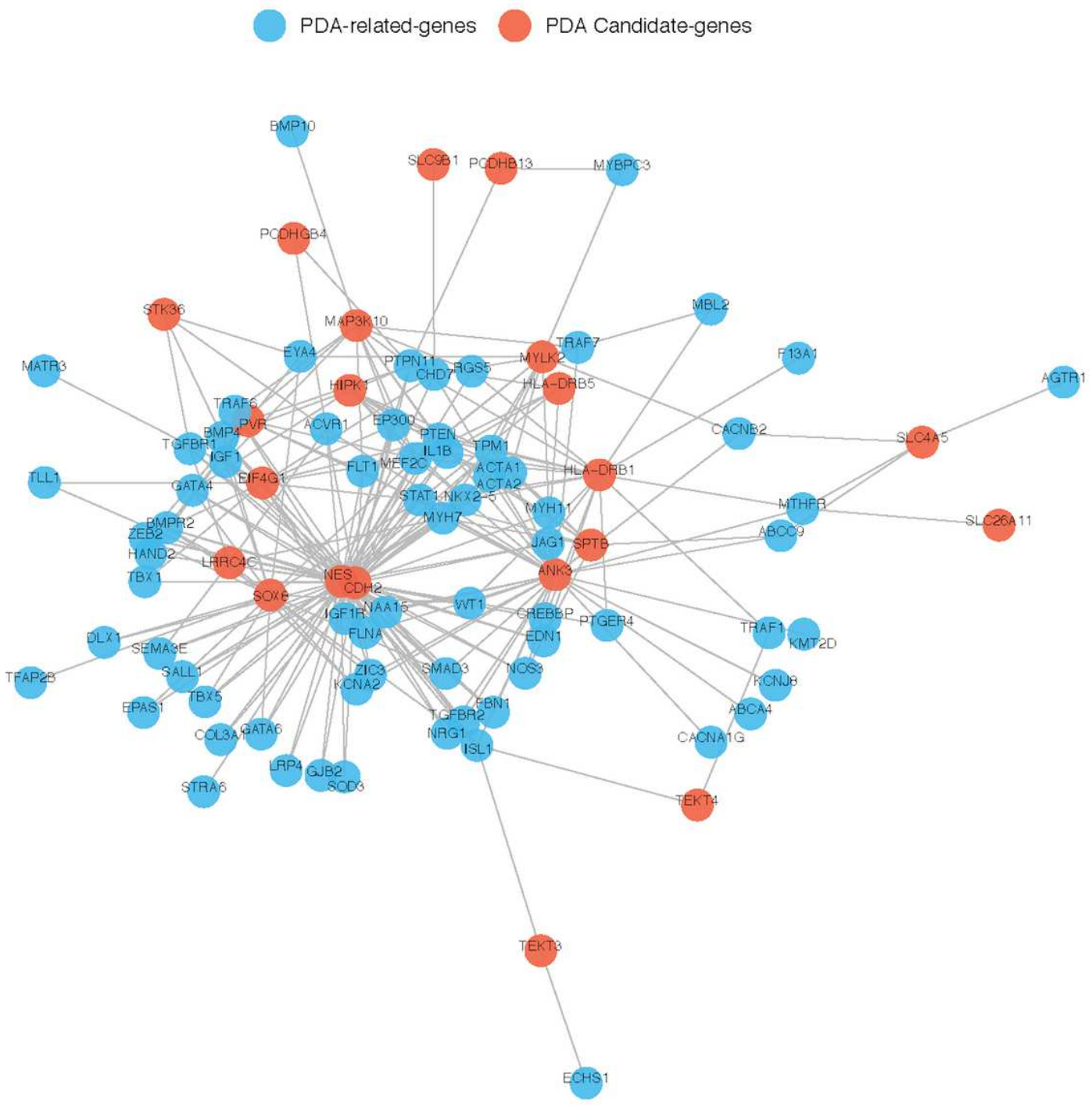


Figure 8

Interaction between our candidate genes and known PDA-related genes. PPI network was generated by Cytoscape software and Our candidate pathogenic genes and the known CHD-related genes were uploaded in STRING database. Each node represents one gene, and each edge represents the protein-protein interaction collected from BioGRID.

