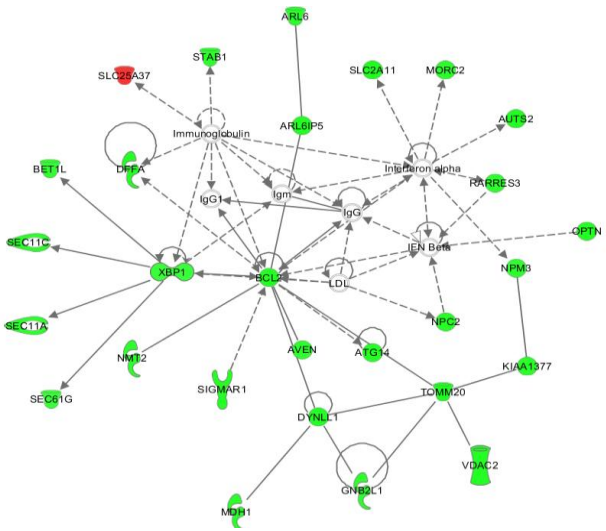
UNIVERSITA' DI NAPOLI FEDERICO II

DOTTORATO DI RICERCA
BIOCHIMICA E BIOLOGIA CELLULARE E MOLECOLARE
XXIV CICLO

Michele Olivieri

Putative transcriptional regulatory elements and gene networks associated to Familial Combined Hyperlipidemia (FCHL)



Academic Year 2010/2011

UNIVERSITA' DI NAPOLI FEDERICO II

DOTTORATO DI RICERCA
BIOCHIMICA E BIOLOGIA CELLULARE E MOLECOLARE
XXIV CICLO

Putative transcriptional regulatory elements and gene networks associated to Familial Combined Hyperlipidemia (FCHL)

Michele Olivieri

Tutor
Prof. Vincenzo De Simone

Coordinator
Prof. Paolo Arcari

Academic Year 2010/2011

Riassunto

L' Iperlipidemia Familiare Combinata (FCHL) è una patologia complessa caratterizzata da elevati livelli sierici di colesterolo e trigliceridi, e riscontrata nel 20% dei pazienti con disturbi coronarici. I pazienti affetti da FCHL presentano inoltre insulino resistenza, obesità e ipertensione. Sono stati condotti diversi studi per la comprensione del meccanismo melecolare e genetico dell'FCHL anche se non hanno portato ad una comprensione chiara ed esaustiva, e sono stati individuati alcuni geni, come USF1 (upstream trascription factor 1), LDL-R (LDL receptor), ApoA1 e CRABP-II (cellular retinoic acid-binding protein 2), che potrebbero giocare un ruolo importante nella rete di geni coinvolti nell'FCHL.

La nostra attività sperimentale è stata improntata all'analisi dei profili di espressione genica in un gruppo di pazienti affetti da FCHL, mediante la tecnica dei DNA microarrays, allo scopo di individuare reti di geni coinvolti nella patologia. Questo lavoro è stato condotto partendo RNA di 10 pazienti affetti da FCHL comparato con quello di 5 controlli con un quadro lipidemico normale (esp. 10vs5), e di 7 di questi pazienti prima e dopo il trattamento con le statine (esp. 7vs7). I dati di espressione ottenuti sono stati analizzati mediante il software GeneSpring 7.3 and 9.0, che ha generato, per entrambi gli esperimenti, liste di geni la cui espressione risultava alterata in maniera significativa. Dall'intersezione delle liste di geni di entrambi gli esperimenti abbiamo ottenuto una lista finale di geni la cui

espressione risultava aumentata o diminuita nei pazienti affetti da FCHL e che cambiava in risposta al farmaco (statine), e alcuni di questi geni sono stati sottoposti a validazione mediante RT-qPCR.

Abbiamo condotto un'analisi dettagliata dei geni la cui espressione risultava alterata seguendo tre strade:

1. Analisi di GeneOntology.
2. Analisi dei network.
3. Analisi dei promotori.

L'analisi di GeneOntology per l'esp. 10vs5 mostra un'arricchimento, molto significativo (p-value 10^{-11}) di una famiglia di geni coinvolti nel metabolismo energetico; nell'esp. 7vs7 risultano arricchiti, con una stringenza media, gruppi di geni coinvolti nell'organizzazione del citoscheletro, nella morfogenesi e nella sintesi dei composti eterociclici.

L'analisi dei network mostra che la patologia maggiormente correlata con il nostro dataset di geni è l'artereopatia cardiaca, seguita da patologie di tipo epatico e renale, che sono complicanze tipiche della patologia.

Infine, l'analisi dei motivi, condotta mediante l'utilizzo del software MEME (Multiple Em for Motif Elicitation), mostra la presenza di ipotetici motivi di regolazione, che sono stati poi sottoposti ad analisi *in vitro* ed *in vivo* per verificare la loro capacità di agire come regolatori dell' espressione genica. Abbiamo effettuato

saggi EMSA (Electrophoretic Mobility Shift Assay) per vedere se questi motivi fossero in grado di legare proteine nucleari, e saggi di espressione transiente (CAT assays) per vedere se questi motivi fossero in grado di indirizzare l'espressione del gene reporter. Questi esperimenti hanno mostrato che, almeno uno di questi motivi, è in grado di agire come regolatore dell'espressione genica. Esperimenti di mutagenesi condotti sul motivo in esame hanno mostrato che la capacità regolatoria dello stesso è sequenza-specifica.

Nel complesso questi dati suggeriscono l'ipoespressione di due network di geni, coinvolti nella produzione di energia e nell'infiammazione, e che potrebbero giocare un ruolo importante nell'FCHL.

Summary

Familial Combined Hyperlipidemia (FCHL) is a complex disease characterized by elevated levels of serum total cholesterol, triglycerides or both, found in 20% of patients with coronary heart disease. FCHL patients are also affected from insulin resistance, obesity and hypertension. Several studies have been performed to understand molecular mechanisms and genetics of FCHL, but still nothing conclusive and exhaustive, although genes have been identified involved in disease, as USF1 (upstream trascription factor 1), LDL-R (LDL receptor), ApoA1 and CRABP-II (cellular retinoic acid-binding protein 2), that may be part a important role in genes networks of FCHL.

Our experimental activity has been focused to the analysis of gene expression profiles in a group of FCHL patients, using the DNA microarrays technique, to identify regulatory networks of genes involved in disease. We conducted a study starting from RNA of 10 FCHL patients compared with 5 normolipidemic controls (exp. 10vs5) and 7 patients by FCHL before and after treatment with statins (exp. 7vs7). Expression data were analyzed using the software GeneSpring 7.3 and 9.0, we generated a series of lists of genes, in both experiments, whose expression was changed significantly. By an intersection of the lists of both experiments we have obtained a final list of genes, whose expression is increased or decreased in FCHL

patients and that changes in response to the drug (statins), of which some were subject to validation by RT-qPCR.

A main focus of my work has been to detailed analysis of the genes whose expression was altered through:

1. Gene Ontology analysis.
2. Network analysis.
3. Motif analysis.

Gene Ontology analysis in 10vs5 exp. showed highly significant enrichment (p-value 10^{-11}) of a genes family involved in energy metabolism; in 7vs7 exp. are enriched, with medium stringency, groups of genes involved in the organization of cytoskeleton, morphogenesis and, synthesis heterocyclic compounds.

Network analysis shows that the disease more related with our genes dataset is cardiac arteriopathy, closely followed by kidney and liver pathologies , also typical complications of the disease.

Motif analysis, performed using the software MEME (Multiple Em for Motif Elicitation), shows some hypothetical motifs of regulation that we submitted to *in vitro* and *in vitro* experiments to verify their ability to act as regulators of gene expression. We performed EMSA assays (Electrophoretic Mobility Shift Assay) to test the ability of these “motifs” to bind nuclear proteins and transient expression assays (CAT assays) to show the ability of these motifs to drive expression of reporter genes. These experiments show that at least one of the motifs considered, its ability to act as gene expression

regulator. Mutagenesis experiments conducted on the motif considered showed that its regulatory function is sequence-specific.

Taken together, our data suggest that the down-regulation of two main gene networks, respectively in Energy production and inflammation, may be an important feature of the FCHL syndrome.

INDEX

Index of Figures and Tables

1. Introduction

1.1 *System Biology*

The, discovery of the double helix structure of DNA, by Watson and Crick, opened the way, to the understanding of the molecular basis of various aspects of biology, like diseases, heredity, development, etc.. Since then, molecular biology has made giant steps with the sequencing of many genomes, from the most simple like that of E.coli, to the most complex as those of mouse, monkey and finally the completion of Human Genome Project at the beginning of the new millennium. These new knowledges are leading us to conceive the molecular machinery of a cell not as the sum of individual components, but as a complex of integrated and interrelated molecular functions, giving rise to a new branch of Biology, the Systems Biology. Systems Biology encompasses several research areas : from molecular biology to bioinformatics, and also involves numerous disciplines such as genomics, transcriptomics, proteomics, interactomics, etc..

Investigation areas of Systems Biology are (1):

1. Structure systems analysis: it includes the interactions of genes, the biochemical pathways connecting them, the mechanisms of these interactions and how they modulate the physical structure of organisms.
2. System dynamics analysis: the understanding of how a system acts in time, how it reacts to specific stimuli and its sensitivity. The use of theoretical analysis and computer simulations are essential for this kind of analysis.

3. Control methods analysis: it identifies the control mechanism necessary to minimize malfunctioning of the system.
4. Systems design methods: the possibility to plan and construct biological systems with the desired properties through simulations.

Progress in one of these four areas of Systems Biology requires 360° innovation not only in molecular biology, but also in computational science and measurement technology, due to the high complexity of biological systems. The main goal of these approach is to identify the regulatory logic of genes and their biochemical networks.

One main tool to understand the complex molecular systems is networks theory. A network is represented as a graph, where the objects (*nodes*) are connected together by different kinds of *links*. Special nodes, with a high number of *links* and connected to a large number of *nodes*, are called *hubs*. In a gene regulatory network, nodes and links represent the effect (activation or repression) exerted by a gene product on the activity of another, and *Hubs* represent the key points of networks. There are several methods to understand and analyze genes regulatory networks, based on different computational and mathematical methods (2,3), which combine gene expression data with the results of proteome, interactome (4) or promoter analysis (5).

In these thesis, we will use the Ingenuity Pathway Analysis (IPA), one of the most commonly used tool for network analysis.

1.2 Familial Combined Hyperlipidemia (FCHL)

The hyperlipidemias include a heterogeneous group of disorders, characterized by an high concentrations of cholesterol and/or triglyceride in the plasma (6,7). Familial combined hyperlipidemia (FCHL), which is the subject of the present study, is the most common genetic disorder associated with premature cardiovascular disease (CVD), with an incidence of 1-3% in Western population. Familial combined hyperlipidaemia (FCHL) is a genetically complex lipid disorder, first recognised in 1973 by Goldstein et al, characterised by increased levels of plasma cholesterol or triglycerides in relatives of the same family. The intrafamilial variability of the lipid phenotype probably results from the interaction of multiple genes, some of which have been identified, with different environmental factors.

Several metabolic abnormalities have been described in FCHL patients, including very low density lipoprotein and apolipoprotein B (ApoB) overproduction, the presence of small dense low density lipoprotein (sdLDL), increased production of apolipoprotein C III, insulin resistance and obesity. All of them may contribute to the increased atherosclerotic risk associated with such a condition. (8).

Familial combined hyperlipidemia is a classic example of multigenic and multifactorial disease, occurring between second and third decade of life. Using the increase of VLDL, LDL, or both as a phenotype for family studies, Goldstein et al. (9) and Brunzell et al. (10), concluded that familial combined hyperlipidemia is an autosomal dominant condition with high penetrance. Brunzell et al.estimated that 10% of premature coronary artery disease is caused by FCHL (9,10).

FCHL patients synthesize and secrete increased amount of very low density lipoproteins (VLDL) in response to an increased flux of free fatty acids (FFAs) through the liver, which derives from the inability of adipose tissue of these subjects to incorporate FFA (11). The consequence of that is that FCHL patients present an interstitial accumulation of FFAs and pro-inflammatory adipokines in the adipose tissue, which in turn to reinforces the metabolic dysfunction.

Adipose tissue is one of the major contributors of free fatty acids (FFA) in the circulation. High levels of FFA in the circulation may lead to both a decrease in insulin-stimulated glucose uptake in skeletal muscle, and to an increase in hepatic lipoprotein synthesis, both characteristics of FCHL syndrome. Therefore, liver, adipose tissue, and muscle are interesting target tissues for differential gene expression studies (12).

The role of inflammatory processes in FCHL is unclear, but there are many inflammatory markers associated with familial combined hyperlipidemia (FCHL), like vascular cell adhesion molecule-1 (sVCAM-1), monocyte chemoattractant protein 1 (MCP-1), interleukin 6 (IL-6), tumor necrosis factor-α (TNF-α). The presence of these markers is independent from age, sex, body weight, insulin resistance, and metabolic syndrome (13).

