EXPLORATORY ANALYSIS OF GENE MICROARRAY DATASETS FOR MOLECULAR DIAGNOSIS

Alexei Leviţchi, Daniela Abduşa, Maria Duca
University of the Academy of Science of Moldova

Summary

The aim of the investigation was exploratory analysis of the free available expression microarray datasets for the identification of candidate genes hypothetically involved in manifestation of cardiovascular diseases. Applying developed methodology, there was determined a set of 6088 of candidate genes. These findings were validated through two sets of genes: cardiovascular priority genes and associated genes. It showed that 10 per cent of the genes belong to one of the list, while the other represents a set of special interest to be demonstrated their involvement in cardiovascular diseases. To be noted more then 2000 genes have no annotation and their functions still need to be clarified. This set of genes potentially can be applied for the development of a cardio-chip for the molecular diagnosis of various CV pathologies.

Rezumat

Analiza explorativă a seturilor microarray de gene pentru diagnosticul molecular

Scopul lucrării a fost evidenţierea genelor implicate în bolile cardiovasculare prin analiza explorativă a seturilor de date microarray. Pentru extragerea şi prelucrarea datelor a fost utilizată o metodologie de lucru, elaborată în cadrul Laboratorului de Bioinformatică, *CUBM, UnAŞM*. În baza nivelului de expresie a genelor analizate, a fost determinat un set de 6088 gene candidate, ipotetic implicate în bolile cardiovasculare. Aceste rezultate au fost validate în baza a două liste de gene: gene cardiovasculare prioritare şi gene asociate bolilor cardiovasculare. În urma validării, s-a stabilit că 10% din gene aparţin uneia din liste, în timp ce celelalte gene reprezintă interes deosebit pentru cercetare, fiind lipsite de dovezi privind implicarea lor în bolile cardiovasculare. Genele identificate în urma analizei, vor putea fi utilizate pentru elaborarea unui *cardio-chip* pentru diagnosticul molecular al patologiilor cardiovasculare.

Introduction

According to statistics, cardiovascular diseases (CVD) remain the leading cause of death all over the world. In 2008, more than 17,3 million people died from CVD **[3, 6]**. The main forms of CVD are coronary heart disease and stroke, which caused 7,3 million deaths **[3]** and, correspondingly, 7,2 million deaths **[11].**

Every year, CVD causes over 4,3 million (48%) deaths in Europe **[1].** Moldova is Europe's leading country in terms of mortality due to vascular disease, registering a high rate of deaths, as cardiovascular and cerebrovascular disease **[7]**. National Bureau of Statistics reports that in 2010 56,2% of all deaths were cardiovascular diseases as a cause. Thus, one of the priority research directions of the scientific community is cardiovascular diseases (http://www.statistica.md/newsview.php?l=ro&idc=168&id=3384).

The cause of the cardiovascular diseases is the result of interactions between genetic and environmental factors, which together contribute to individual susceptibility. Genetic studies of cardiovascular diseases allowed identification of hundreds of associated loci, elucidating the specific mechanisms of CVDs and paved the way for the development of new diagnoses and treatments, major goal leading to the development of personalized medicine. Genomic data and bioinformatics tools offer new approaches in studying cardiovascular pathologies caused by a single gene or a multigene complex. Genomic databases provide both access to useful biological information, as well as analytical tools and methods to apply these data to solve specific biological problems.

At present, a growing number of bioinformatic tools allow the analysis of genes and proteins large sets, analysis of biomarkers at the molecular pathways and processes, development

of the cardiovascular network. Computational analysis of large data sets and their interpretation, allow their representation in new and unique forms that give researchers a better understanding of the basic problems facing humanity: cancer, cardiovascular disease, tuberculosis and liver cirrhosis.

Thus, the aim of the investigation was exploratory analysis of the free available expression datasets for the identification of candidate genes hypothetically involved in manifestation of cardiovascular diseases.

Advantage of free available microarray profiles investigation is systematization of the information about cardiovascular diseases manifestation based on gene expression and their integration into the study of gene involvement in manifestation of cardiovascular pathologies. Genes identified by the analysis can be used in the development of a *cardio-chip* for molecular diagnosis of cardiovascular pathologies.

Materials and methods

The investigation was realized according to the pipeline, elaborated for the microarray datasets extraction, analysis and interpretation, in frame of Laboratory of Bioinformatics, University Centre of Molecular Biology, UnASM **[5] (*figure 1*)**.

