Review

A review on cardiovascular genetics and its in silico methods

Beutline Malgija and Shanmughavel Piramanayagam*
Department of Bioinformatics, Bharathiar University, Coimbatore 641046, Tamil Nadu, India

Received 12 November 2014; revised 13 July 2016; accepted 18 July 2016

Cardiovascular disease (CVD), encompassing a range of disease from myocardial infarction to congenital heart disease has become an ubiquitous cause of morbidity and mortality worldwide. The major cardiovascular diseases, including coronary artery disease, cardiomyopathy, myocardial infarction, cerebrovascular disease and atherosclerosis are due to the multiple environmental and genetic factors and the probable interactions between them. Despite remarkable growth in the understanding of critical cellular and molecular mechanism during the development and pathogenesis of CVD and the improvements in cardiovascular research, current preventive treatments, diagnostic procedures, and therapies for CVD remains insufficient. Even though several research focuses on the genetics of the disease, many cardiovascular genes still awaits to discovery. Identification of these genes, analysing their complex regulatory and metabolic interactions providing in depth insights into the molecular mechanisms behind these diseases is of great importance. This offers tools for genetic counselling, gene therapy, pharmacogenomics and medicine, which attempt to prevent and treat these diseases. To understand fundamental mechanisms of the disease and to develop and target new therapies, information on genetic factors, gene expression and protein expression patterns are essential. Hence, this review focus on the current perspectives in genome wide expression profiling and in silico approaches in analysing gene expression data.

Keywords: Cardiomyopathy, pharmacogenomics, atherosclerosis, cardiovascular genes, genetic risk factors, myocardial infarction, pathway analysis

Introduction

Cardiovascular diseases (CVD) are an important cause of global morbidity and mortality among the non communicable diseases (NCDs). It is found to be a single larger cause of death globally accounting for about more than a third of all deaths¹. Recent surveys suggest that the disease has risen to younger age groups also. According to recent statistics, the incidence of CVD related deaths in developing countries has grown at an alarming pace. Of the estimated 16.6 million deaths attributed to CVD worldwide, 80% are in developing countries². In India, overall CVDs accounts for about one third of all deaths in 2008. CVDs are expected to be the fastest growing chronic illnesses between 2005 and 2015, rising at 9.2% annually, and accounting for the second largest number of NCD patients after mental illness³. It is expected that, CVD would be the leading cause of death among infectious diseases globally by the year 2020⁴. Hence, controlling and preventing CVDs and their complex pathogenesis, subjective to genetic and environmental factors, have gained great importance⁵.

Great improvements have been achieved in research, during the last decade due to the annotation availability of human genome sequence and characterization of its natural variability, concerning to diseases. Apart from the 25,000 to 30,000 protein coding genes in the human genome, the intergenic non coding sequences may also contain key elements for the regulation of gene expression. Among which single nucleotide polymorphisms (SNPs), which is the differences in individual base pairs, are the most common type of variation. Whereas other sequence variations comprising variable number of tandem repeats such as mini and microsatellites, insertion and deletions of variable lengths may also exist. Since majority of sequence variation are located in the non-coding region, phenotype would not be affected⁶. However, variations in the coding sequences or regulatory regions may affect gene expression and cause changes in phenotype.

Current Limitations in Cardiovascular Research

Clinical biomarkers possess a variety of functions in assessing the standard diagnostic and prognostic mechanism in the management of CVD. Recent plasma and serum markers that fall into diagnostic biomarkers, including troponin I and troponin T have been successfully
incorporated into cardiovascular research especially for myocardial damage\(^7\). Although these biomarkers are effective in diagnosis and management of irreversible myocardial infarction, they found to be unsatisfactory in reversible myocardial ischemia and acute coronary artery syndrome\(^8\). Moreover, currently there are no potential markers for cardiovascular screening, since the prevention of events in individuals with higher risk for CVD (primary prevention) tends to have significant impact on the overall health burden. Also a number of markers developed during the last decade were not found to be successful\(^9\), which thus requires integrative research. A recent study\(^10\) provides an evidence for the association between cardiovascular traits and polymorphic markers located in the genes encoding for inflammatory mediators (interleukins, chemokines, selectins and adhesion molecules). Moreover it requires further investigation in finding the functional role of these polymorphisms in the development of CVD.