To clearly define a disease so complex as FCHL is quite difficult. There are two basic criteria to diagnosticate this pathology:

1. Clinical: Findings of hypertriglyceridemia, hyperapobetalipoproteinemia, increased small and dense LDL (sdLDL) are crucial for FCHL diagnosis, however, non-essential but frequent characteristics as low plasmatic

concentrations of HDL, increased cholesterol, obesity and insulin resistance may help for diagnosis.

2. Genetics: Several studies have been performed to understand the molecular mechanisms and genetics of FCHL, but there are no conclusive answers yet. However, some genes involved in FCHL disease have been identified, as USF1 (upstream trascription factor 1), LDL-R (LDL receptor), ApoA1 and CRABP-II (cellular retinoic acid-binding protein 2) (14, 15, 16), may partecipate to genes networks that are affected in FCHL patients.

It is now clear that the genetics of FCHL is complex and the phenotype is heterogeneous. It is possible that the heterogeneity of the FCHL phenotype arises from a defect in more than one gene or, alternatively, that the disorder results from the interaction between one or more major genes and their "genetic" environment (12).

Recent trials show that high concentrations of small dense low-density lipoprotein (sdLDL) are highly specific markers of FCHL, independently of a concomitant metabolic syndrome (MS). In FCHL patients high levels of sdLDL are related to history of cardiovascular (CVD) events, independently of MS, total cholesterol and apo B (17).

Due to insulin resistance and obesity concomitance, a first approach to FCHL treatment is the achievement of an ideal weight through a balanced diet (complex carbohydrates, monounsaturated fats, etc.). If the balanced diet fails and the lipidic profile is still altered, drug therapy may be used to normalize lipidic profile. The most used drugs are the "fibrates" for a hypertriglyceridemic phenotype, and the "statins" for a

hypercholesterolemic phenotype. In this study we focalized the attention on the effect of statins on FCHL disease.

Statins act as inhibitors of the enzyme HMG-CoA reductase (3-hydroxy-3-methyl-glutaryl-CoA reductase), an enzyme that plays a central role in the biosynthesis of cholesterol in the liver, blocking the conversion of HMG-CoA in mevalonate, a precursor of cholesterol (Figure 1).

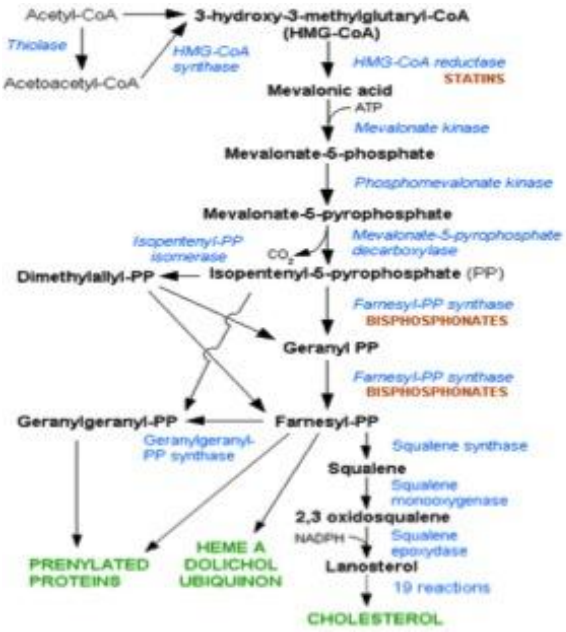


Figure 1: Cholesterol metabolism: the figure shows how the liver metabolism of cholesterol is blocked by statins, that inhibit the conversion of HMG-CoA in mevalonate, a precursor of cholesterol.

Statins also act by increasing the expression of LDL receptors, promoting a greater uptake of plasmatic LDL by hepatocytes, therefore resulting in a reduction of cholesterol levels.

Literature evidence also suggest a pleiotropic effect of statins, probably connected to the inhibition of isoprenoids synthesis, requiered for prenylation of several important proteins associated to the plasma membrane (18). An important consequence is the inhibition of isoprenylation of small GTP-ase (or small G proteins) like Rho, Ras, Rac and Rap (13) and the consequent inhibition of enzyme NADP-oxidase, resulting in a reduced production of superoxide ions (free radicals), probably at the basis of the alleged anti-inflammatory action of statins. All these actions would result in plaque stabilization and the improvement of endothelial functions.

Transcriptome analysis and gene expression profiles are a powerful tools for the identification of genes involved in complex diseases. This approach has already been used in several conditions, including diabetes mellitus (19), heart failure (20), and cancer (21), as well as in Tangier disease, a monogenic dyslipidemia (22).

1.3 DNA Microarrays

The main trend of the current postgenomic *era*, or the *era* of functional genomics is to expand the scale of biologic research from studying single genes or proteins to simultaneously studying all genes or proteins using a systematic approach. Among the methods for obtaining genome-wide mRNA expression data, DNA microarrays are particularly powerful since they can provide a global view of changes in gene expression patterns in response to physiologic alterations or manipulation of transcriptional regulators.

DNA microarray technology is a technique for comprehensive gene expression analysis and has been used in various fields, such as basic biology, medical science, and agriculture. In biomedical research, such an approach will ultimately define the transcriptional difference between normal and diseased tissues, which may provide insights into disease mechanisms and identify novel markers and candidates for diagnostic, prognostic and therapeutic intervention. However, microarray technology is still in a continuous state of evolution and development, and it may take time to implement microarrays as a routine medical device (23).

DNA microarrays can be divided into 1) small custom arrays designed to monitor expression of a few hundred genes, 2) very large arrays that represent tens of thousands of genes, 3) arrays that represent entire genomes. Presently, there are several applications for DNA microarray-based mRNA expression profile data, like the identification of genes whose mRNA levels are different under different biological conditions, e.g., in response to drugs treatments, in different cell types, or in particular mutants. Such genes are often considered candidates for playing an important role in the biological process of interest.

Due to the size and complexity of mRNA profile data, computational tools are required for analysis. These tools must be tailored according to the type of analysis being carried out. If the goal is to identify genes that show differential expression among different samples, statistical tools for significance tests and multiple tests correction are needed to sort genes based on the degree of likelihood that they are actually differentially expressed (24).

Arrays, or DNA *chips*, are a collections of microscopic DNA spots attached to a solid surface (glass, plastic, nylon or silicon chips), each DNA spot containing a specific DNA sequence, the *probes* (in-vitro or in-situ synthesized). These sequences, often synthetic oligonucleotides, are complementary to a portion of the transcribed region of a gene, and are used to hybridize, under high-stringency conditions, cDNA samples (*targets*), representing the transcriptome.

Oligonucleotide probes can be divided into two types: long (50 to 70 -mer) and short (25 -mer). Long oligonucleotide probes are common among DNA microarrays produced in house and are used on some commercial DNA microarrays, but they offer poorer discrimination than short oligonucleotide probes. As the hybridization specificities of short oligonucleotides are lower than long ones, having multiple probes per gene is essential. It also provides an advantage as expression values can be calculated more precisely.

Two-color and single-color methods: two-color or "two-channel" method is a procedure that analyzes two RNA samples on a single array, while a method that analyze one RNA sample on one array is called "single-channel", or one-color method. In two-channel methods, two mRNA samples are labeled with two different fluorochromes, typically Cy3 (green) and Cy5 (red), and the two labeled samples are competitively hybridized to a single array to obtain the ratio of the two mRNA levels. Since the array must be scanned at two wavelength channels because of the two different colors, this method is called "two-channel". In contrast, in the one-channel method each cDNA is hybridized to one microarray slide, obtaining a single

dataset for each sample. In this case the comparison between the two samples is performed "in silico" by the data analysis software. The "two-channel" method was initially developed because in the first spotted DNA microarrays the probe amounts could differ substantially from spot to spot. With the new commercial DNA microarrays the probe amounts are fairly homogenous, so the more robust single-channel methods can be used (Figure2).

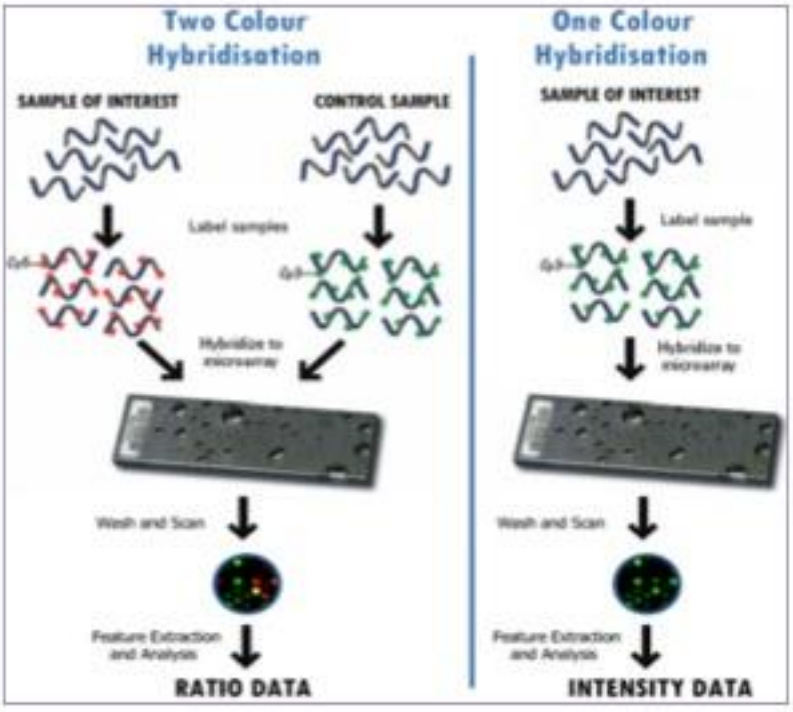


Figure2: "Two colours" (competitive) and "one colour" (single) procedures for cDNA hybridization on microarray slide. The data generated by the competitive procedure are expressed as ratio of fluorescence intensities, while those generated by the "one colour" approach are absolute intensity data.

There is a specific challenge to the two-channel method. While comparison of two samples hybridized to the same array can give good precision, comparison of two samples based on results of two arrays increases the error substantially, since two additional measured values are used in the calculation: To compare sample A and sample C based on the two sample ratios A/B and C/B, two sample B measurements are needed. Moreover, hybridizing all the sample pairs of interest would increase the number of arrays rapidly (25).

After hybridization of the fluorescent targets to DNA microarrays, and washing to remove aspecific material, the array is scanned and fluorescence image of the array are generated. Data pre-processing consists of the procedures that convert the fluorescence image of the array into the expression level values for each gene of the array. The pre-processing methods are different for single-channel and two-channel methods, however, there are some common steps used in both methods. The measured fluorescence intensity values in each array must be corrected for the background, which is caused by optical noise, non-specific hybridization, probe-specific effects, and measurement error.

Data from different arrays are usually not directly comparable even after background adjustment. For example, the overall fluorescence intensity among arrays typically varies. Ultimately, the goal is to compare the expression values of each gene from different arrays. The data from different arrays must first be “normalized”, so that they are comparable to each other. This “between-array” normalization is performed under certain assumptions. It is crucial that the assumptions made in a particular

normalization method are true for the data set of interest; otherwise, a normalization method can introduce artificial expression value changes. Common assumptions are that data from different arrays share the same value(s) of some descriptive statistics. The pre-processing and normalization procedures generate the expression data (dataset), that in a two-colours experiment are in form of expression ratios, while in a one-colour experiment consist in absolute intensity data. These data may then be subjected to further analysis, that are of two types:

I level analysis. The generated dataset of different samples are grouped in "experiments", in which two or more conditions are compared. Usually this analysis consists in a measurement of the expression changes (increase, decrease, no change) for each gene of the array in the two or more conditions being compared. In this step the extend of variations (fold change) and the statistical significance of the observed variations are assessed. The result of these analysis is usually a list of genes whose expression changes above or below a given threshold in a statistically significant way.

II level analysis. These list of "altered" genes can be subject to further analysis, in order to evidentiate group of genes whose expression profiles are similar (hyerarchical clustering), or belonging to a gene category that results significantly enriched (Gene Ontology analysis), or to networks in which the occourrence of these genes that belong to the list is significantly higher than expected (network analysis).

The **G**ene **O**ntology (**GO**) project is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases (26).

The project aims to:

- the development and maintenance of the ontologies themselves;
- annotate genes and gene products, and assimilate and disseminate annotation data;
- development of tools that facilitate the creation, maintenance and use of ontologies.

The ontology covers three domains: *cellular component*, *molecular function* and *biological process.* These three domains are composed of various sub-groups. This allows the user a subdivision of its genes list of analysis with different levels of specificity. Each domain, group and sub-group have a special symbol that will identify it.

In our study, we have performed all these analysis, and in addition we have carried out the search of putative DNA regulatory sequences in the promoters of these genes.

A key step of all studies of gene expressions is validation to confirm the obtained data. The principal methods used are: qRT-PCR and Northern blot.

2. Materials and Methods

My research activity has been directed to analyze the expression profiles of FCHL patients compared to that of normolipidemic controls, and that of the same patients after treatment with statins. Patients were selected by the team of Professor Paolo Rubba, Department of Clinical and Experimental Medicine University of Naples "Federico II". Out of 27 FCHL patients provided by the clinical team, we have selected 10 in order to minimize age (45-55) and severity of disease (middle) heterogenicity in the patients. We have therefore analized a group of 10 FCHL patients and compared their expression profiles with those of 5 normolipidemic controls (Exp. 10 vs. 5). A second experiment has been carried out with a group of 7 FCHL patients before and after treatment with statins.