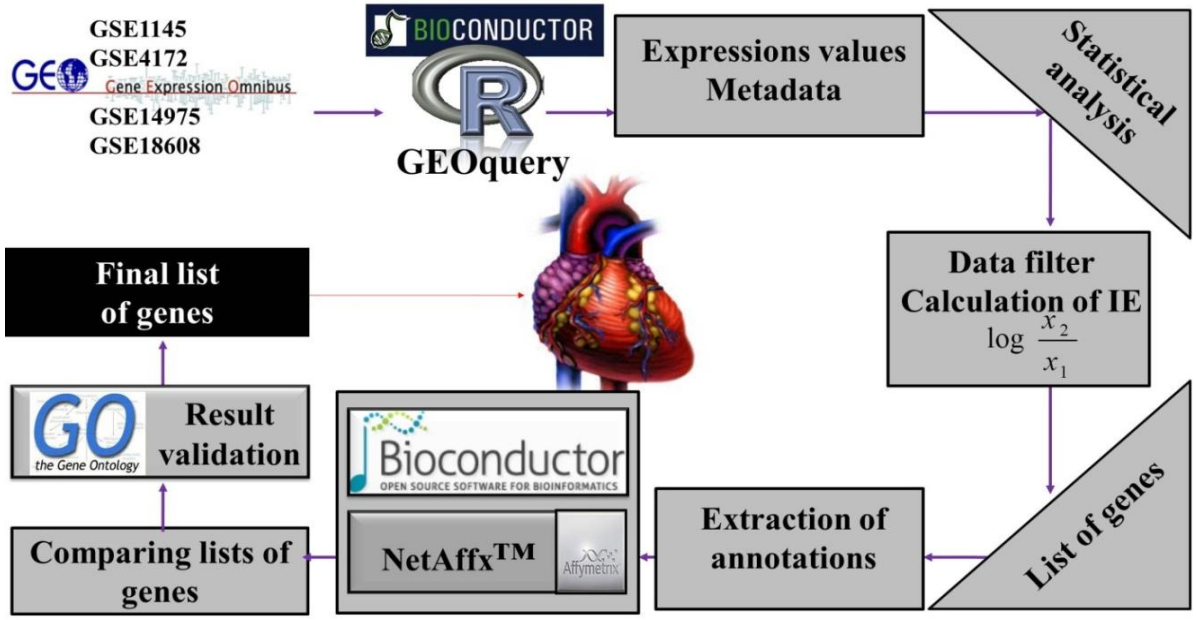

Figure 1. *Strategy of genes involved in cardiovascular pathologies determination.*

Microarray datasets are freely available for downloading from Gene Expression Omnibus, GEO (*http://www.ncbi.nlm.nih.gov/geo/*) [10]. Only datasets refereed to CVD were used in analysis.

Dataset extraction, statistical and exploratory analysis were done under *R environment* [2, 8]. There were used R packages stored on Bioconductor, which advantage is open source and free availability (www.bioconductor.org) [4].

Microarray data extraction

For the dataset extraction there was used *GEOquery* R package **[9]**, while annotation was downloaded by ***NetAffx Analysis Center*** (http://www.affymetrix.com/analysis/index.affx) from Affymetrix website.

Basically there were used data from two kinds of microarray chips: *Affymetrix 95A version 2* and *Affymetrix U133 plus 2.0.* Respectively, two annotation databases were downloaded from BioConductor - *hgu95av2.db* and *hgu133plus2.db.*

The data are stored as *GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array* and GPL8300 *Affymetrix Human Genome U95 Version 2 Array*, and comprise about 54000 and, respectively, 12000 genes. These GPLs include most of all expression datasets from GEO: GPL570 includes 60356 GSMs and 2173 GSEs, while GPL 8300 – 4950 GSMs, 277 GSEs). For the present study GSE4172, GSE14975, GSE18608, GSE1145 were selected as referred to the experiments on the CVD investigation (*table 1*).

Table 1

Selected dataset description

GSE	Title	GPL	Description
***GSE*1145**	Gene expression profiling of human inflammatory Cardiomyopathy	*GPL*8300 *GPL570*	The study results provide data on the gene expression profile in aortic stenosis and different subtypes of cardiomyopathy
***GSE*14975**	Rac1-Induced Connective Tissue Growth Factor regulates Connexin 43 and N-Cadherin Expression in Atrial Fibrillation	*GPL570*	Signal transduction was studied in atrial remodeling, which contributes to the pathogenesis of atrial fibrillation
***GSE*18608**	Transcriptional Profiling of CD133+ Cells in Coronary Artery Disease and Effects of Exercise on Gene Expression	*GPL570*	Differential expression genes were identified in CD133 + cells in patients with coronary disease
***GSE*4172**	Changes in cardiac transcription profiles brought about by heart failure	*GPL570*	The GWAS studies, expression profiles were obtained for different subtypes of dilated cardiomyopathy

These contained datasets regarding gene expression in case of several forms of cardiomyopathies. GSE1145 and GSE4172 data, though belonging to different platforms, were joined as contains data on similar experiments.

