Yazdan Rahmati¹, Hasan Mollanoori¹, Soodabeh Shafiee², Sajjad Esmaeili³*
¹Department of Medical Genetics, Iran University of Medical Sciences, Tehran, Iran
²Department of Biology, Sciences & Research Branch, Islamic Azad University, Tehran, Iran
³Medical Biology Research Center, Health Technology Institute, Kermanshah University of Medical Sciences, Kermanshah, Iran

CORRESPONDENCE AUTHOR
Mr. Sajjad Esmaeili
E-mail: sajad.smaeili@kums.ac.ir

CITATION

ABSTRACT
Kawasaki disease (KD) is an inflammatory vasculitis of unknown etiology that occurs in infantile and shows a higher incidence in younger children. This work aimed to investigate the correlation between gene expression status and age of disease onset by integrated bioinformatics analysis. In the first step by using metaDE package, meta-analysis performed for three expression arrays including GSE73464, GSE18606, and GSE68004. In the next step, WGCNA package applied on meta-analysis results for the detection of genes in modules that correlate with the gender and the disease age of onset. Finally, Functional annotations of these genes were carried out to highlight the KD-associated molecular pathways. Our meta-analysis identifies 2417 differentially expressed genes as the most important genes. By applying WGCNA upon these genes we identified a module that has a negative correlation with disease age of onset, so genes in this module show an important role in KD. Functional annotations revealed inflammatory pathway and immune response pathways to infections as the most enriched terms. Genes including, CHUK, TLR5, TICAM2, LY96, MYD88, IRAK4, and JAK2 identified as hub genes. Among them the role of LY96, TLR5, MYD88, and IRAK4 in KD had been confirmed in previous studies but for the first time we introduce TICAM-2, CHUK, and JAK2 as genes that can have important role in KD.

Keywords: Kawasaki disease, metaDE, WGCNA, gene expression, age of onset
1. INTRODUCTION
KD is an inflammatory vasculitis of unknown etiology that occurs in infantile and toddlers. Its main complication is the development of coronary artery lesions (CALs) including coronary arterial dilation, stenosis, and aneurysms [1-3]. Kawasaki disease (KD) exhibits many signs such as high temperature, erythema which occurs in the mucosa of lips or mouth, alterations in the body organs, and large lymph nodes in the neck [4-6]. KD among Japanese and Japanese American children is clearly more common (265/100,000) than other populations like other Asian countries (51 to 194/100,000), Europe (8.39/100,000) or American children (20.8/100,000) [7-12]. To date, the etiology of KD has remained an undetermined question and the causative agents are also unexplained [13]. But, based on researchers’ findings, a genetic predisposition considered as a cause to the development of KD and also the interaction of an unknown infectious source can predispose to this disease [14]. The observation of familial occurrence and raised frequency in Asian populations, taken together suggest the existence of a genetic basis [15, 16]. Associations with genetic variation in several genes such as BLK, CASP3, CD40, FCGR2A, IPTKC, and HLA class II have been ascertained in diverse populations. Also, it is known that the increased risk for the development of coronary aneurysms in populations with European ethnicity is correlated with the genetic variation in the TGF pathway (TGFβ2, TGFβR2, SMAD3) [17].

Heart damage is visible in a significant number of untreated KD patients and they can develop myocardial infarction, unexpected death, or ischemic heart disorder from ectasia (Coronary artery aneurysms) [18, 19]. Early diagnosis is crucial for effective treatment which results in abolition of the inflammatory process to reduce the risk of coronary artery aneurysms (CAA) rates to approximately 5% to 10% [20]. CAA is the most important complication of KD patients and untreated KD children, the disease-associated inflammation modifies the arterial wall which leads to CAA in 25% of them [21]. In developed countries, KD has been reported as the most prevalent cause of acquired heart disease in children [22]. KD shares certain features to other childhood febrile conditions, including, infectious (e.g. staphylococcal and streptococcal toxic shock syndromes, measles and other viral illnesses) and inflammatory conditions which makes differential diagnosis difficult [20]. Although nowadays there are guidelines to assist diagnosis based on clinical signs and symptoms, such as echocardiography, and laboratory variables, rapid diagnosis of KD from other mimicking conditions for treatment and impediment of CAA development remains a momentous mission [23].