Genetic studies of atherosclerosis and ischemic stroke (IS) have not reached reliable results till now due to limited sample sizes, inappropriate hypothesis of candidate genes, differences in the methods used, reproducibility of phenotypes across the studies and the misinterpretation of various interactions among genes and to their environment\(^11\). The variation in findings is due to the influence of multiple genes. But the advent of new methodologies for the evaluation of gene expression profiling, proteomic and metabolomic studies opened a new era in human genetic diseases, including IS\(^12\). The advancement in high-throughput technologies, namely gene expression profiling and genome wide association studies opens path for the development of necessary tools to pinpoint the genes responsible for disease susceptibility and progression and thus their complex networks and pathways triggered by these genes.

**Genome Wide Association Studies (GWAS) to CVD**

GWAS helps in studying the genetic variation across the human genome to identify their association with diseases. The suscep of GWAS led to the detection of susceptibility alleles for many of the common “complex” diseases including CVD. In the past decade, many different loci associated with diseases have been identified and provided insights into the architecture of complex traits. Most of the cardiovascular diseases are polygenic involving both genetic and environmental factors\(^13\). Hundreds of loci associated with CVD were identified after the completion of draft human genome sequence. Thirty common loci have been identified using this technique for their independent association with coronary artery disease (CAD) and myocardial infarction\(^14\). Identification of such locus associated with CVD risks have provided possible insights behind the molecular mechanism and the associated pathways, whereas the results are premature which further awaits extensive research.

Several important loci for CVD have reported from various genome wide linkage analysis. Rampazzo et al\(^15\) identified a locus on chromosome 1q42-43 in association with arrhythmogenic right ventricular cardiomyopathy, which was later mapped to the RYR2 gene encoding the ryanodine receptor in cardiac muscle, a component of calcium channel that supplies ions to the cardiac muscle having role in cardiomyocyte contraction\(^16\). Another locus associated with dilated cardiomyopathy, diffuse myocardial fibrosis and sudden death was found on chromosome 10q25 - q26.1\(^17\). Further mutation in TNNI3, having role in contraction and relaxation of cardiac muscle was found for its association with familial restrictive cardiomyopathy\(^18\). Another study reported mutation in TTN gene to have association with cardiomyopathy. Several other genes MYH7, MYBPC3, TNNI3, TNNNT2, MYL2, MYL3, TPM1, ACTC, and TNNT1 have also been reported for their role in hypertrophic cardiomyopathy\(^19\).

Many GWAS of atherosclerosis reported a most important locus on chromosome 9p21 that contains a gene for a long non coding RNA\(^20\). Studies also reported the chromosomal region 9p21.3 in myocardial infarction, lacking the involvement of protein coding genes and the underlying mechanism remains unclear\(^21\). MEF2A gene on chromosome 15q26 was found to have their role in CAD, by their active involvement in atherosclerosis\(^22\). Mani et al\(^23\) identified a missense mutation in LPR6 on chromosome location 12p13, in autosomal dominant early CAD. Locus 9p21 has found in stroke GWAS, further the genes CDKN2B, CDKN2A and MTAP were reported as the prime candidate genes at the locus, whereas the locus was initially recorded in CAD and myocardial infarction cohorts of European decent\(^20\). A single SNP (rs2200733-T) on chromosome 4q25 was extensively associated with cardioembolic stroke. Also SNPs in paired-like homeodomain transcription factor 2 (PITX2) and zinc finger homeobox 3 (ZFHX3) were found to have their association with both ischemic stroke and atrial fibrillation and further pathway
analysis reveals their role in cardiac development\textsuperscript{24}. The NADPH oxidase 4 (NOX4) gene was also found to have a role in cardiovascular research, but the ideal target for these is not yet defined\textsuperscript{25}.

Even though these methods recorded several genes and their locus, it additionally prompts the confounded issue of multiple testing. As more GWAS are accounted for, many of the loci are replicated across studies and still there should be a place for previously identified biochemical or genetic results. Hence analysing the pattern of gene expression, which renders knowledge about the critical genes, thus the identification of crucial pathways is beneficial. The availability of public gene expression and genome repositories like gene expression omnibus (GEO)\textsuperscript{26}, Stanford microarray database (SMD)\textsuperscript{27}, The human genome database (GDB)\textsuperscript{28}, The National Centre for Biotechnology Information (NCBI) etc. have opened up a realm of challenging possibilities provided by efficient integration of gene expression data, and analysing these data can be plausible by the development of bioinformatics tools and approaches.