Statistical analysis of microarray datasets

All datasets included samples in several repetitions. In order to receive a unique value of gene expression for a certain stare, average values were calculated and further used in proceeding:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where: $\bar{x}$ - average value of gene expression; $\sum_{i=1}^{n} x_i$ - sum of expression values of a gene; n – number of different stares, gene expression was estimated; i = 1, 2,…, n.

Exploratory analysis

In order to solve the problem of inter-platform reproducibility, all the values of expression were transformed in *Index of Expression* (IE), as the measure of magnitude change of gene expression in a pathological stare comparative to normal stare:

$$IE = \log \frac{x_2}{x_1}$$

In the end, each of the genes was characterized by IE values for each of the analysed pathologies.

Data filtering

In order to identify genes with differential expression, it was empirically stated IE≥1, which means that a gene modified it's expression at least twofold. The sigh of the IE value indicate increase of expression in case of "+", and decrease – in case of "-". For general purpose, IE absolute value was used for the gene ranking, and the sign – for the explanation of the produced effects in different stares.

Results

On the all, there were identified 3400 GPLs for *Homo sapiens* in GEO. Affimetrix, Agilent Technologies and Illumina technologies are basic providers of expression datasets stored in GEO. Most of them belonged to Affymetrix microarray chips (*figure 3*).

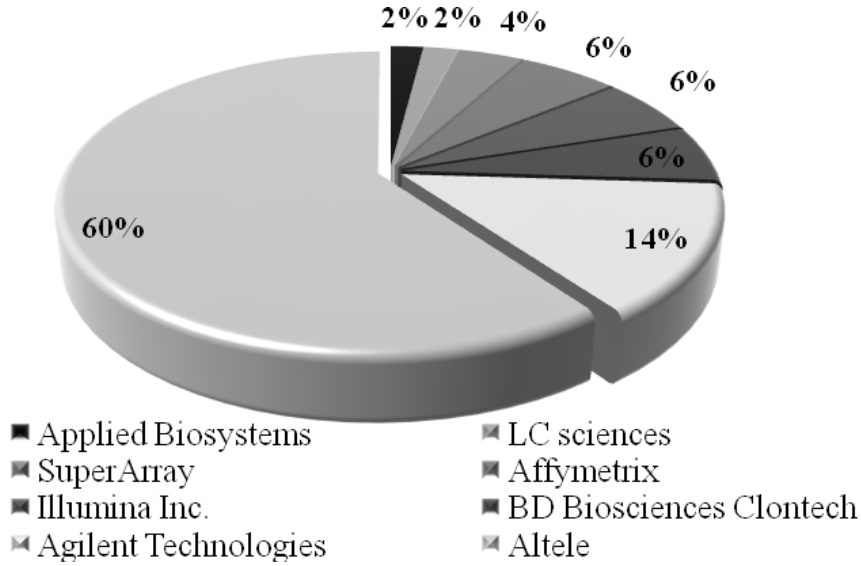

Figure 3. *Microarray chip datasets from different providers.*

Still, over 60% of platforms include unsystematic data, primary datasets which were not published in articles, or chips produced by individual scientific centres. Although, potentially, these data can provide some information, but it should be a complex methodology necessary for their analysis and interpretation.

Selected *GPL*570 was used to produce 2173 GEO series, only 4 of them were identified with the reference to CVD: GSE4172, GSE1145, GSE14975, GSE18608. Another platform, *GPL8300*, was applied for 277 GSEs, only one (GSE1145) being selected for the analysis.

From each of the datasets there were extracted a different number of GEO samples, according to the aim of the investigation. The final dataset comprised whole set from GSE1145, GSE4172, GSE14975, totally 129 GSMs. Eleven GSMs were extracted from GSE18608.

After calculation of IE values, there was estimated its variation in correlation to number of filtered genes. The correlation showed to be high and negative (r=-0,97), that means the increase of the IE value, as the threshold for the filtering, will decrease the number of candidate genes.

This step is important for the avoidance of null results after filtering, thus, being ensured that each of the list of candidate genes will contain at least one result. There were over 6000

genes which passed the threshold as genes with differential expression in those four CV pathologies (*table 2*).

Over 75 % of the results referred to the *Atrial fibrillation* pathology, that can be explained by the complexity of the processed involved in the support of normal functioning of atriums.

The list of candidate genes, noted as Probe_IDs identifiers from the chips, were supplemented with annotation data. In case of the lack of annotation data, Probe_ID was still used. For each of the probes there were extracted Entrez IDs, gene name and symbol, the first one being unique identifier from Entrez system of NCBI databases for all sequences.

