Screening of Candidate Key Genes Associated with Congenital Heart Disease Using Bioinformatics Data Analysis

Jingshu Zhang1,*, Xuyao An2, Xiaohan Sun1,3, Kun Yu1 and Tianxing Gong1

1College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China
2Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
3MSc Bioengineering, University of Nottingham, Nottingham, England

*Corresponding author email: zhangjs@bmie.neu.edu.cn

Abstract. Congenital heart disease (CHD) is one of the most dangerous diseases seen in daily life. Aim of this study is to find the deep causes of congenital heart disease. The GSE35776 chip data was extracted from the Gene Expression Synthesis Database (GEO). Analyzing above data was using the R language. The enrichment pathways of differentially expressed genes were processed using the Kyoto Encyclopedia of Genes and Genomics (KEGG) and gene ontology (GO) database. Then this study uses Cytoscape and GCBI to structure protein-protein interaction (PPI) networks, gene regulation networks. 257 differentially expressed genes (DEGs) were found out, mainly focusing on cell cycle, oocyte meiosis, p53 signaling pathway and progesterone mediated oocyte maturation. By constructing gene regulation network, 12 hub genes were screened, including NUF2, BUB1, CENPI, CCNB2, SGO1, SMC4, NCAPD2, TUBB and NCAPH. We hypothesized that these 12 genes may be key factors in CHD.

Keywords: Gene Expression Omnibus; Congenital heart disease; functional enrichment analysis; protein-protein interaction networks.

1. Introduction

CHD is the most serious congenital malformation seen in daily life, accounting for about 28% of all malformations [1], among which nearly one-third of infants need intervention [2]. According to statistics, in the world, 135-150 million babies born every year have congenital heart disease, of which about 45% are moderate or even severe, and the other 55% are common heart malformations. Although more than 90% of children can successfully live to adulthood [3], the mortality rate of congenital heart disease among adults is increasing year by year, and is associated with higher risk of stroke, asymptomatic cerebral infarction and vascular cognitive impairment [4]. Among them, patients with complex conditions may suffer from hypoxia, shock or even death. With the further study of CHD, the study of gene has become more and more important for curing diseases [5]. Therefore, it is very urgent and important to study the genesis and development mechanism of congenital heart disease at the genetic level. Many researchers have studied genes that affect the occurrence of congenital heart disease in many ways, according to statistics, there are about 400 genes related to congenital heart disease [6]. Previous studies have shown that mutations in the SGO1 are associated with a new syndrome characterized by chronic heart disease and intestinal discomfort, known as CAID syndrome. Wang et al. pointed out the relationship between the changes of functional elements in the regulatory region and the occurrence of congenital heart disease [7]. The regulatory elements mainly include promoters, enhancers and so on. Their changes provide new ideas for clinical diagnosis.
and treatment. In addition, studies have shown that GDF1 plays a crucial role in left-right mode formation. We believe that the gene mutation in the coding region is closely related to congenital heart disease [8]. Meanwhile, Samira Kalayinia et al. confirmed that GATA4 gene has a key part in the composition of the heart. The researches show that mutation of NOTCH1 gene has been recognized worldwide as one of the most common causes of CHD, which can be passed down from generation to generation in families [9]. So far, there has been a little research into the key genes responsible for congenital heart disease. In this study, the genes related to cardiac development were analyzed by bioinformatics. The differentially expressed genes were analyzed by GO and KEGG pathways, and key node genes were screened by constructing protein PPI network. A series of bioinformatics methods were used to select genes that may be involved in cardiac development, providing a theoretical basis for prenatal diagnosis and follow-up gene therapy.

2. Methods

2.1. Recognition of DEGs
The gene expression profiles of congenital heart disease(GSE35776) was derived from the GEO database. This is a sample dataset describing the expression of non-coding RNA in congenital heart defects, consisting of 16 non-syndromic tetralogy of Fallot (TOF) infants without 22q11.2 deletion, 3 fetal heart samples, and 8 normal developing infants with right ventricular myocardial gene expression. This study normalized the probe horizontally and mapped the probe based on the annotation platform. The expression profiles of over 20,000 genes were analyzed using R language, and 16 children with tetralogy of Fallot in the data set were selected as the experimental group, and the other 3 fetal specimens and 8 normally developing infants were selected as the control group. Finally, DEG identification was carried out according to the standard screening results of P value <0.005 and logFC absolute value >1.5.