Recently, analysis algorithms for the meta-analysis of array data and differential co-expression network, advanced and implemented to study the expression data of genes and microRNAs [24-26]. There is a new functional strategy in systems biology, the weighted gene co-expression network analysis (WGCNA) algorithm, which identifies the most important genes in co-expressed genes that are associated with a sample trait [27, 28]. WGCNA makes modules, sub-network regions, based on similarities in expression profiles of samples, and detects associated genes [29, 30]. By analysis of these modules in two different conditions, healthy control against disease samples, we aimed to identify gender and age of onset -related genes that may represent potential diagnostic biomarkers as well as therapeutic targets with clinical utility.

2. MATERIALS & METHODS
2.1. Expression array datasets and primary processing
Three expression datasets including GSE73464, GSE18606, and GSE68004 in total studied for gene expression of 86 controls 194 patients. These datasets obtained from GEO NCBI at https://www.ncbi.nlm.nih.gov/geo by searching Kawasaki disease and Homo sapiens in the query line. GSE73464 and GSE68004 are produced using Illumina HumanHT-12 V4.0 expression beadchip and included 77 healthy and 154 disease samples. To normalize these datasets, quantile normalization method in limma package was used. GSE18606 produced using Agilent-014850 Whole Human Genome Microarray 4x44K G4112F and included 9 healthy and 40 disease samples. To normalize this dataset quantile
normalization method in limma package was used too. For better results, we included only single measure of each gene by the aggregate function in the S4Vectors package, which gives an average measure of the probes of each gene.

2.2. Quality control, comparability and detection of outlier samples
By metaQC [31] package quality of datasets are assessed and samples that have low quality removed from the rest of analysis, p-value cutoff 0.05 was used for the selection of differentially expressed genes, and p-value cutoff 0.05 was used for the selection of pathways by metaQC package. To evaluate the comparability of control and disease samples, soft Connectivity function in WGCNA package was used. Soft Connectivity evaluates two factors, 1) correlation of expression level of each gene between two datasets, 2) correlation of connectivity of each gene between two datasets. The datasets are comparable if two mentioned correlations are positive and have significance p.value. To remove outlier samples, standardized connectivity (Z. K) method used and samples which had Z. K score<−2 deleted from the rest of the analysis.

2.3. Meta-analysis and detection of significant genes
Fisher method in metaDE package was used to recognize significant genes. FDR cutoff 0.001, used to access the most important genes in this study.

2.4. Network construction and module detection by WGCNA package
The results of metaDE set as input for WGCNA package. The disease dataset of GSE73464 used as the main dataset for the rest of analysis. One weighted gene co-expression networks according to disease samples was made. Using pickSoftThreshold function that helps to choose proper soft-thresholding power, we selected soft thresholding power of 6 for providing scale-free topology fit index that reaches values above 0.9. After calculation of adjacencies, to minimize the effect of noise and spurious associations, adjacency results transformed to Topological Overlap Matrix (TOM), then scaling of Topological Overlap Matrices was used to mitigate the effect of different statistical properties. We used Quantile-quantile plot of the TOMs in the dataset to see what scaling is achieved. Results of TOM set as input to produce dendrogram of genes. CutreeDynamic function used for branch cutting. Minimum module size of 30, and the module detection sensitivity deep Split 2 in block-wise Consensus Modules function were used for network construction.

2.5. Identification of clinically significant modules and functional annotation
Module membership (MM) and gene significance (GS) for evaluating correlation between genes and traits were measured. GS is the correlation between genes and traits and MM is the correlation of the module eigengene and gene expression profile. By measuring MM and GS, we can identify modules with high module membership as well as significant genes related to gender and age of onset. We also evaluated the correlation between GS and MM for genes in modules with high module membership in which central genes considered as the most important elements of the module associated with the trait.