Table 2

Filtered and validated results

Cardiovascular pathology	Probes	Cardiovascular-associatedgenes	Cardiovascular prioritized for annotation gene
Coronary artery disease	975	157	12
Atrial fibrillation	4542	36	2
Aortic stenosis	482	170	14
Cardiomyopathy	89	89	62
Total	**6088**	**452**	**90**

Following, it was possible to compare candidate gene lists between the platforms and those met in *cardiovascular associated* and *cardiovascular prioritized for annotation* gene lists, elaborated by *Cardiovascular Gene Ontology Annotation Initiative*, in collaboration with *University College London* and *European Bioinformatics Institute* (EBI), *British Heart Foundation* (http://www.ucl.ac.uk/cardiovasculargeneontology/). Such validation allows verification of the applied methodology for the true positive results.

Total number of genes considered as true positively determined was 542, from which most (over 80%) were genes with proved association to CVD. For 90 genes the association was not yet shown. Over 30% of genes had no annotation, meaning the lack of determined function in biological processes.

Coronary artery disease involve modification of expression of 975 genes, from which 157 were shown as CV associated genes, and 12 still waiting for the proof from association studies. Though *Atrial fibrillation* pathology was characterized by more then 4,5ths of genes, less then 10% of them were associated with CVD. *Aortic stenosis* pathology possesses the highest content of CVD associated genes to total number of candidate genes, almost 40%.

Conclusions

All these set of candidate genes can characterize differences in gene expression modification specific to the studied pathologies, and could serve for molecular diagnostic. Each of the pathologies possess from 2 (aortic stenosis) to 62 (cardiomyopathy) genes which were shown previously to be associated to CVD, while the rest cca 5000 could be useful for the precision of the diagnostic. These findings are in concordance with the multigenic complexity of CVDs and each of genes has its contribution to the variability of pathological processes, which still needs to be determined.

These genes can be applied for the proposal of the cardio-chip for the molecular diagnosis of CV pathological stares. Besides, candidate genes' expression can be used by doctors as additional indices of associated pathologies, with still unknown engagement in other diseases.

References

1. Apetrei E., Societatea Română de Cardiologie, Publicaţie a Societăţii Române de Cardiologie, numărul 2/2009.

2. Baayen RH. languageR: Data sets and functions with “Analyzing linguistic data: a practical introduction to statistics". 2008. R package version 0.953, http://cran.r-project.org/web/packages/languageR/languageR.pdf.
3. Causes of death 2008, World Health Organization, Geneva, 2008, http://www.who.int/ healthinfo/ global_burden_disease/ cod_2008_sources_methods.pdf.
4. Gentleman RC, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5.
5. Martea R., Abduşa D., Dragomir L., Levițchi A. Analiza datelor microarray pentru evidenţierea aspectelor moleculare legate de caracterele valoroase în ameliorarea plantelor. Culegere de teze, Chişinău, 2011, p.43.
6. Mendis S, Puska P, Norrving B editors. Global Atlas on Cardiovascular Disease Prevention and Control. World Health Organization. Geneva. 2011.
7. Protocol clinic naţional „Accident vascular cerebral ischemic”. Chişinău. 2008. p. 8.
8. R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. 2005. Vienna, Austria. ISBN 3-900051-07-0.
9. Sean Davis and Paul S. Meltzer. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics (2007) 23 (14): 1846-1847.
10. Tanya Barrett, Dennis B. Troup, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Rolf N. Muertter, Michelle Holko, OluwabukunmiAyanbule, AndreyYefanov and Alexandra Soboleva. NCBI GEO: archive for functional genomics data sets—10 years on. Nucl. Acids Res. (2011) 39 (suppl 1): D1005-D1010).
11. World Health Organization, Fact Sheet 317: Cardiovascular Diseases. Updated September 2009.

Web references

http://www.affymetrix.com/analysis/index.affx
http://www.ncbi.nlm.nih.gov/geo/
http://www.statistica.md/newsview.php?l=ro&idc=168&id=3384
http://www.ucl.ac.uk/cardiovasculargeneontology/
www.bioconductor.org

SEMNIFICAŢIA MODERNĂ A IMUNOREABILITĂRII ŞI PRINCIPIILOR DE IMUNOTERAPIE
(Revista literaturii)

Elena Berezovscaia[1], Doina Barba[2], Lucia Andrieş[1]
Laboratorul Alergologie și Imunologie Clinică (1), Clinica Medicală nr. 6 (2),
Universitatea de Stat de Medicină şi Farmacie «Nicolae Testemiţanu»

Summary

The modern significance of the Immunorehabilitation and principles of Immunotherapy

The scientific analysis of the fundamental immunological sources of the modern literature showed difference and sometimes controversial views according to the definition of the terms, such as immunorehabilitation, immunocorrection, immunomodulation. Deciphering and revealing the procedures, methods, immunotrope remedies used in the treatment of many diseases have been the subject of the essence of the work presented.

Rezumat

Analiza ştiinţifică a surselor fundamentale imunologice denotă viziuni diverse, iar uneori şi controversate cu referire la imunoreabilitare, imunocorecţie, imunomodulare. Descifrarea