2.2. Enrichment Analysis
In this study, the annotation, visualization and integration discovery database DAVID was used for GO and KEGG pathway analyse to set the gene regulatory network, and then the database was used for Pathway and functional enrichment analysis, and the difference analysis results were obtained. The P <0.05 was selected as the result of significant enrichment.

2.3. PPI Network Construction
Proteins usually work together and rarely work independently, so it is important to study the relationships between proteins and how they work. In order to accomplish this task, String database was selected in this study and PPI network was structured according to difference analysis results. In a PPI network, nodes represent genes, edges represent associations between genes, and established networks can be used to predict associations between specific proteins. In the final study, source nodes and target nodes will be selected and imported into Cytoscape software to construct a gene regulatory network and visualize the results.

3. Result

3.1. Identification of DEGs
In general, this study used bioinformatics method and R language to analyze the differential expression of the sample data. 257 DEGs were obtained from the tissue samples and normal samples of congenital heart disease, including 164 up-regulated genes, 93 down-regulated genes. The volcano map (FIG. 1) was drawn to show the distribution of gene expression.
Figure 1. Volcano plot of DEGs. There are 93 down-regulated genes in the green region, 164 up-regulated genes in the left red region and non-DEGs in the middle gray region.

3.2. Enrichment Analysis

Table 1. Significantly enriched GO terms of DEGs.

| Term                                      | Count | PValue      |
|-------------------------------------------|-------|-------------|
| GO:0000279-M phase                        | 44    | 3.70E-29    |
| GO:0022403-cell cycle phase               | 47    | 4.43E-28    |
| GO:0000278-mitotic cell cycle             | 44    | 4.98E-27    |
| GO:0007067-mitosis                        | 36    | 1.02E-26    |
| GO:0000280-nuclear division               | 36    | 1.02E-26    |
| GO:0000087-M phase of mitotic cell cycle  | 36    | 1.93E-26    |
| GO:0048285-organelle fission              | 36    | 4.20E-26    |
| GO:0022402-cell cycle process             | 51    | 6.01E-26    |
| GO:0007049-cell cycle                     | 58    | 1.46E-25    |
| GO:0051301-cell division                  | 34    | 2.83E-20    |
| GO:0007059-chromosome segregation         | 20    | 3.15E-18    |
| GO:0005819-spindle                        | 22    | 4.07E-16    |
| GO:0000793-condensed chromosome           | 21    | 4.15E-16    |
| GO:0044427-chromosomal part               | 32    | 5.91E-16    |
| GO:0015630-microtubule cytoskeleton       | 37    | 1.16E-15    |

In this study, DAVID database was used in our pathway and GO analysis in order to reveal biological performance of gene differential expression. In this study, 177 GO enrichment results and 10 KEGG pathways were obtained. Table 1 lists the ten most remarkable enrichment pathways, and all consequences of KEGG pathway analysis are listed in Table II. This circumstance showed that DEGs mainly affected cell mitosis, which was related to cycle process, nuclear division, chromosome segregation and concentration, and cytoskeleton formation.
Table 2. Conspicuous enriched KEGG pathways of DEGs.

| KEGG ID  | Description                                   | Count | PValue      |
|----------|-----------------------------------------------|-------|-------------|
| hsa04110 | Cell cycle                                    | 10    | 9.77E-05    |
| hsa04114 | Oocyte meiosis                                 | 9     | 2.54E-04    |
| hsa05322 | Systemic lupus erythematosus                  | 9     | 9.01E-04    |
| hsa04115 | p53 signaling pathway                         | 6     | 0.003296005 |
| hsa05034 | Alcoholism                                    | 9     | 0.005167713 |
| hsa04914 | Progesterone-mediated oocyte maturation       | 6     | 0.009950109 |
| hsa05203 | Viral carcinogenesis                           | 8     | 0.035165111 |
| hsa01130 | Biosynthesis of antibiotics                   | 8     | 0.041013697 |
| hsa04540 | Gap junction                                  | 5     | 0.04376639  |
| hsa00561 | Glycerolipid metabolism                       | 4     | 0.056964896 |

3.3. PPI network construction

As can be seen from Figure 2, the PPI network consists of 27 nodes and 328 edges using STRING database. We used the MCODE plugin to simplify it into two modules. Module 1 has seven nodes, which are NUF2, BUB1, CENPI, CCNB2, CENPF, SGO1 and CENPE (Fig. 3). Module 2 has five nodes, which are SMC4, NCAPG, NCAPD2, TUBB, NCAPH (Fig. 4).