2.1 RNA extraction

RNA extraction was performed starting from lymphocytes of selected patients with 30 ml of blood according by using the Qiagen RNeasy midi-extraction kit. Erytrocyte were lysed by the Qiagen EL buffer. RNA was then subjected to spectrophotometic quantitation and qualitative analysis through electrophoresis on formaldehyde agarose denaturing gel (FA).

2.2 DNA Microarrays

RNA samples extracted and purified were sent to the genomic service "Genecore" of the *European Molecular Biology Laboratory* (EMBL, Heidelberg, Germany), where they were subjected to hybridization, scanning and pre-normalization, using a single hybridization procedure ("*one color*") on CodeLink glass slides containing 56,000 oligonucleotides (60 bp) complementary to human transcribed sequences.

The samples subjected to microarray analysis were then analysed into two experiments:

- 10 vs 5: 10 FCHL patients compared to 5 healthy controls;
- 7 vs 7: 7 FCHL patients before and after treatment with statins (atorvastatin).

2.3 Data analysis: GeneSpring GX 7.3 software

Datas resulting from array scanning were subjected to pre-normalization using the analysis software CodeLink slides, and then organized in two different experiments "in silico":

- 10 vs 5: to analyze the changes of the transcriptome associated with the disease.
- 7 vs 7: to analyze the changes of the transcriptome after treatment with statins.

The analysis procedure used is identical in both experiments.

The pre-normalization was performed by using the *GeneSpring GX 7.3* analysis software of *Agilent Technologies*. Normalization phase was carried out using the default parameters of the software that normalizes the

total intensity of the sample "for Genes" and "for Chip", whereby it is assumed that all samples have an equal amount of starting mRNA. A filter to prevent a high number of false positives axcludes all the genes whose value was <0.01001, thus filtering off all the negative samples, to which the scanner assignes a default value of 0.01. First-level analysis of the dataset consisted in two main analytical procedures:

- *Fold Change (FC):* a filter that excludes those genes whose ratio between the expression values in the two conditions (FCHL vs. control in 10 vs 5; before and after treatment with statins in 7 vs. 7) is ≤ 2 or ≥ 0.5, in order to retain only those genes whose expression is increased or decreased by a factor 2.
- *Anova (analysis of variance):* performed on the genes lists already filtered by Fold Change. This analysis calculates the variance among the different measurements of each gene in a given group (ex. FCHL) and compares it with the variance obtained for the same gene of another group (ex. controls) determining whether the differences between the two variances are statistically significant, with a cutoff of p-value ≥ 0.05.

The first-level analysis generated two lists of hypo- and hyper-expressed genes, one for each experiment (see results).

Using a tool provided by the GeneSpring software (*Venn diagrams*) we then performed an intersection between the lists of two experiments "[10v5] ∩ [7v7]", generating a final list of 41 genes whose expression is altered in both experiments (see results).

2.4 Gene Ontology

We then submitted the lists from the two experiments to GO analysis in order to identify genes groups belonging to GO categories and subcategories, which are significantly enriched (p-value ≤ 0.05) .

For this analysis we used both the Gene Ontology tool provided by the GeneSpring software and a public domain tool, GOrilla (***G****ene* ***O****ntology en****RI****chment ana****L****ysis and visua****L****iz****A****tion tool*), which provides an output of Gene Ontology analysis in the form of *DAGs* (*Direct Acyclic Graphs*) simple and informative graphical representation (27).

2.5 Network analysis

We submitted the final lists of genes to IPA software (*Ingenuity Pathway Analysis*), a biological data analysis software from Ingenuity® Systems, which analyzes data from a variety of experimental platforms and provides accurate biological insight into the interactions between genes, proteins, pathways, cellular phenotypes, and disease processes in specific systems.

We were especially interested in analysing with this software the genes lists of the 10 vs 5 experiment, to establish if there are associations between this genes and pathological conditions related to FCHL that could corroborate the reliability of our experimental dataset.

2.6 RT-q PCR

The most commonly method used for microarrays data validation is qRT-PCR technique.

Reverse Trascriptase: Starting from RNA of both experiments, we prepared c-DNA sample to be used in pPCR reaction. For each sample we started with 10 μg of total RNA using 200U of M-MLV RT for each μg of RNA (Invitrogen ®) using Invitrogen protocol.

Primers: For each gene subject to validation was synthesized, at the oligonucleotides service of CEINGE-Advanced Biotechnology of Naples, a pair of 20bp primers, with a Tm (melting temperature) of 60° C and allowed to amplify a region of ≈ 200bp corresponding to the center of the transcribed region to be analized.

PCR: To test oligonucleotides and the presence of homogeneous amplification we performed qualitative PCR reactions using Taq DNA polymerase (EuroClone ®) following the EuroClone protocol. Based on this screening, five pairs of primers were excluded from the final analysis of "Real-Time" PCR due to absence of amplification signal, or heterogeneity, of the amplified DNA.

qPCR: Validation was then performed on 11 genes, starting from 100 ng of c-DNA using 11 primer pairs plus a pair of primers for the Abl gene as an internal control.

qPCR was performed with the SYBR Green method, using iQ SYBR Green Supermix enzyme of BioRad®.

qRT-PCR consists of two main steps:

- Denaturation at 90°C.
- Annealing and extension at 60°C.

The samples were subjected to 40 amplification cycles. To check the efficiency of primers we used a tool of the I-Cycler software which measures the "Melting Curves" of the sample, allowing us to pinpoint heterogeneous amplificates.

Two RNA samples from each patients or control were analysed, and each measurement was performed in triplicate. Therefore, each gene was analysed 10x2x3= 60x in FCHL patients and 30x in controls.

The qPCR analysis allows the identification of the cycle where the fluorescence exceeds the background. This cycle is called "threshold cycle" (**Ct**). The difference between the Ct value of the sample of interest (the average of the 2x3 measurements) and the *housekeeping* gene (Abl gene) gave us the **ΔCt** value. The next step is the comparison of the expression levels in the two conditions (ex. Ct of FCHL gene less Ct of control gene), obtaining the ΔΔCt. Finally, everything is expressed as a negative exponential in base two to get the value of Fold Change: $\mathbf{FC= 2^{-\Delta\Delta Ct}}$**.**

2.7 Motif analysis (MEME)

The analysis of the promoters of our genes in order to identify putative regulatory sequences was carried out with the help of the software on-line MEME (**M**ultiple **E**m for **M**otif **E**licitation), which uses an algorithm that is able to identify, within a set of user-supplied data, short sequences which are over-represented significantly.

We performed this analysis with the 41 genes resulting by the intersection of the lists of the two main experiments [(10vs5) ∩ (7vs7)]. We uploaded into the program the promoters sequences (FASTA format) of

the intersection gene lists in an ranging from -400 to +1 relative to trascription start site, setting the parameter "length of motifs" to 6-15bp. For the other parameter, like repetitions number, sites number and maximum number of searched motifs the default settings are used. The output of the software, both graphical and tabular, provides the following informations:

- Motifs sequences, with a *core* and flanking sequences.
- E-value of identified motif in the set of genes, gives by the sum of p-values of the motif of the individual genes, an important information about the significativity of the motif.
- Preservation of identified motif in the set of genes and its position (comparison matrix).

This software also enables to compare the motifs found with the binding motifs for transcription factors already present in databases such as Jaspar and Transfac, using the TOMTOM tool.

2.8 Nuclear extracts

Nuclear protein extraction was performed starting from HepG2 cells (Hepatocellular carcinoma) using the Mini Dignam protocol (28):

- Cell coltures at 80% confluence were harvested using a cell scraper, washed in 30 volumes of phosphate-buffer-saline (PBS), centrifuged for 5 minutes at 1.200 rpm, resuspended in 1 volume of buffer A (10 mM Hepes pH7.9, 1,5 mM MgCl2, 10 mM KCl, 0,5 mM DTT) for lysis and incubated 15 minutes on ice. Cells were then lysed by 5-10 rapid strokes through a narrow gauge hypodermic needle (25-g ⅝).

- The cell homogenate is centrifuged for 20" at 12.000 rpm, the nuclear pellet is resuspended in 2/3 V of Buffer C (20 mM Hepes pH7.9, 25% Glycerol, 420 mM NaCl, 1,5 mM MgCl2, 0,2 mM EDTA, 0,5 mM PMSF, 0,5 mM DTT, 1% Aproteinin) and incubated on ice with stirring for 30 minutes.
- The nuclear debris are then pelleted by spinning for 5 minutes at 12.000 rpm and the supernatant (nuclear extract) can be stored at -80°C. The concentration of nuclear extract eas determined with the "Bio-Rad Protein Assay" based, on Bradford method (29).

2.9 EMSA

EMSA assay (***E****lectrophoretic* ***M****obility* ***S****hift* ***A****ssay)* is one of most used methods to study sequence-specific interactions between DNA and proteins. This assay requires the use of a terminally labeled DNA fragment which is interacting with the proteins of interest (in our case a crude extract of nuclear proteins). If the protein recognizes the binding site to the DNA fragment it binds and we observe a delay in electrophoretic migration of the DNA fragment due to the binding protein (or mixture of proteins) significantly delayes the bound DNA compared to the same unbound DNA fragment.

Oligonucleotides synthesis:

MEME analysis allows to identify a series of "motifs" present in the genes of our interest. We designed and sinthetized a pair of complementary DNA oligos containing the identified motifs. Two version for each sequence, based on the most conserved flanking region of the genes

containing the motif, were chosen and sintetized. The two olinucleotides (25bp) were designed so as to have, after annealing of the complementary sequences, a 5'-GATC- single strand sequence at both ends, such as to permit their use in the experiments later described. The oligonucleotides were synthesized at the oligonucleotides service of CEINGE-Advanced Biotechnology of Naples.

Oligo 1: 5′-GATCGTAGCTGGGCGTGGTCGGGTC-3′
3′-CATCGACCCGCACCAGCCCAGCTAG-5′
Oligo 2: 5′-GATCGCTCCTGGGGGTGTTCGGGGT-3′
3′-CGAGGACCCCCACAAGCCCCACTAG-5′
Oligo 3: 5′-GATCGCAGCTCTTTGCTTTGACAGA-3′
3′-CGTCGAGAAACGAAACTGTCTCTAG-5′
Oligo 4: 5′-GATCGGAGCTCTTTCCTTAGTTTGT-3′
3′-CCTCGAGAAAGGAATCAAACACTAG-5′
Oligo 5: 5′-GATCAAGATGAATTTCAGCAGGACC-3′
3′-TTCTACTTAAAGTCGTCCTGGCTAG-5′
Oligo 6: 5′-GATCCTGAAGAATTTCTGAAACAAG-3′
3′-GACTTCTTAAAGACTTTGTTCCTAG-5′

Labeled of probes using fill-in:

This reaction allows the labeling of the 3' termini of each filament of double-stranded DNA using the Klenow fragment of E. coli DNA polymerase I.

The labeled protocol is the following:

- 1 μg of *ds*DNA
- Appropriate Buffer for Klenow fragment
- 2 μCi (2000 Ci/mmole) of [α-^{32}P] dATP
- 10 Units of Klenow fragment of E. coli DNA polymeraseI, incubated for 45 minutes at room temperature.
- Stop the reaction by heating it for 5 minutes at 70°C.
- Separate the labeled DNA from unincorporated dNTP by chromatography on small columns of Sephadex G-50.

- Apply the radioactive probe (200 µl) to the top of G50 column. Fill the column with TE 1X buffer.
- Start to collect fractions (≈ 200 µl) in microfuge tubes, measuring the radioactivity in each tubes. The leading peak of radioactivity consists of nucleotides incorporated into DNA, and the trailing peak consists of unincorporated [α-^{32}P] dATP.
- Pool the radioactive fractions in the leading peak and store at -20°C.
- Alternatively, use a spin-column procedure.

DNA Fragment Retardation Assay:

3000 cpm of the endlabelled probe are incubated with the nuclear extract of interested (≈15 µg) and the appropriate buffer (10X binding buffer; buffer D). As carrier are used 100-300 µg/ml of poly [dI/dC]. The sample are incubated at 25°C for 25 minutes and 5 minutes on ice. After incubation are added the appropriate volume of loading buffer and the samples are loaded onto a 5% (29:1 Acrylamide/Bis-Acrylamide) native polyacrylamide gel. Electrophoresis is carried out in 1X TBE buffer at 10V/cm for ≈ 2-3h. After electrophoresis the gel is dried with a gel dryer and exposed on Bio-Rad Imaging Screen K. After exposure, image is acquired by scanning with Bio-Rad phosphor-imager.

Buffer conditions:

10X Binding Buffer (BB):

- 1M KCl
- 20mM MgCl2
- 40mM Spermidine
- 1mg/ml BSA

Buffer D:

- 20mM Hepes-KOH pH 7.9

- 20% Glycerol
- 20mM KCl
- 2mM MgCl2
- 0.2mM EDTA
- 0.5mM DTT

Loading buffer:

- 40% Glycerol
- 5x TBE pH 8.3
- 50mM EDTA
- Bromophenolblue

2.10 Preparing the plasmids for transient expression assays

The oligonucleotides used in EMSA experiments were also used to perform transient expression assays in human cells, to determine if these motifs are capable to act, *in vivo*, as regulators of transcription.