2.6. Detection of hub genes and their functional annotations
Within detected module, hub genes were screened according to criteria cor.geneModuleMembership > 0.8 and cor.geneTraitSignificance > 0.25. Next we carried out functional enrichment analyses of known genes in order to facilitate the interpretation of the biological mechanisms related to these genes. KEGG and STRING databases used to functional and biological interpretation of detected genes, using the KEGG database, important signaling pathways with P.value < 0.05 and combined score > 10 were identified. PPI network of detected genes was constructed by Search Tool for the Retrieval of Interacting Gene (STRING10.5; https://string-db.org/) with a combined score >0.4 as the cut-off point. The coexpression network of the genes was constructed by Cytoscape software[32].
3. RESULTS
3.1. Quality control, comparability and detection of outlier samples
The quality of the datasets was examined using MetaQC package. According to the results of the metaQC package and especially SMR score, we decided to keep all selected datasets in the process of analysis (Table. 1). Dataset comparability evaluated by the soft Connectivity function that provides positive correlation values when there is a high comparability between control and disease datasets for each expression array. Our datasets were comparable as the overall gene expression correlation for GSE73464 was (cor=0.99, p<1e−200), the overall gene expression correlation for GSE68004 was (cor=0.98, p<1e−200) and for GSE18606 was (cor=0.72, p<1e−200) (Fig.1). After selecting GSE73464 as the final dataset for WGCNA package analysis, based on standardized connectivity (Z. K) method, 3 outlier samples removed from this dataset. Figure 3A shows the correlation between clinical traits and the sample dendrogram depicted for the disease dataset of GSE73464.

3.2. Meta-analysis and detection of significant genes
The results of MetaDE package, a heatmap that present up and down regulated genes is summarized in Figure 2. 2417 DEGs using FDR cutoff 0.001 detected from which 930 are upregulated and 1487 are downregulated. These genes set as input for WGCNA package.

---

**Table 1:** quality control of three datasets using metaQC package. . IQC, internal quality control; EQC, external quality control; CQCg, consistency quality control of gene; CQCP, consistency quality control of pathway; AQCg, accuracy quality control of gene; AQCp, accuracy quality control of pathway, SMR, standardized mean rank.

<table>
<thead>
<tr>
<th>datasets</th>
<th>Groups</th>
<th>IQC</th>
<th>EQC</th>
<th>AQCg</th>
<th>AQCp</th>
<th>CQCg</th>
<th>CQCp</th>
<th>SMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE73464</td>
<td>Control-patient</td>
<td>5.615</td>
<td>6.52</td>
<td>248.5</td>
<td>5.622</td>
<td>568.6</td>
<td>410</td>
<td>1.5</td>
</tr>
<tr>
<td>GSE68004</td>
<td>Control-patient</td>
<td>3.3</td>
<td>3.594</td>
<td>204.5</td>
<td>24.77</td>
<td>454.4</td>
<td>410</td>
<td>2.25</td>
</tr>
<tr>
<td>GSE18606</td>
<td>Control-patient</td>
<td>5.615</td>
<td>5.661</td>
<td>68.24</td>
<td>1.161</td>
<td>94.86</td>
<td>410</td>
<td>2.5</td>
</tr>
</tbody>
</table>
3.3. Network construction and module detection by WGCNA package

The KD dataset related to GSE73464 exhibited a scale-free topology as the power was set to 6 and the Scale-free Topology Fit Index reached values above 0.9 for low powers (< 30). This also shows that the batch-effects were not present in our dataset (Fig. 4). For module identification, weighted gene co-expression networks constructed for this dataset. Based on the WGCNA package, a module is a group of strongly co-expressed genes, these genes have similar biochemical and functional properties or belong to similar pathways. By hierarchical clustering, 8 modules identified for the KD dataset, these modules have different sizes in terms of the number of genes which are labelled by the different colors and shown in figure 3B.

Figure 4: Analysis of network topology for various soft-thresholding powers.