Figure 2. Gene regulatory network.

Figure 3. Subnet a.
Figure 4. Subnet b.

GCBI is a bioinformatics platform which integrates data processing, document retrieval, gene analysis and other functions. In this study, we used gene radar in GCBI platform to establish the regulatory network among gene-related proteins, lncRNAs, miRNAs, activation and indirect relationship. Figure 5 and Figure 6 show the regulatory network related to NCAPD2 and TUBB respectively.

Figure 5. NCAPD2-related regulatory network.

Figure 6. TUBB-related regulatory network.
4. Discussion

Diagnostic or prognostic signs are currently got by performing comparative experiments on high-throughput cases to extract the most significant DEGs[10]. According to the results of our study, there mainly existed four enriched pathways: Cell cycle, Oocyte meiosis, p53 signaling pathway and Progesterone-mediated oocyte maturation [11-13]. Cell cycle, Oocyte meiosis and Progesterone-mediated oocyte maturation all have something to do with the growth and development of cells [14]. P53 signaling pathway regulates cell senescence and death, which is the pathogenic factor of some congenital diseases [15]. Therefore, we believe that these four pathways may be factors affecting congenital heart disease and the growth and development of fetal heart.

Through the establishment of gene regulatory network, we obtained 12 key node genes. We found that nine genes, NUF2, BUB1, CENPI, CCNB2, CENPF, SGO1, CENPE, SMC4, NCAPG, have been accounted to be concerned with congenital heart disease by literature mining [16-18]. Thus, we have carried out a more in-depth study of the remaining three genes NCAPD2, TUBB and NCAPH by using GCBI bioinformatics analysis platform.

The NCAPD2 gene encodes the D2 subunit of the non-SMC coacertin I complex and has a potential part in the upgrowth of the central nerve system. Its new homozygous splicing site mutation causes primary microcephaly, and it is also associated with human SMC family proteins HCAP-C and HCAP-E . Therefore, NCAPD2 may be related to cell mitosis and chromosome structure, additionally affecting the growth and development of cardiac cells. It can be seen from Figure. 5 that there is a protein interaction between SMC4 and NCAPD2. According to the literature mining of NCAPG, SMC4 and NCAPG participate in the regulation of DNA methylation process, which is the influencing factor of some congenital diseases. We predict that NCAPD2 may also be involved in DNA methylation and affect the development of congenital heart disease.

TUBB has been proved to be closely related to centrosomal abnormalities, which can lead to precancerous lesions and tumours . As can be seen from Figure 6, there is an indirect relationship such as activation between ARHGEF2 and TUBB. We found that the single nucleotide polymorphism of ARHGAP9 is related with coronary lesion, which has a vital role in the pathogenesis of acute coronary syndromes such as variable angina [19]. Therefore, the Rho family that is abnormally activated in cardiovascular disease is also closely related to ARHGAP9. The Rho family is associated with microtubules and becomes active under microtubule-depolymerization, and ARHGEF2 participates in gene regulation as a microtubule-related exchange factor. To sum up, ARHGEF2 and TUBB are indirectly related to activation and other factors, and it is speculated that TUBB may also participate in heart development and play a regulatory role.

NCAPH has been proved to be associated with the integrity of chromosome structure. Studies have shown that missing of chromosomal integrity is an vital determinant of mitotic death. The protein interaction between NCAPH and FOXL1 can be seen from Figure.7. Left ventricular dysplasia syndrome (LVDS) is a serious life-threatening congenital heart malformation. The patient data were analyzed by means of DNA sequencing of genes and high-resolution array comparison, and finally mutations of five genes including FOXL1 were found to affect the development of heart valves [20]. The protein interaction between FOXL1 and NCAPH indicates that NCAPH gene may also be related to cardiac development or congenital heart disease formation.