After annealing the oligonucleotides are cloned into a expression vector as monomers or polymers. The vector used is pTKsh-CAT (4421 bp), in which reporter gene (CAT gene), coding for enzyme Cloramfenicol Acetyl Transferase, is located under the control of a minimal promoter (SP1+TATA) derived from the Thymidine Kinase gene of Herpex Simplex Virus (HSV-TK). Upstream of the promoter is located the pUC19 polylinker, that has many unique restriction sites, including the BamH1 site whose sticky ends are complementary to the *ss* extremities of our oligos. After oligos annealing we performed a phosphorylation reaction, to add a phosphate group at the 5 ' of oligos with the enzyme polynucleotide kinase T4, that catalyzes the transfer of terminal phosphate (γ) from ATP to the 5'-OH of our ends oligos. pTKsh-CAT was digested with BamH1 restriction enzyme, purified on low-melting agarose gel and treated with alkaline

phosphatase CIP (Calf Intestinal Phosphatase) to remove 5' phosphate of vector to avoid closing of the empty digested vectors.

We then performed a ligase reaction between vector and oligos, in a 1:1 ratio, using T4 DNA ligase,. Ligase reaction DNA mixture was then inserted, by transformation, into "competent" bacterial cells of E. coli, strain DH5α, to assumption of exogenous DNA. DNA from picked positive colonies was extracted by QIAprep ® Spin Miniprep Kit using Qiagen protocol.

To verify if the fragment of interest was inserted in properly we digested the mini-preps of plasmid DNA with the restriction enzyme Pst1, which cuts into our vector upstream and downstream of the BamH1 site, used to cloning DNA fragment. DNA presenting insert were sequenced at the sequencing service of CEINGE-Advanced Biotechnology of Naples.

2.11 Transient expression assays

Calcium phosphate transfection:

Transfection was performed in two eukaryotic cell lines: HeLa and HepG2. The cells are grown at 37°C and 5% of CO2 in DMEM (Dulbecco's Modified Eagle Medium, Gibco) culture medium, supplemented with fetal bovine serum (FBS 10%), antibiotics (penicillin/streptomycin 1%) and L-glutamine (1%).

Transfection was performed using the DNA co-precipitation with calcium phosphate technique. Each sample was transfected in duplicate. Cells were trypsinized and cells are resuspend in the amount of DMEM desired, distributed in the plates, and incubated for approximately 3 hours

before transfection. 5 μg of DNA are transfected in each plate (6cm diameter), using the calcium phosphate methods (30) and incubate O/N. After 12-16 hours transfected cells are washed with PBS-EGTA (3 mM) in order to remove the precipitate that did not enter the cells and that may be toxic and supplemented with fresh culture medium. After 24 hours the cells, washed with PBS, are harvested by adding directly to the plate a solution containing 40 mm Tris pH 7.6, 1 mM EDTA, 150 mM NaCl (T*en* solution). After incubating for 5 minutes at room temperature, the cells are collected by pipetting, centrifuged at 13,000 rpm for 1 minute (4 °C) and the cell pellet is resuspended in 200μl of 0.25 M Tris pH 7.8 (cold). The cells are then lysed by 4 cycles of freezing and thawing. After centrifugation for 10 minutes at 13,000 rpm (4 °C) the supernatant containing cytoplasmic protein us collected and stored at -80 ° C.

Chloramphenicol AcetylTransferase Assay (CAT assay):

The chloramphenicol acetyltransferase (CAT) is a bacterial gene that encoding for an enzyme that neutralize the antibiotic chloramphenicol, by the addition of acetyl groups. This gene is not present in eukaryotes, therefore, eukaryotic cells-extracts do not show any background CAT activity. CAT gene is one of the first reporter genes used for transient expression experiments in mammalian cells. Important feature, of this system, is the extreme sensitivity of the assay and the absolute lack of background activity.

CAT assays offer an indirect but quantitative way to measure the amount of transcription driven from any given promoter.

Introduction of putative regulatory sequences in vectors containing the CAT gene and transfection in cells culture allows to evaluate the ability of these sequences to drive the expression of CAT gene and therefore to act as transcriptional activators.

CAT gene activity can be monitored using TLC (Thin Layer Cromatography), a chromatography based on distribution of different substances between a mobile (a mixture of solvents: chloroform-methanol 19:1) and stationary (cellulose) phase.

Protocol:

- Heat at 37°C for 5 minutes a mixture with:
  - 95 µl cellular extract
  - 60 µl Tris 0.25 M pH 7.8
  - 2.5 µl ^{14}C-chloramphenicol (25 µCi)
- Add 20 µl of 4 mM acetyl CoA
- Keep at 37°C for 30 minutes
- Extract with 500 µl ethyl acetate
- Spin for 1 minute (4°)
- Extract with another 500 µl ethyl acetate. Respin for 1 minute (4°)
- Dry 1 ml in the Savant speed vac
- Dissolve with 20 µl of ethyl acetate
- Spot 20 µl on Polygram SIL-G thin layer chromatography sheet
- Run the chromatography with 200 ml of a (95:5) chloroform-methanol solution for 2h.
- Dry the sheet on air
- Expose to Bio-Rad Imaging Screen K.
- Scanning image with Bio-Rad phosphor-imager.

3. Results

3.1 First-level analysis

First-level data analysis was performed using *GeneSpring GX 7.3* software of *Agilent Technologies* (see Materials and Methods), applying the *Fold Change(FC)* and *Anova* filter (Figure 3 a-b) to both 10vs5 and 7vs7 experiments. For details, see Materials and Methods section 2.3, page 17.

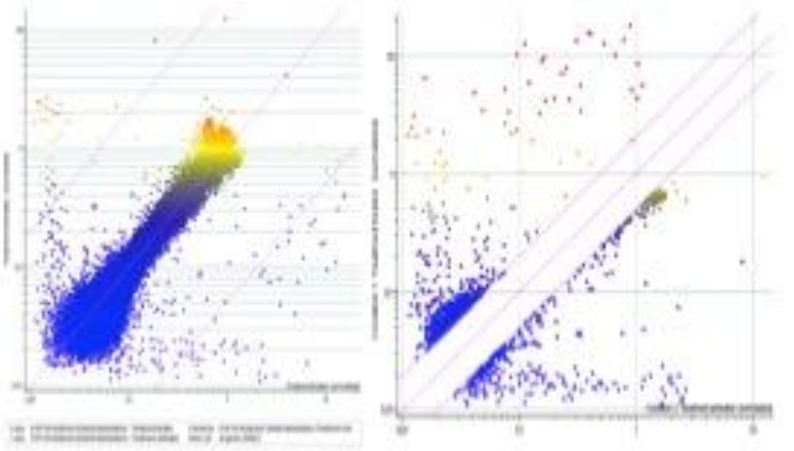


Figure 3a: Scatter plot of the experiment 7vs7 before and after FC filter

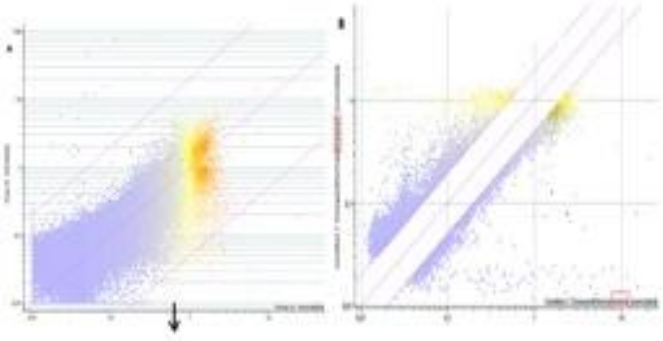


Figure 3b: Scatter plot of the experiment 10vs5 before and after FC filter

This analysis generated two gene lists: a 1747 genes list, for the "10vs5" experiment (1534 genes hypo-expressed and 213 genes hyper-expressed), and a list of 460 genes for the "7vs7" experiments (296 genes hyper-expressed and 164 genes hypo-expressed). Using the GeneSpring "*Venn diagrams*" tool (Figure 4) we have intersected the two gene lists and generated a final list "[10v5]∩ [7v7]" (table 1) of 41 genes differentially expressed in both experiments (Figure 5). The two "main" list (10vs5 and 7vs7) contain too many genes to be really useful. Therefore, any further discussion about these lists will be refered to the II-lvel analysis (GO, IPA and MEME). On the other hand, the "intersection" list contain a workable (41) number of genes, whose expression is altered in FCHL patients and restored upon statin treatment. Among these genes, there are some noteworth cases:

1. ATP11B, a class VI ATPase, hypoexpressed (0,07x) in FCHL patients, that recovers after treatment (9.16x).
2. SLC25A37, a Fe^{++} carrier protein, which is heavily hyperexpressed (70x) in patients and promptly reduced by statins (0,008x)
3. HNF1α, a liver-specific transcriptional regulator, repressed in patients (0,08x) and restored by statins (6,7x).

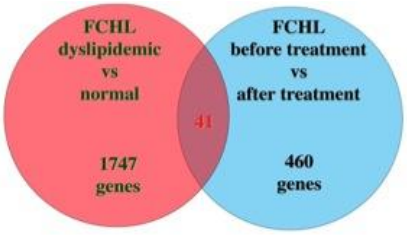


Figure 4: *Venn diagrams* showing the intersection of two gene lists [10v5]∩ [7v7], resulting in a final list of 41 genes (see table 1).

Genbank	Description	FC 10v5	FC 7v7
NM_014243	ADAM metallopeptidase with thrombospondin type1 motif,3	0,387	3,178
CD245683	Aquaporin 1 (Colton blood group) (AQP1)	0,265	3,318
BI757158	ATPase, class VI, type 11B (ATP11B)	0,0715	9,159
NM_033028	Bardet-Biedl syndrome 4 (BBS4)	4,307	0,157
NM_030809	Cysteine-serine-rich nuclear protein 2 (CSRNP2)	30,48	0,0148
NM_000716	Complement component 4 binding protein, beta (C4BPB)	0,452	3,203
NM_001257	Cadherin 13, H-cadherin (heart) (CDH13)	0,328	2,522
NM_000076	Cyclin-dependent kinase inhibitor 1C (p57, Kip2) (CDKN1C)	0,376	0,476
NM_003663	CGG triplet repeat binding protein 1 (CGGBP1)	14,1	0,0487
NM_000497	Cytochrome P450, family 11, subfamily B, pp 1 (CYP11B1)	0,377	2,752
X62515	Heparan sulfate proteoglycan 2 (HSPG2)	0,41	2,013
AI680974	WDR45-like (WDR45L)	3,157	0,16
NM_145258		2,608	0,473
AA916572	Solute carrier family 25, member 37 (SLC25A37)	69,6	0,00837
NM_002864	Pregnancy-zone protein (PZP)	0,066	7,033
AF153482	Ribosomal protein L41 (RPL41)	10,37	0,0921
NM_014412	Calcyclin binding protein (CACYBP)	0,00297	100,5
BE265261	Solute carrier family 25 member 3 (SLC25A3)	0,00356	142,7
NM_000545	HNF1 homeobox A (HNF1A)	0,0845	6,716
NM_003394	Wingless-type MMTV integration site family (WNT10B)	0,0963	10,35
BM727911	Transcribed locus	70,35	52,69
BE857924	Transcribed locus	51,82	25,61
AI827457		3,553	16,56
BC023640	Homo sapiens, clone IMAGE:4890344, mRNA	2,195	12,05
BQ187437	Hypothetical LOC728804 (LOC728804)	2,044	11,71
AK056249	CDNA FLJ31687 fis, clone NT2RI2005473	0,453	6,812
AW295445	Transcribed locus	0,442	6,145
AI732747	Transcribed locus	0,433	4,724
AI365358	Transcribed locus	0,424	3,336
AI376656	Transcribed locus	0,39	2,902
AW139602	Transcribed locus	0,376	2,617
CB047896	Transcribed locus	0,373	2,505
AI744743	Transcribed locus	0,276	2,383
BG151547	Transcribed locus	0,274	2,353
AW445216	Transcribed locus	0,0904	2,274
AA776532	Transcribed locus	0,0793	2,148
AI652378	Transcribed locus	0,0652	0,467
BF195078		0,0459	0,372
AI809873		0,0321	0,357
AW291006	Small nucleolar RNA, H/ACA box 58 (SNORA58)	0,0186	0,0155
BX096208		0,00834	0,00908

Table 1: The 41 genes differentially expressed in the intersection of the two experiments (10v5 and 7v7).