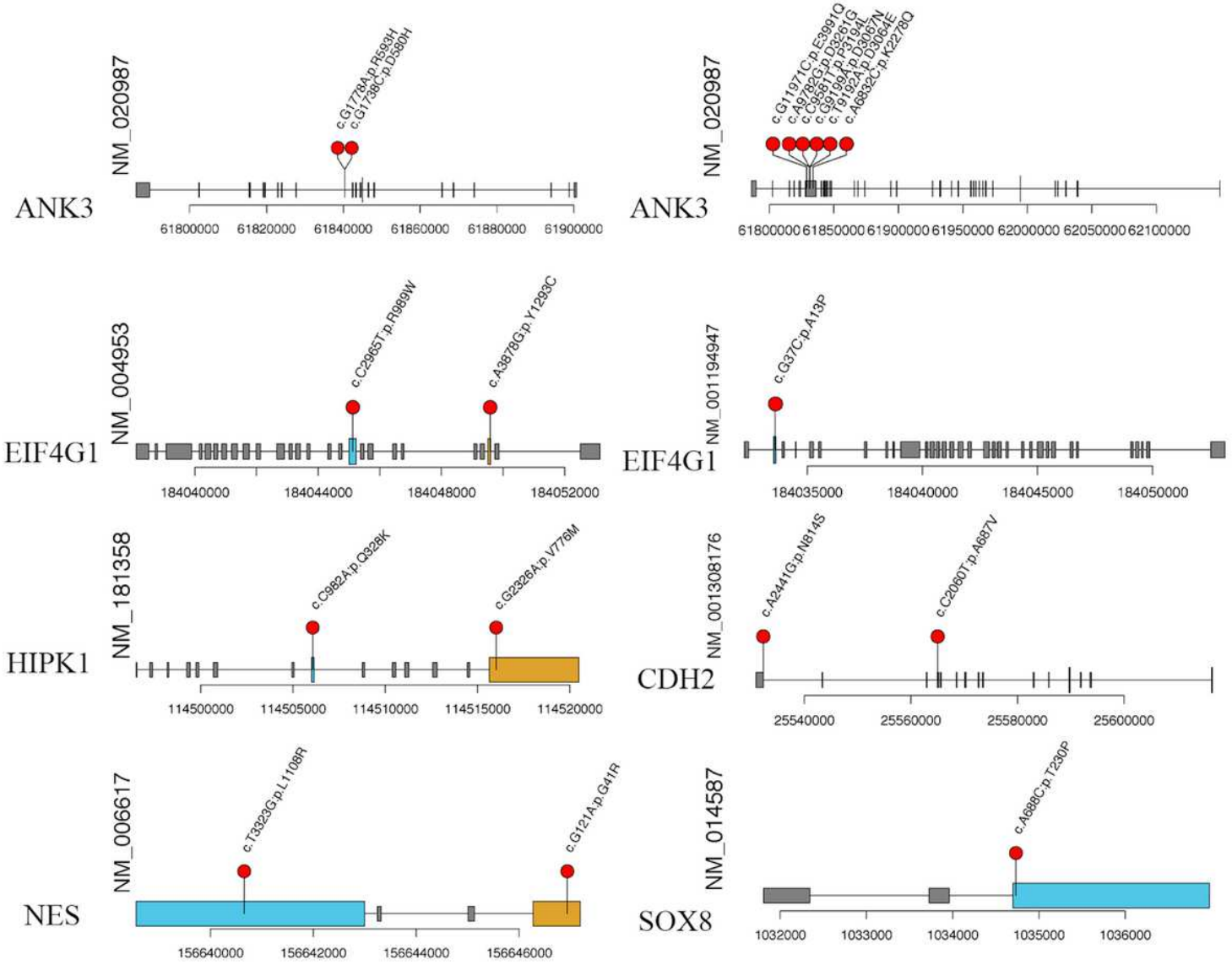


Figure 9

The specific amino acid sites of variants of our candidate gene. The red balls represent the location of rare variant on the encoded proteins or protein domains.