3.4. Identification of clinically significant modules and functional annotation

To identify modules that are significantly associated with the measured clinical traits (Age of onset, Gender), we used Module-trait association plot. The analysis identifies several significant module-trait associations. As it can be seen from the plot, the best correlation exists between age of onset and the yellow module (Fig. 5A). We quantified association between genes of the yellow module and age of onset by measuring GS and MM (Fig. 5B). A heatmap was made to analyze the interaction between 8 modules, the heatmap can depict adjacencies or topological overlaps, with light colors denoting low adjacency (overlap) and darker colors higher adjacency (overlap) (Fig. 5C). We used plot Eigengene Networks function in WGCNA package to generates a summary plot of the eigengene network and also add a clinical trait (Age of onset) to the eigengenes to see how the traits fit into the eigengene network (Fig. 5D).
3.5. Identification of hub genes and functional analysis

We find yellow module as the most important one in correlation of KD and age of onset, genes related to the yellow module can be seen in Table 2. According to the results of KEGG database analysis, the most important signaling pathways related to the genes of yellow module are shown in Figure 6A, the cenplot that shows the relation between signaling pathways and yellow module genes are shown in Figure 6B. We set out to identify hub genes with the highest module membership scores according to the criteria cor.geneModuleMembership > 0.8 and cor.geneTraitSignificance > 0.25 in yellow module, then to identify the most significant clusters of the hub genes, PPI network was constituted by STRING as shown in Figure 7. Genes like CHUK, TLR5, TICAM2, LY96, MYD88, IRAK4, and JAK2 are the most interacted genes which are identified as hub genes.

Table 2: the list of genes identified in yellow module related to KD

<table>
<thead>
<tr>
<th>Module</th>
<th>genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yellow</td>
<td>SC02, WAC, ZFP36L1, KDM3B, SPG11, PJA2, SYNJ1, ZNF281, PARP9, NT5C2, UBQLN2, RFWD2, SUMO1P3, WDR51B, RAB33B, FEM1C, KBTBD7, MTMR6, TMEM71, FAM45A, CYB5R4, TLR5, HMGR, WDR26, JAK2, GOLPH3, TP53INP1, STXB5, SLK, IRAK4, TMEM188, NIN, SCYL2, PAPD4, USP15, TLE4, CCDC128, AADACL4, MIER1, FBXO30, CHUK, FCHO2, SMNDC1, USP6, SEL1L, PTEN1, CMTM6, OSBP1L1, PCMT1, RHOT1, LY96, VCP, ZFYVE16, MTHFD2, RP5-1022P6.2, BAZZ2B, DCP2, LYPPLA1, SLC16A6, RAB11A, ZMPSTE24, IPO11, ZEB2, CNIIH4, IFRD1, TICAM2, MYD88</td>
</tr>
</tbody>
</table>
Figure 6: Functional enrichment analysis of yellow module: (A) KEGG pathway analysis of all genes in yellow module. (B) cnetplot of all genes in yellow module that depicts the linkages of genes and most important signaling pathways.
4. DISCUSSION
KD is one of the rare diseases that can become very serious and deadly due to the involvement of human inflammatory pathways and blood vessels [1, 18]. It is important to note that KD is seen only in people of a low age and its prevalence in Asia is higher than in other parts of the world. According to studies in the field, inflammatory pathways are activated in innate immune cells. Studies have shown that inflammatory pathways are activated in innate immune cells and stimulate pro-inflammatory cytokines such as TNF-β.
In this regard, there have been studies that target these signaling pathways and genes that have influenced the course of the disease. However, the disease is still considered pathologically unknown [1, 33, 34]. In this study by selecting three datasets that all of them evaluate the difference of expression between control and disease samples, the correlation between DEGs and Age of onset of KD explored. At first step, it tried to identify the most important genes by metaDE package. We identified 2417 DEGs in two up and down regulated groups. Because of high quality and much number of disease samples, GSE73464 dataset selected for the rest of the analysis, so expression status of DEGs in disease samples of the mentioned dataset used as input for WGCNA package, we also considered age of onset and gender of disease samples as clinical data in our analysis. As seen in figure 3-A, the most valuable correlation between clinical traits and detected modules is a negative association between age of onset and yellow module. It seems genes in yellow module play an important role in disease with early age of onset. We constructed modules and subgrouped regions based on similarity in expression profiles of samples and gene links using WGCNA method. These modules are designed to identify the strong relationships with the onset and development of KD so that they can be used disease treatment. According to Figure 3, the yellow modulus was considered as a module with high association with the disease. All 67 yellow module genes are listed in Table 2. The genes with the highest module membership scores and also with the most number of interactions with other genes related module considered hub genes of that module. Enrichment was then carried out on the basis of signaling pathways (by KEGG pathways; https://www.genome.jp/kegg/pathway.html) and protein-protein interactions (PPI) by STRING software.