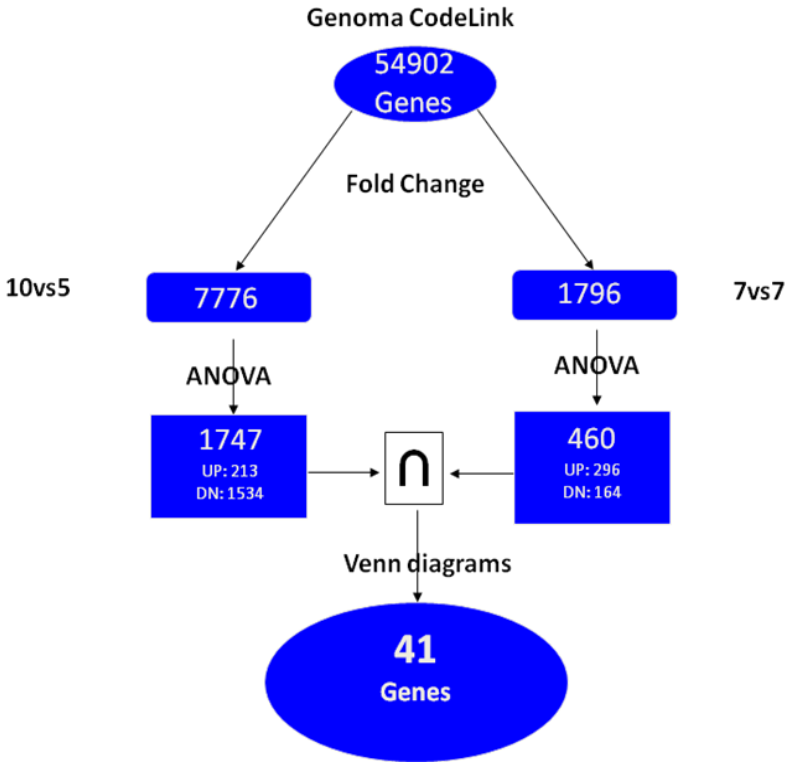


Figure 5: Workflow of I level data analysis of both experiments.

3.2 Gene Ontology analysis

We have submitted the gene lists of the two experiments to Gene Ontology analysis tools. Figure 8 reports the results obtained with the software on-line GOrilla (***G**ene **O**ntology en**RI**chment ana**L**ysis and visua**L**iz**A**tion tool*) that provides an output in form of *DAGs* (*Direct Acyclic Graphs*). Genes clusters significantly enriched have a more intense color based on p-value.

The analysis of the 10vs5 list shows that a group of 14 ATP synthase genes, involved in the transport of protons in the mitochondria, is significantly (p-value $<10^{-11}$) enriched. (Figure 6).

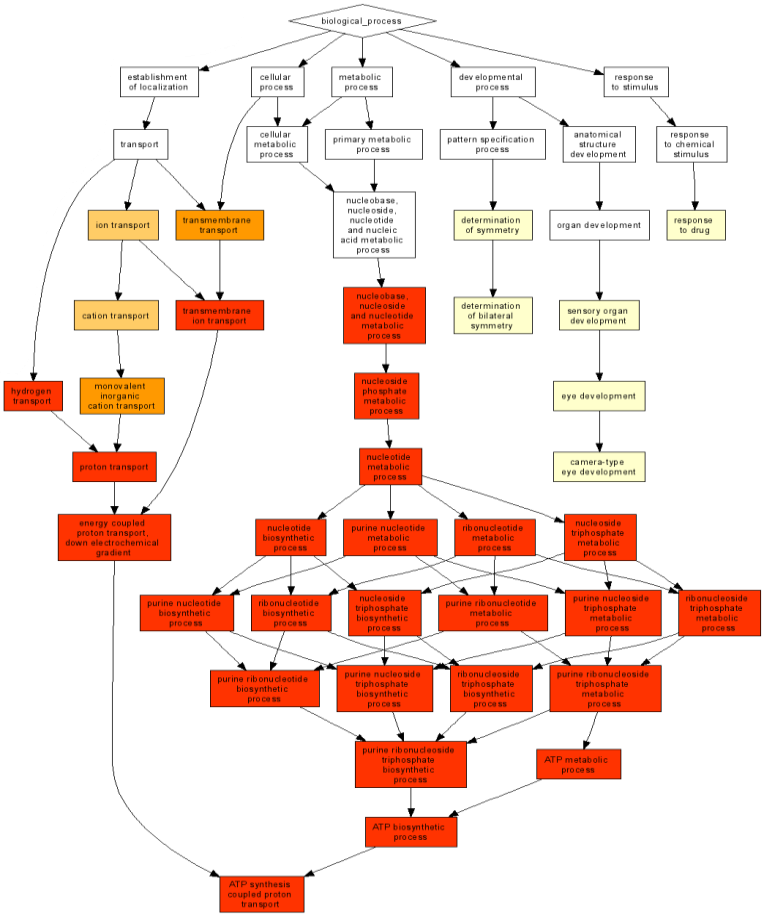


Figure 6: The DAG (Direct Acyclic Graphs) of the biological process GO category resulting from the analysis of the 10vs5 gene list. A group of 14 genes, belonging to the ATP synthase family, is strangly enriched (p-value $<10^{-11}$).

Gene Ontology analysis of the 7vs7 list shows an enrichment of three groups of genes (p-value $<10^{-6}$), respectively involved in cytoskeleton organization, morphogenesis and synthesis of heterocyclic compounds (Figure 7).

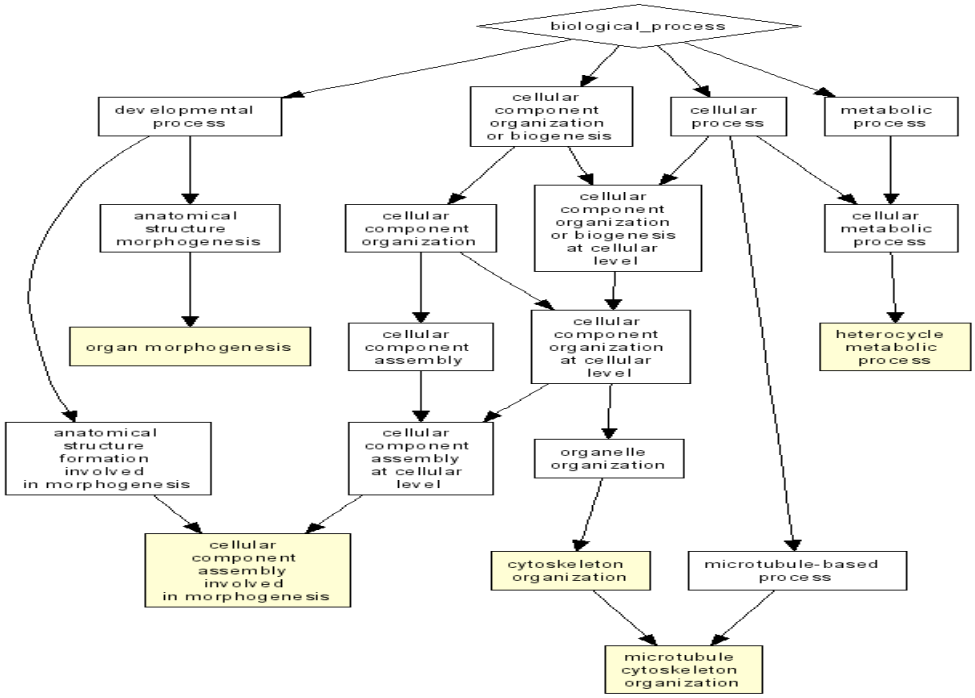


Figure 7: The DAG (Direct Acyclic Graphs) of the biological process GO category resulting from the analysis of the 7vs7 gene list. There are 3 group, of 4 genes, involved in cytoskeleton organization, morphogenesis and synthesis of heterocyclic compounds.

3.3 Pathway analysis

The 10vs5 gene list has been subjected to pathways analysis by using the IPA (Ingenuity Pathway Analysis) software. The results can be summarized as follows:

A) The search for pathologies significantly associated to our list of genes, whose results are reported in figure 8, gave us as a result three pathologies well above the significativity treshold: cardiac, liver and renal disease. These pathologies correspond exactly to the three main complications of the FCHL syndrome, thus confirming that our expression profiles identify correctly the gene expression changes associated to the FCHL condition.

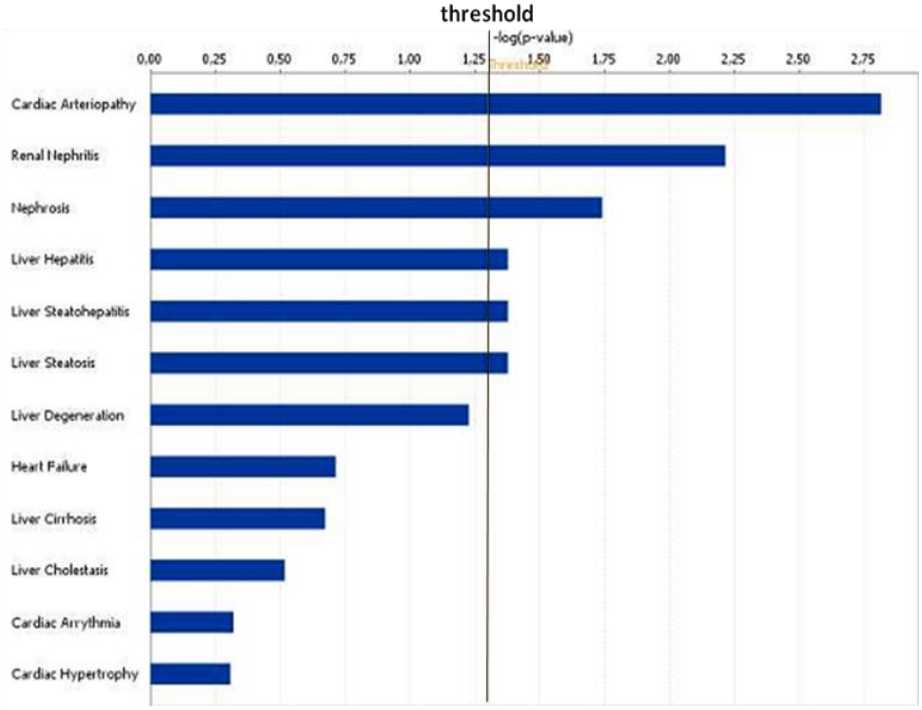


Figure 8: disease related to gene list of intersection (**I**ngenuity **P**athway **A**nalysis).

B) The search for metabolic pathways indicate that 4 out of the 5 enzyme complex involved into mitochondrial electron transport are affected, in FCHL patients, in agreement with the gene-ontology data (Figure 9).

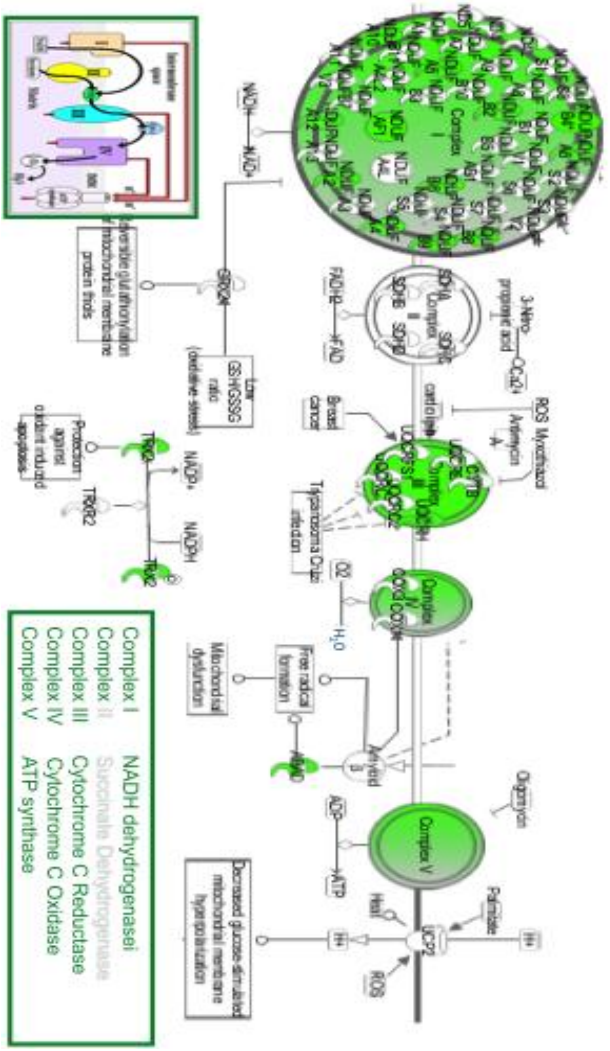


Figure 9: Metabolic networks associated to 10vs5 gene lists.

C) Among the other regulatory pathways that are affected in FCHL patients there is a network connecting lipid metabolism inflammation and apoptosis, in which the heavily hyperexpressed SLC25A37 gene plays a pivotal role (figure 10).