**Gene Expression Profiling**

Gene expression refers to the process of transcribing a gene's DNA sequence into RNA, which provides a measurement of the activity of gene under certain biochemical conditions\textsuperscript{29}. Since cardiovascular diseases encompass a group of diseases and is multigenic, it is essential to find a set of common genes that are differentially expressed. The opportunity to identify the set of causative genes for these multigenic diseases was limited in the pre-genomic era. Traditional genetic approaches were designed to find the genes responsible for Mendelian cardiovascular disorders, which results from a single base change in gene, which leads to alteration of protein function. Gene expression profiling is an approach in finding the function of a gene and to profile the expression pattern of these in parallel. This approach is based on the assumption that genes sharing similar patterns of expression under different conditions are functionally related and that changes in gene expression reflect the demands placed on cells by changing physiological conditions. There are several methods of gene expression profiling which includes serial analysis of gene expression (SAGE)\textsuperscript{30}, microarray\textsuperscript{31}, massive parallel signature sequencing \textsuperscript{32}. In case of microarrays, the expression data can be presented as a matrix of fluorescent intensities, corresponding to each spot on the array, whereas rows represent the genes and columns the cDNA samples. Computational methods and tools provide knowledgeable information from these values.

**In Silico Approaches**

The *in silico* analysis of these data with the integration of transcriptomics, proteomics and metabolomics studies might provide better insights about the knowledge of CVD and the workflow in analysing the potential pathway through gene expression data is shown in Figure 1. Finding the significant genes from gene expression data and the analysis of their regulatory regions, further integration of knowledge from SNP microarray, miRNA microarray and GWAS will provide significant insights into the molecular mechanisms of cardiovascular disease. The construction of gene networks is essential to obtain a clear knowledge about the interactions of among various genes and towards the environment.

**Identification of Differentially Expressed Genes**

For the identification of genes that are differentially expressed, the genes are grouped into positive and negative value groups and ranked according to their absolute values. The top genes from the two groups are selected as the informative genes based on statistical cut off. For example, Li *et al*\textsuperscript{33} proposed a genetic algorithm/k-nearest neighbour (GA/KNN) method for the identification of marker genes with an accuracy of 88.5%.

![Flowchart depicting the analysis of pathways by the knowledge from various sources.](image_url)

Fig. 1 — Flow chart depicting the analysis of pathways by the knowledge from various sources. The flowchart depicts the analysis of metabolic pathways from various microarray gene expression data namely the cDNA microarray, SNP, miRNA microarray and genome wide association studies by the identification of genes that are differentially expressed.
genes. Linear models such as t-test, F-test, analysis of variance (ANOVA) and linear regressions have been proposed. Even though these models are flexible enough for analysing microarray data, they perform results ineffectually due to the low number of replicates. One way to get better performance is to deal with gene specific variance estimates. To the variance of t-statistic a small number called fudge factor should be added and the optimal value of the fudge factor is derived using cross validation. Since the analytical distribution of t-statistics is not known, the null distribution was determined by the permutation of the arrays between both the conditions and this approach has been implemented in the popular significance analysis of microarrays (SAM) package. Prediction analysis for microarrays (PAM) is another statistical package implemented using R for class prediction and survival analysis based on the method of nearest shrunken centroids which identifies subsets of genes that characterize each class. Methods based on Bayesian model assumptions can also be used to control the variances. Perez-Sanchez et al. identified the differentially expressed genes (DEGs) between monocytes of antiphospholipid syndrome (APS) and systemic lupus erythematosus (SLE) patients in the area of inflammation, atherosclerosis and cardiovascular disease, which might pave way for the development of targeted therapies.