5. Conclusion

A total of 257 DEGs were screened by bioinformatics analysis. Additionally, 12 hub genes were selected out, including NUF2, BUB1, CENPI, CCNB2, SGO1, SMC4, NCAPD2, TUBB and NCAPH, which could be the pivotal genes relevant to CHD. The above analysis provides a abundant view for understanding underlying causes of congenital heart disease. However, bioinformatics analysis usually requires a large number of experimental and clinical researches as evidence. Therefore, the functions of these key genes need to be confirmed by experimental researches.
Acknowledgments

This study was supported by the Fundamental Research Funds for the Central Universities (No. N2019009) and the Fundamental Research Funds for the Central Universities (N180719020). J.Z. and X.A. contributed equally.

References

[1] Rhoia Neidenbach, et al. Improving medical care and prevention in adults with congenital heart disease—reflections on a global problem—part I: development of congenital cardiology, epidemiology, clinical aspects, heart failure, cardiac arrhythmia. Cardiovasc Diagn Ther. 2018, 8(6): 705-715.

[2] Asimina Tsintoni. Nutrition of neonates with congenital heart disease: existing evidence, conflicts and concerns[J]. The Journal of Maternal-Fetal & Neonatal Medicine, 2020, 33(14).

[3] Margarita Brida, Michael A Gatzoulis. Adult Congenital Heart Disease: Past, Present and Future. Acta Paediatr. 2019 Oct; 108(10):1757-1764.

[4] Jonathan M C Smith. Adults With Complex Congenital Heart Disease: Cerebrovascular Considerations for the Neurologist. Front Neurol. 2019 Apr 4; 10:329.

[5] Aline Saliba. Genetic and Genomics in Congenital Heart Disease: A Clinical Review. J Pediatr (Rio J). May-Jun 2020; 96(3): 279-288.

[6] Kylia Williams. Genetics of Congenital Heart Disease. Biomolecules. 2019 Dec 16; 9(12): 879.

[7] Wang F. Progress in the study of gene regulatory region changes associated with congenital heart disease [J]. International Journal of Pediatrics, 2017, 44 (2): 77-79.

[8] Gao X. Association of functional variant in GDF1 promoter with risk of congenital heart disease and its regulation by Nkx2.5. Clin Sci (Lond). 2019 Jun 06; pii: CS20181024. doi: 10.1042/CS20181024.

[9] Samira Kalayinia. A Novel De Novo Dominant Mutation of NOTCH1 Gene in an Iranian Family With Non-Syndromic Congenital Heart Disease. J Clin Lab Anal. 2020 Apr; 34(4): e23147.

[10] Bi, D., 2015. Gene expression patterns combined with network analysis identify hub genes associated with bladder cancer. Comput. Biol. Chem. 56, 71 – 83.

[11] KANEHISA M. New approach for understanding genome variations in KEGG[J]. Nucleic Acids Research, Oxford University Press (OUP), 2018, 47(D1): D590-D595.

[12] KANEHISA M. KEGG: new perspectives on genomes, pathways, diseases and drugs[J]. Nucleic Acids Research, Oxford University Press (OUP), 2016, 45(D1): D353-D361.

[13] KANEHISA M. KEGG: Kyoto Encyclopedia of Genes and Genomes[J]. Nucleic Acids Research, Oxford University Press (OUP), 2000, 28(1): 27-30.

[14] Belloc E. Sequential waves of polyadenylation and deadenylation define a translation circuit that drives meiotic progression. Biochem Soc Trans. 2008, 36: 665-700.

[15] Resnick-Silverman L. Two Faces of SIVA. Cancer Discov. 2015, 5: 581-603.

[16] Bharadwaj R. Identification of two novel components of the human NDC80 kinetochore complex. J. Biol. Chem. 2004, 279: 13076-13085.

[17] Liu XS. Germinal Cell Aplasia in Kif18a Mutant Male Mice Due to Impaired Chromosome Congression and Dysregulated BubR1 and CENP-E. Genes Cancer. 2010, 1: 26-39.

[18] Gong D. The roles of cyclin A2, B1, and B2 in early and late mitotic events. Mol Biol Cell. 2010, 15(9): 3149-3161.

[19] Takefuji M. Mutation of ARHGAP9 in patients with coronary spastic angina. J Hum Genet. 2010, 1: 42-49.

[20] Iascone M. Identification of de novo mutations and rare variants in hypoplastic left heart syndrome. Clin Genet. 2012, 6: 542-554.