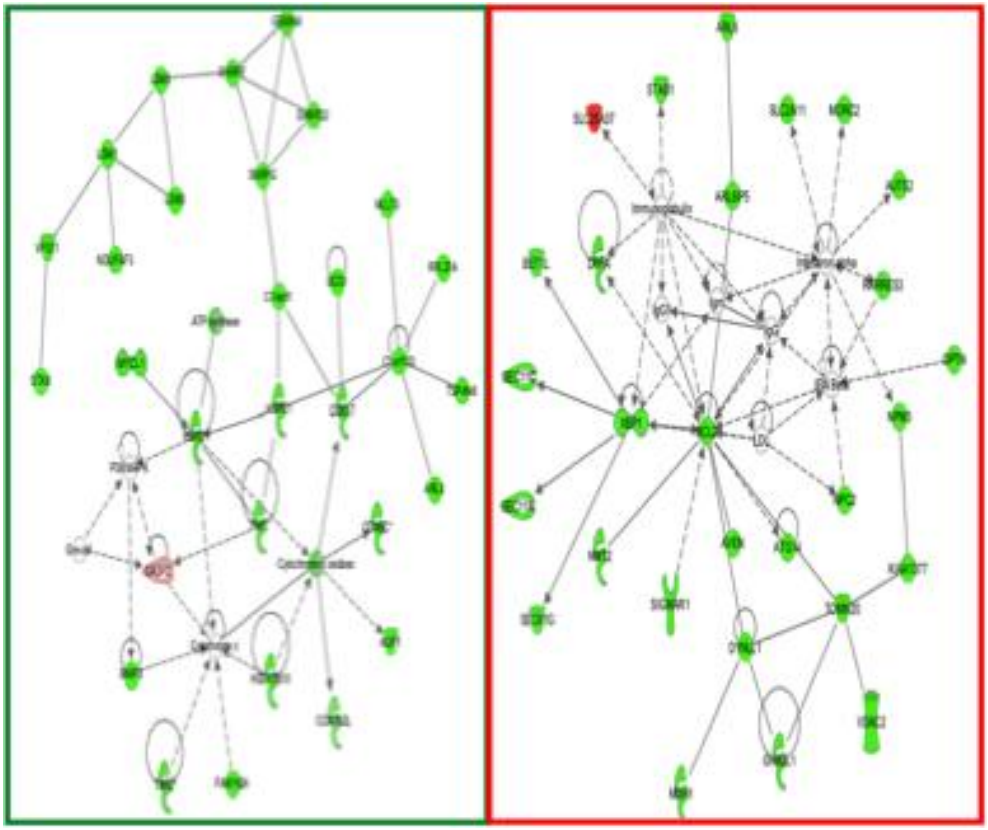


Figure 10: Networks of gene hypoexpressed in FCHL patients that are involved in energy metabolism, lipid metabolism, inflammatory molecules and apoptosis.

3.4 Validation

Microarrays data have been validated by qRT-PCR technique. Validation was performed on 16 genes (table 2), some of them belonging to the intersection list of two experiments, other to the 10vs5 list, in particular those with extreme FC values.

Gene Bank	Common name	Description
AW938887	SIP	Siah-interacting protein
NM_016355	DDX47	DEAD (Asp-Glu-Ala-Asp) box polypeptide 47
BE265261	SLC25A 3	solute carrier family 25 , member 3
BM727911	Transeq	Transcribed locus
NM_000006	TNFRSF 21	tumor necrosis factor receptor superfamily, member 21
NM_000545	TCF1	HNF1 homeobox A
AF533922	HLA-DQA1	major histocompatibility complex, class II, DQ alpha 1
NM_000011	ZNF143	zinc finger protein 143
NM_021104	RPL41	ribosomal protein L41
NM_005546	ITK	IL2-inducible T-cell kinase
NM_000075	ALAS1	aminolevulinic acid synthase 1
NM_000002	PIG3	tumor protein p53 inducible protein 3
AF158747	GALNT1 0	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 10 (GalNAc-T10)
NM_00007	CXCL9	chemokine (C-X-C motif) ligand 9
NM_002207	ITGA9	integrin, alpha 9
BU675991	HRB2	HIV-1 Rev binding protein 2

Table 2: Lists of the genes submitted to validation by qPCR.

Five out of the 16 genes (PIG3, GALNT10, CXCL9, ITGA9 e HRB2) were excluded from the final step of validation because RT-PCR pre-

screening gave negative results (amplified DNA band heterogeneous or absent).

RT-qPCR experiments resulted in the validation of the expression pattern of 7 out of the 11 genes analyzed, 5 hypo-expressed and 2 hyper-expressed (Figure 11). We assumed as validated the expression pattern of a gene when the two methods, namely microarrays analysis and qPCR, gave the some qualitative (i.e. significantly hyper or hypo-expressed) results. However, quantitative data also show the same relative range of FC varitions. More than >60% of the genes analyzed were validated, in line with the results published in literature for similar experiments (31).

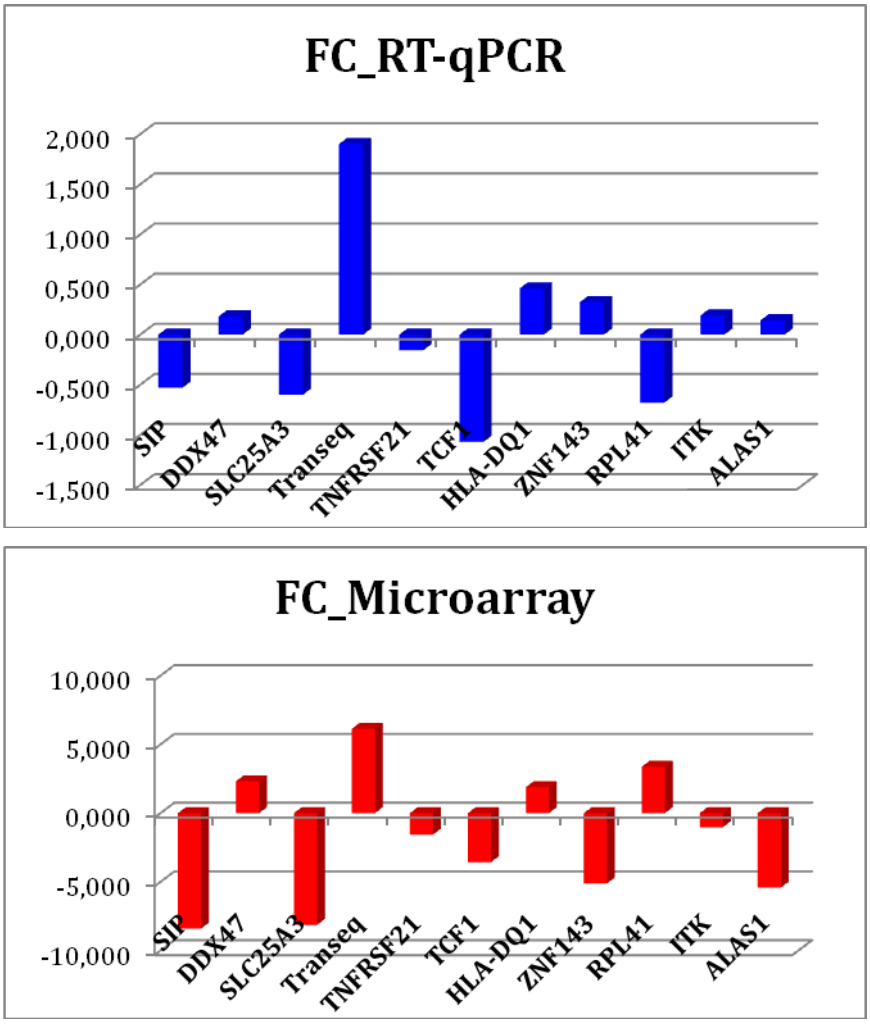


Figure 11: The Fold Change values of the 11 genes submitted to validation by qRT-PCR (blue) as compared to Microarrays Fold Change values (red): 7 out of 11 genes were validated (>60%).

3.5 Promoter analysis

Promoter analysis was performed "in silico" by MEME software (http://meme.sdsc.edu/meme/cgi-bin/meme.cgi). We submitted to the MEME software the 5' flanking sequences (-400 to +1 from trascription start site) of the 41 genes of the intersection list. The putative "motifs" (Figure 12) were analysed with the CisRed software, and each "motif" was used as a query to search Jaspar and TransFac transcriptional regulatory element databases for similarity to known regulatory sequences. From these results, we selected 3 putative regulatory motifs (Figure 13), to be submitted to in vivo and in vitro analysis.

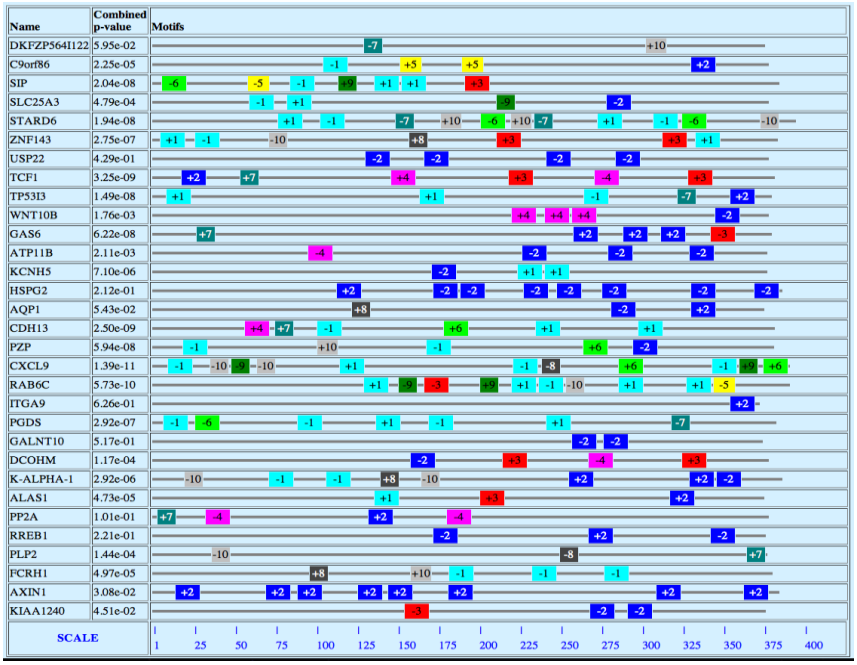


Figure 12: Combined block diagram generated by MEME analysis for the -400/+1 regions of the promoters of intersection genes list [(10vs5) ∩ (7vs7)].

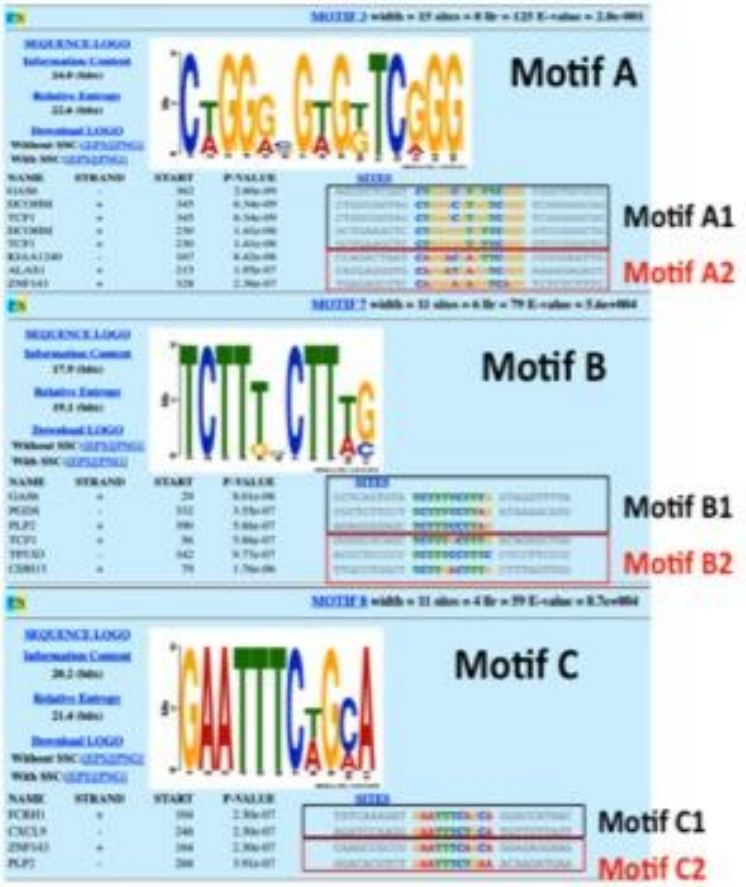


Figure 13: Three putative regulatory "motifs". The sequence of the original context in which these motifs have been found are shown (BOX). For each motif, two "versions" have been selected for oligonucleotide sinthesys.

3.6 EMSA analysis

The oligonucleotides synthesized for each motif (two versions for everyone, see Mat. and Met. section), were used for *in vitro* analysis by EMSA (Electrophoretic Mobility Shift Assays), in order to verify their ability to bind proteins present in nuclear extract from different cell lines. All the oligonucletides were positive to EMSA assays (figure 14).