Clustering

Clustering refers to the grouping of genes in such a way that genes with the same functionality are grouped in the same cluster. A number of computer algorithms and software's have been developed for clustering gene expression data. The most prevalent approaches include: hierarchical clustering, K-means clustering, clustering through self-organizing maps (SOMs). Hierarchical clustering creates a dendrogram, where clusters at one level of the hierarchy are joined into a cluster at the next level. The partitional clustering methods such as K-means or SOMs assigns each gene to a single cluster, which fails to provide information about the influence of a particular gene for the overall shape of clusters. Hence, Doulaye and Phillippe applied a fuzzy partitioning method; Fuzzy C-means (FCM) which offers an efficient way to define subset of genes which are more tightly associated to a given cluster.

Gene Regulatory Networks

The biological cells are found to possess mRNAs and proteins that arise from gene expression, which interacts with each other or bound to cell membranes, associating further with molecules in the environment or pass through cell membranes and mediate long range signals to other cells comprising a gene regulatory network. Hence, estimating a gene network is an important task. Since microarray experiments offer tremendous amount of gene expression data, constructing gene networks for differentially expressed genes helps in elucidating the relationships of various genes. The groups of genes, their regulatory proteins and their interactions are often referred to as regulatory networks, whereas the complete set of metabolites and the enzyme driven reactions are termed as the metabolic networks.

Bonnet et al. reconstructed a regulatory network using data's from gene expression, microRNA (miRNA) expression and clinical parameter of prostate cancer patients. This led to the identification of several genes associated with cell cycle as well as novel functional categories and also the novel miRNAs that might be linked to prostate cancer.

The most important steps in reconstructing networks from gene expression data are selecting the network model and fitting the available data into the network's structural parameters. Several models have been proposed for both continuous and discrete variables include Boolean network models, probabilistic Boolean networks, Bayesian models, differential equation model, neural network models and hybrid models. The reconstruction of gene networks is found to be crucial due to the high-dimensionality of gene expression data. i.e., a small number of time points with respect to large number of data. Hence Alakwaa et al. developed a bi-clustering functional enrichment analysis tool kit to study the effect of bi-clustering in reducing the dimension of data.

Pathway Analysis

Pathway analysis or functional enrichment is becoming a popular method for the analysis of microarray data. This method integrates the normalized data and their annotations, such as metabolic pathways and functional classification using gene ontology. The detection of low levels of expression and smaller changes in expression is becoming possible in pathway analysis. This method also allows the use of different biological data, such as gene expression data and metabolomic and proteomic data, which will help integrate data from these sources into metabolic networks.
Challenges

Although there have been a number of substantial findings to date in cardiovascular genomic research, still there are some significant challenges in genomics of cardiovascular diseases. Since this involves a group of disorders or condition, it is beneficial in finding the common genetic causes responsible for the disease. Also the analysis of gene variants or polymorphisms that do not directly cause disease but may influence the regulatory factors such as promoter regions, thus altering gene expression or influence the function of key enzymes responsible for normal functioning is important. With the advent of high-throughput technologies, necessary tools are available to pinpoint the genes responsible for disease susceptibility and progression and thus their complex networks and pathways triggered by these genes. Finding the genetic determinants for CVD is essential for individualized medicine, since each individual affected by the same type of CVD having the same symptoms, may be caused by the altered combinations of genetic and environmental factors, hence requires different therapies. This may also provide new biomarkers for diagnosis, and therapeutic strategies according to genetic risk factors to acquire early preventive measures. Even though many CVD linked loci has been mapped successfully by linkage analysis, still there is a need for finding the responsible gene and sequence variation at the loci, which is little complicate since the region may sometimes include tens and hundreds of genes.

Reconstruction of gene regulatory networks from gene expression, miRNA expression and clinical parameters, offers several genes and novel functional categories and miRNAs\(^{28}\). The expression of miRNAs was found to be altered during the development of cardiovascular disease\(^{31}\), emphasising the importance of miRNAs in cardiovascular research. Since this is a complex disease, involving the combination of various genetic, environmental and developmental factors, the integration of multiple areas of research is needed. This requires the combined analysis of gene expression and miRNA expression which provide novel therapeutic candidates. Even though the current methods improves the understanding of the disease, more work is needed to identify the genetic factors and for the better understanding of the phenotypic and genotypic variations. Hence, the integration of next generation sequencing data with transcriptomics, proteomics and metabolomic studies will provide a new paradigm in cardiovascular research by the identification of common genetic factors and the assessment of molecular events behind cardiovascular risk and its progression which further afford tremendous knowledge about the molecular mechanisms of CVD.

References