---

**Figure 7:** The PPI network of detected hub genes that was analyzed by String software. Genes like CHUK, TLR5, TICAM2, LY96, MYD88, IRAK4, and JAK2 have important role in this network.
Finally, 7 of these genes including CHUK, JAK2, TLR5, TICAM2, LY96, MYD88 and IRAK4 were selected because of the high score. MYD88 (Myeloid differentiation primary response 88) is involved in the signaling pathway within the immune cells as an adapter, it activates transcription factor NF-κB by TLRs and IRAKs [35, 36]. One study found that the activation of MyD88-dependent inflammatory signals in macrophages and vascular cells led to a negative effects on KD [37]. MYD88 also activates proinflammatory cytokines such as TNF-α by activating the TLR4-MYD88 signaling pathway [38]. TLR5 (Toll-like receptor 5) have been found to be involved in the onset of many diseases, such as inflammatory bowel disease [39]. TLR5 is present in the cell membrane and through MYD88 and IRAK4, it activates the innate immune system and secretes cytokines such as TNF-α [40-42]. One study showed that the Toll-like receptor family, including TLR5, was up-regulated in KD [43]. LY96 (lymphocyte antigen 96) as called myeloid differential protein-2 (MD-2) along with TLR5 is involved in the regulation of TNF-α and nuclear factor-κB signaling pathway [44, 45]. JAK2 (Janus kinase-2) is a non-receptor tyrosine kinase involved in cytokine receptors signaling pathways such as interferon receptors, and GM-CSF receptor family. Past studies have shown that JAK2 has been implicated in many diseases, such as leukemia [46] and myelofibrosis, but has not been reported in KD [47]. CHUK (Conserved helix-loop-helix ubiquitous kinase), also called IKKa, is a nuclear factor-κB transcription activator, and it leads to stimulates lipogenesis in the context of hepatitis C virus infection [48]. Also it is involved in signaling pathways of the MAPK, adipocytokine, PI3K-Akt, and mTOR associated with lipid metabolism, but has not been reported in KD [49].

TICAM-2 (TIR-containing Adapter Molecule), also called MYD88-4 or TIRP involved in toll receptor signaling. One study concluded that TICAM-2 is involved in the process of INF-B production through TICAM-1 [50].

Our results shows that LY96, TLR5, MYD88, and IRAK4 genes are important in KD etiolo. On the other hand, TICAM-2, CHUK, and JAK2 genes are the first to be reported in relation to KD. According to the purpose of the study, these genes are predicted to play a critical role in the onset and development of KD and can be potential target in the therapeutic process. Finally, this study should be substantiated by evaluating the expression levels of these genes and proteins in clinical samples.

5. CONCLUSION
In this study for the first time, correlation of several genes with Kawasaki disease in earlier ages detected, these genes are TICAM-2, CHUK, LY96, TLR5, MYD88, IRAK4 and JAK2. This information helps understand the molecular mechanism and treatment of KD.

REFERENCES


[34] Makino, N., et al., Descriptive epidemiology of Kawasaki disease in Japan, 2011-2012: from the