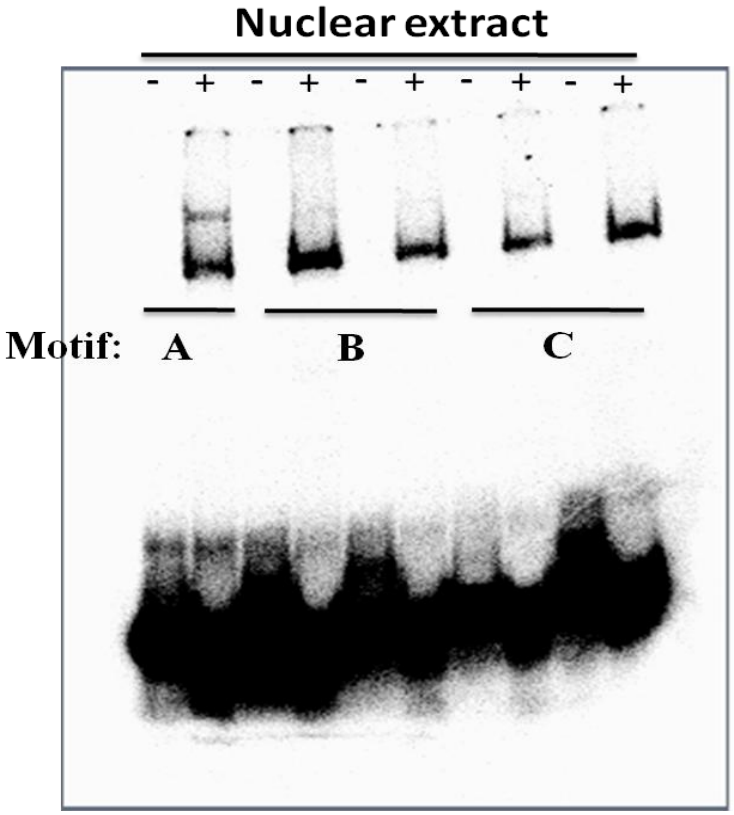


Figure 14: EMSA assay showing that all putative motifs (A, B and C) present a retarded band when challenged with HepG2 nuclear extracts. Sample are loaded in presence (+) and absence (-) of protein extracts.

Figure 15 reports an EMSA experiment carried out with the A1 Motif, showing that the retarted band is specifically competed by increasing amount the same or a related (A2) unlabeled oligo, but not by a different sequence. These results show that the binding of protein present in the nuclear extract is sequence-specific.

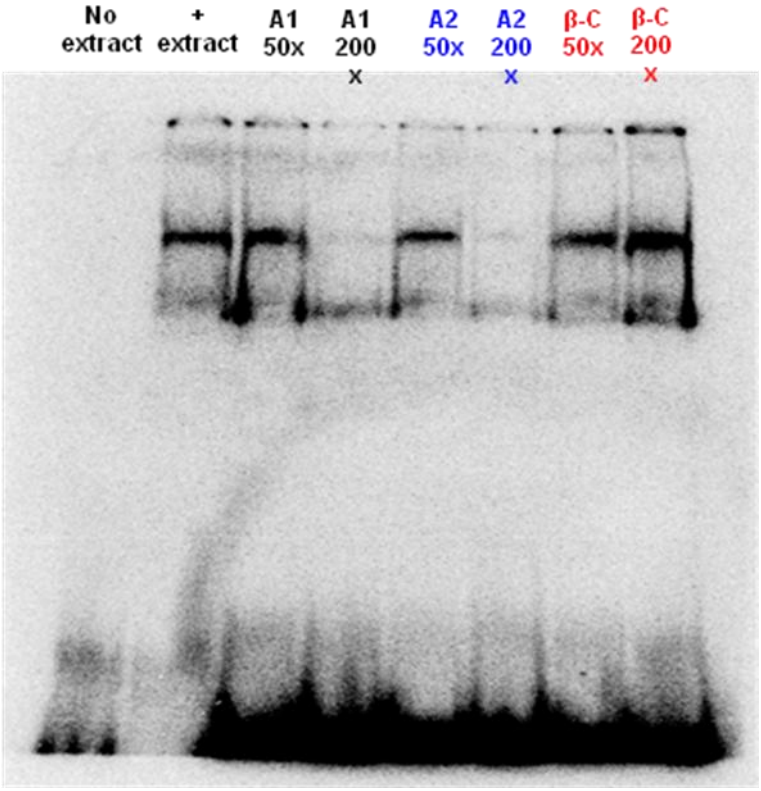


Figure 15: EMSA assay for Motif A1: the data show that in presence of molar excess of specific competitor (A1-A2: ≈ 200x) the retarded band disappears. This competition effect does not occours with the aspecific competitor (β-casein promoter oligo).

3.7 CAT assays

In order to investigate whether the selected motifs were able to act as transcriptional activators, we performed a serie of transient expression assays, with expression vectors containing the CAT reporter genes under the control of single or multiple copies of these motifs (figure 16).

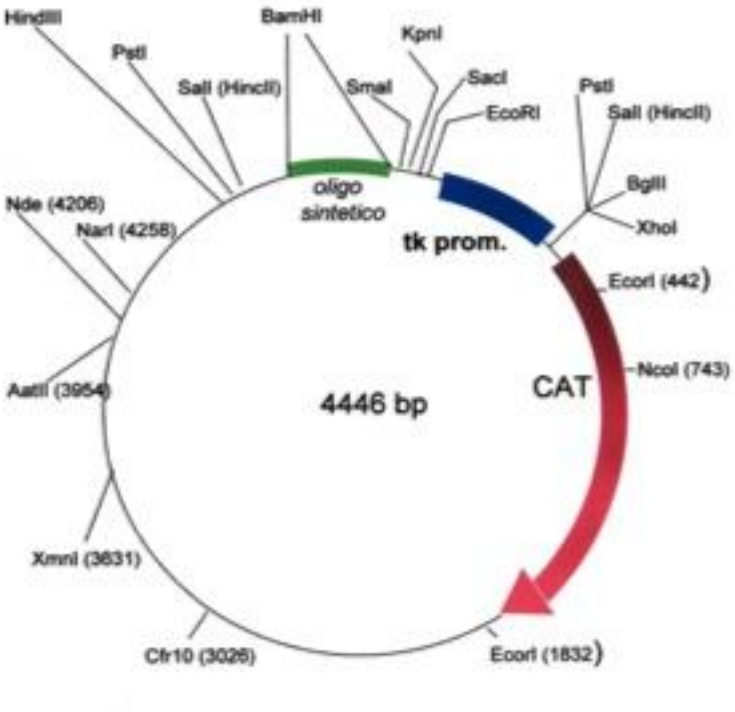


Figure 16: Map of expression vector pTKsh-CAT, 4446 bp, that used to clone the motif (in the BamHI site).

DNA sequencing to verify number and orientations of the inserts was described in the Mat. and Met. section. Table 3 shows the results of the cloning experiments, while in figure 17 is reported a typical transient expression.

Motif	Orientation and copy numbers
A1	Dir → → →
A1	Dir →
A1	Dir → → →
A2	Dir → →
A2	Rev ←
A2	Dir →
A2	Dir →
B1	Dir → →
B1	Dir → →
B1	Dir → →
B1	Dir → →
B2	Dir →
B2	Rev ← ← ← ←
B2	Rev ←
B2	Dir →
C1	Rev ←
C1	Dir-Rev → ←
C1	Rev ←
C2	Rev ←
C2	Dir →

Table 3: Copy numbers and orientation of oligonucleotides inserted in the pTKsh-CAT vectors.

The results show that, in HepG2 cells, the constructs containing the motif A1 are significantly more active than the reference vector (pTKsh-CAT), and that this activity increases depending on the copy numbers (motif A1 >> motif A1).

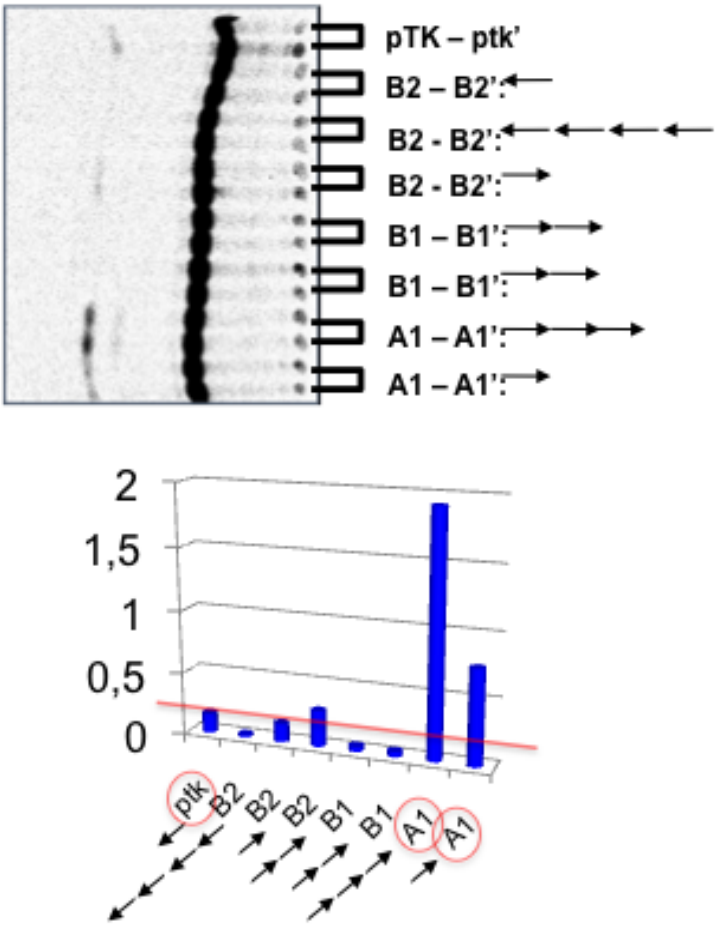


Figure 17: Transient expression assay of constructs containing one or more copies of motif A or B. The picture shows the autoradiography and the phosphor-imager quantization of a CAT assay. The CAT activity of the A1 sample is significantly higher than that of the pTK control plasmid (HepG2 cells).

3.8 Site-direct Mutagenesis of the A1 motif

The results of transient expression assays show that the A1 motif is transcriptionally active in HepG2 cells. To prove that this effect is sequence-specific and depends on the "CORE" motif of the A1 oligo we performed a site-directed mutagenesis analysis of motif A1 (Figure 18):

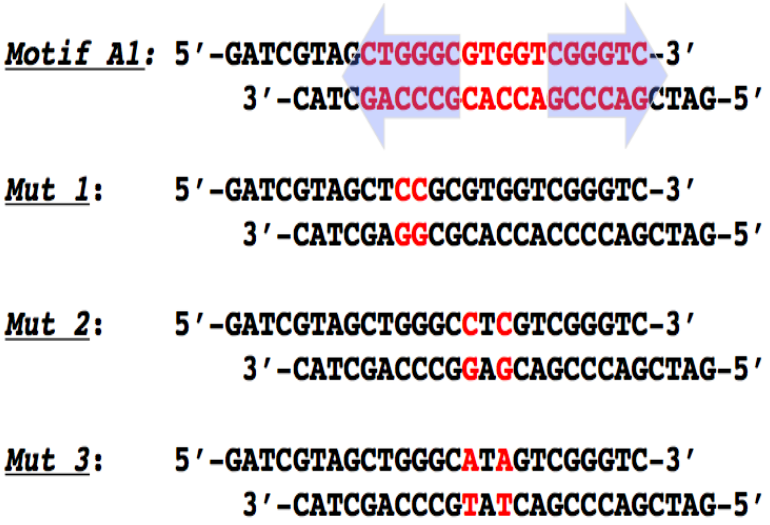


Figure 18: The three mutant oligos for mutagenesis analysis of motif A1. The divergent arrows indicate the two halves of the palindromic sequence found in the motif A1.

The transient expression assays, reported in figure 19 a-b, show clearly that all the mutations affect the transcriptional activity of the sequence, reducing it at levels comparable (if not lower) to those of the pTKsh_CAT control.

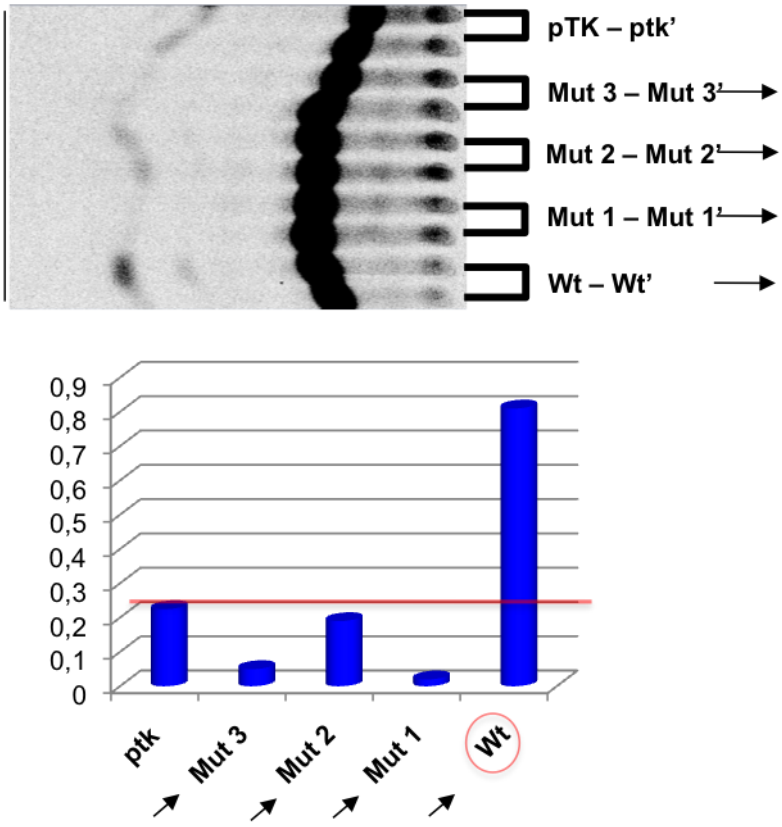


Figure 19a: Transient expression assay showing that all the mutations reduce drastically the transcriptional activity of motif A1(in HepG2 cells).