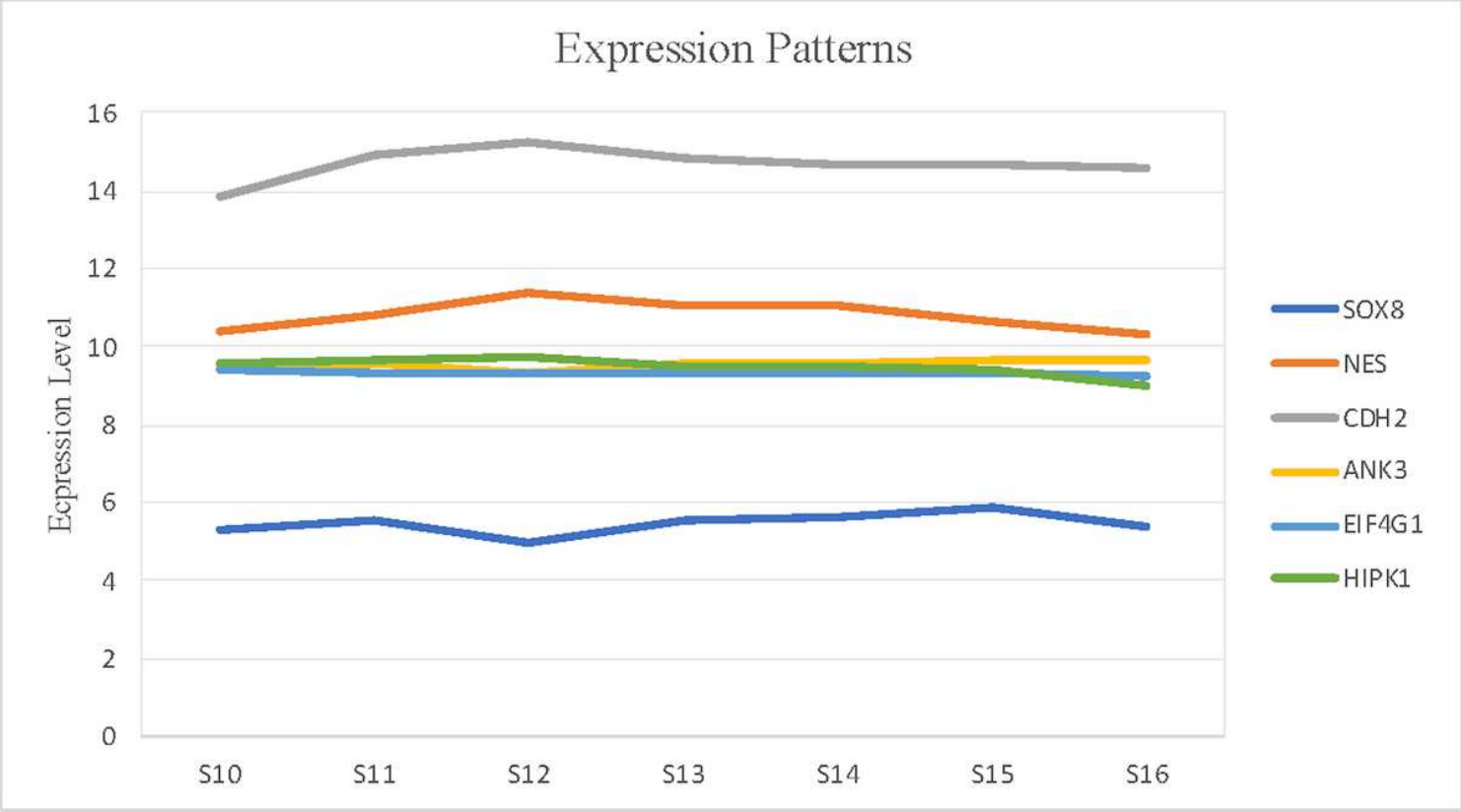


Figure 10

Expression of candidate genes in human embryonic heart. The expression patterns of candidate genes in human embryonic heart at different stages of S10 to S16 were analyzed by microarray. X-axis represents the different stages of human embryonic heart, while the Y-axis indicates the level of gene expression.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementaryfile.docx