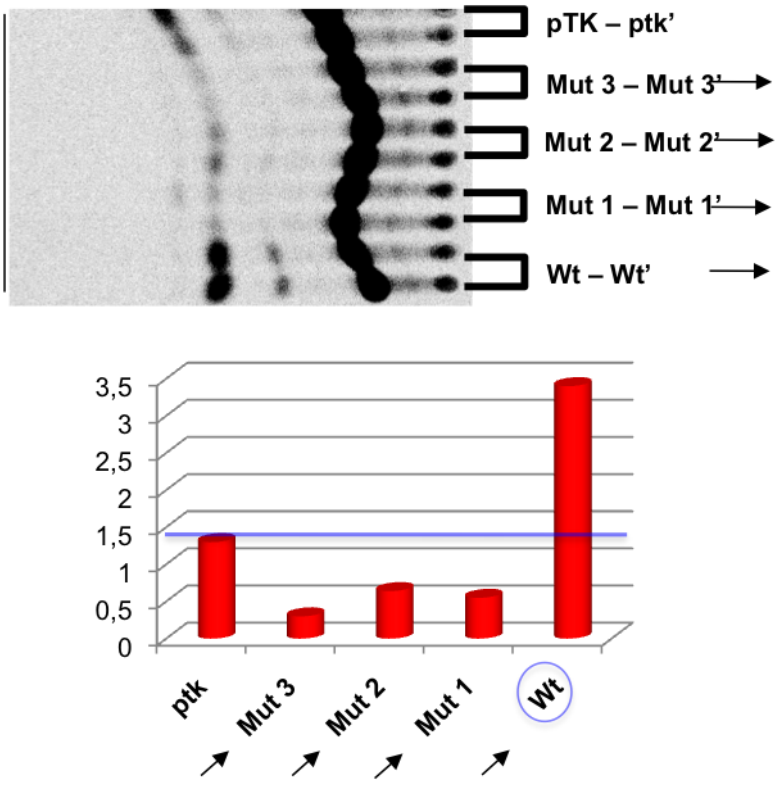


Figure 19b: Transient expression assay showing that all the mutations reduce drastically the transcriptional activity of motif A1(in HeLa cells).

4. Discussion

The ultimate goal of this study was to identify groups of genes whose expression levels in FCHL patients are significantly altered as compared to those of normolipidemic individuals and within the same FCHL patients, after treatment with statins. We have generated two lists of genes, corresponding to the 10vs5 experiment (FCHL patients vs. controls: 1747 genes) and 7vs7 experiment (FCHL patients before and after treatments with statins: 460 genes). The Intersection of the two experiments has generated a final list of 41 genes, mostly hypo-expressed in FCHL, that after treatment with statins regain the original levels of expression, switching, sometimes, from hypo- to hyper-expression. The expression profiles of 11 of these genes have been subjected to validation using the qRT-PCR tecnique. We have validated the expression patterns of 7 of the 11 genes analyzed, namely more than 60%, in agreement with what other authors find in similar experiments (31). We have confirmed the hypo-expression of the TCF1 gene (HNF1 homeobox A), which encodes for the transcription factor HNF1α, a major regulator of hepatic differentiation TCF1 alteration appears to be associated with a reduction of the APO-lipoprotein M (ApoM), and thus to a reduction of HDL (32), a typical symptom of FCHL. We also found the hypo-expression of the SIP gene (Siah-interacting protein), involved in protein degradation, which results in a slowdown of protein metabolism and in the alteration of the ATP synthesis. This may result in the accumulation of fatty acids, causing the FCHL symptoms. Finally, we found a reduced expression of the HLA-DQ1 gene (major histocompatibility

complex, class II, DQ alpha 1), responsible for the activation of lymphocytes B, which alters the immune response mediated by this factor, that is also partially reduced in FCHL patients.

These genes lists have been the starting material for a detailed analysis of the functional and regulative relationships between the genes whose expression was altered in FCHL patients or after statins treatment.

Gene Ontology (GO) analysis highlights coherent groups of functionally related genes that are significantly enriched. Gene Ontology analysis of the 10vs5 list shows a group of 14 hypo-expressed genes, that belong to the ATP synthase family involved in the protons transport in the mitochondria and a group of genes involved in protein degradation. Gene Ontology analysis of the 7vs7 list shows three enriched groups involved in cytoskeleton organization, morphogenesis and synthesis of heterocyclic compounds, respectively.

Network analysis showed clearly that the human diseases more related with our gene dataset are cardiac arteriopathy, consistent with the high risk of coronary heart disease in FCHL patients, closely followed by kidney and liver pathologies, typical complications of FCHL (Figure 10).

The promoter analysis of genes belonging to enriched GO categories and to the intersection list allowed us to identify putative regulatory sequences possibly involved in the pathology and/or in response to statins. The most interesting of these putative regulatory motifs were selected for further experiments *in vitro* (EMSA assay) and *in vivo* (CAT assay) to verify their ability to act as regulators of gene expression. All the tested motifs are positive to *in vitro* analysis, and one of them (motif A1) is able to

act as a regulator of gene expression (as also shown by site-directed mutagenesis experiments). A search for possible homologies between the consensus sequence of motif A1 and the binding sites of known transcription factors in the TRANSFAC and JASPAR databases showed a significant homology to the sequence of the binding site of the transcription factor Ikaros, which encodes a family of proteins belonging to the superfamily of zinc finger (33) and that play an key role in the cholesterol absorption. Motif A1 also shows significant homology to the binding site of factor Lun-1, that mediates proteins ubiquitination, regulates cell cycle, and inhibits cell proliferation (34-35).

Our future experiments will be directed in the following directions:

- To analyze the expression profiles of the validates genes in higher numbers of patients. This will serve to consolidate a "panel" of genes whose altered expression can be used as diagnostic tools for FCHL.
- To analyze the putative regulatory motifs identified *in silico* in their biological context (endogenous promoter). We are aware of the limitations of the transient expression experiments carried so far, and it is very likely that many of these motifs, while failing to activate a reporter gene, may still have an importante regulatory role in their original context. It is to be taken into account that our "in vivo" assay only reveals putative activatory elements, while it is not suitable, in its present form, to identify negative regulatory sequences.
- Finally, we plan to search for and to, identify the transcription factor(s) that interact with the regulatory motifs, identified *"in silico"*, and validated in expression assay.

Bibliography:

1. Kitano, H. (2002). "Systems biology: a brief overview." Science 295(5560): 1662-1664.
2. Lee, W. P. and W. S. Tzou (2009). "Computational methods for discovering gene networks from expression data." Brief Bioinform 10(4): 408-423.
3. Margolin, A. A., K. Wang, et al. (2006). "Reverse engineering cellular networks." Nat Protoc 1(2): 662-671.
4. Huang, S. S. and E. Fraenkel (2009). "Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks." Sci Signal 2(81): ra40.
5. Pilpel, Y., P. Sudarsanam, et al. (2001). "Identifying regulatory networks by combinatorial analysis of promoter elements." Nat Genet 29(2): 153-159.
6. Beaumont, J. L., L. A. Carlson, et al. (1970). "Classification of hyperlipidaemias and hyperlipoproteinaemias." Bull World Health Organ 43(6): 891-915.
7. Havel, R. J. (1969). "Pathogenesis, differentiation and management of hypertriglyceridemia." Adv Intern Med 15: 117-154.
8. De Michele, M., A. Iannuzzi, et al. (2007). "Impaired endothelium-dependent vascular reactivity in patients with familial combined hyperlipidaemia." Heart 93(1): 78-81.
9. Goldstein, J. L., H. G. Schrott, et al. (1973). "Hyperlipidemia in coronary heart disease. II. Genetic analysis of lipid levels in 176 families and delineation of a new inherited disorder, combined hyperlipidemia." J Clin Invest 52(7): 1544-1568.
10. Brunzell, J. D., J. J. Albers, et al. (1983). "Plasma lipoproteins in familial combined hyperlipidemia and monogenic familial hypertriglyceridemia." J Lipid Res 24(2): 147-155.
11. Castro Cabezas, M., T. W. de Bruin, et al. (1993). "Impaired fatty acid metabolism in familial combined hyperlipidemia. A mechanism associating hepatic apolipoprotein B overproduction and insulin resistance." J Clin Invest 92(1): 160-168.
12. Eurlings, P. M., C. J. van der Kallen, et al. (2001). "Genetic dissection of familial combined hyperlipidemia." Mol Genet Metab 74(1-2): 98-104.

13. Pauciullo, P., M. Gentile, et al. (2008). "Tumor necrosis factor-alpha is a marker of familial combined hyperlipidemia, independently of metabolic syndrome." Metabolism 57(4): 563-568.
14. Pisciotta, L., T. Fasano, et al. (2008). "A novel mutation of the apolipoprotein A-I gene in a family with familial combined hyperlipidemia." Atherosclerosis 198(1): 145-151.
15. Salazar, J., M. Guardiola, et al. (2007). "Association of a polymorphism in the promoter of the cellular retinoic acid-binding protein II gene (CRABP2) with increased circulating low-density lipoprotein cholesterol." Clin Chem Lab Med 45(5): 615-620.
16. Lee, J. C., D. Weissglas-Volkov, et al. (2007). "USF1 contributes to high serum lipid levels in Dutch FCHL families and U.S. whites with coronary artery disease." Arterioscler Thromb Vasc Biol 27(10): 2222-2227.
17. Pauciullo, P., M. Gentile, et al. (2009). "Small dense low-density lipoprotein in familial combined hyperlipidemia: Independent of metabolic syndrome and related to history of cardiovascular events." Atherosclerosis 203(1): 320-324.
18. Jasinska, M., J. Owczarek, et al. (2007). "Statins: a new insight into their mechanisms of action and consequent pleiotropic effects." Pharmacol Rep 59(5): 483-499.
19. Mootha, V. K., C. M. Lindgren, et al. (2003). "PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes." Nat Genet 34(3): 267-273.
20. Hwang, J. J., P. D. Allen, et al. (2002). "Microarray gene expression profiles in dilated and hypertrophic cardiomyopathic end-stage heart failure." Physiol Genomics 10(1): 31-44.
21. van de Vijver, M. J., Y. D. He, et al. (2002). "A gene-expression signature as a predictor of survival in breast cancer." N Engl J Med 347(25): 1999-2009.
22. Lawn, R. M., D. P. Wade, et al. (1999). "The Tangier disease gene product ABC1 controls the cellular apolipoprotein-mediated lipid removal pathway." J Clin Invest 104(8): R25-31.
23. Ewis, A. A., Z. Zhelev, et al. (2005). "A history of microarrays in biomedicine." Expert Rev Mol Diagn 5(3): 315-328.

24. Katagiri, F. and J. Glazebrook (2009). "Overview of mRNA expression profiling using DNA microarrays." Curr Protoc Mol Biol Chapter 22: Unit 22 24.
25. Vinciotti, V., R. Khanin, et al. (2005). "An experimental evaluation of a loop versus a reference design for two-channel microarrays." Bioinformatics 21(4): 492-501.
26. (2008). "The Gene Ontology project in 2008." Nucleic Acids Res 36(Database issue): D440-444.
27. Eden, E., D. Lipson, et al. (2007). "Discovering motifs in ranked lists of DNA sequences." PLoS Comput Biol 3(3): e39.
28. Lee, K. A., A. Bindereif, et al. (1988). "A small-scale procedure for preparation of nuclear extracts that support efficient transcription and pre-mRNA splicing." Gene Anal Tech 5(2): 22-31.
29. Bradford, M. M. (1976). "A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding." Anal Biochem 72: 248-254.
30. Bacchetti, S. and F. L. Graham (1977). "Transfer of the gene for thymidine kinase to thymidine kinase-deficient human cells by purified herpes simplex viral DNA." Proc Natl Acad Sci U S A 74(4): 1590-1594.
31. Morey, J. S., J. C. Ryan, et al. (2006). "Microarray validation: factors influencing correlation between oligonucleotide microarrays and real-time PCR." Biol Proced Online 8: 175-193.
32. Richter, S., D. Q. Shih, et al. (2003). "Regulation of apolipoprotein M gene expression by MODY3 gene hepatocyte nuclear factor-1alpha: haploinsufficiency is associated with reduced serum apolipoprotein M levels." Diabetes 52(12): 2989-2995.
33. Molnar, A. and K. Georgopoulos (1994). "The Ikaros gene encodes a family of functionally diverse zinc finger DNA-binding proteins." Mol Cell Biol 14(12): 8292-8303.
34. Rajendra, R., D. Malegaonkar, et al. (2004). "Topors functions as an E3 ubiquitin ligase with specific E2 enzymes and ubiquitinates p53." J Biol Chem 279(35): 36440-36444.
35. Saleem, A., J. Dutta, et al. (2004). "The topoisomerase I- and p53-binding protein topors is differentially expressed in normal and malignant human tissues and may function as a tumor suppressor." Oncogene 23(31): 5293-5300.