UNIVERSITA' DEGLI STUDI DI PARMA

Dottorato di ricerca in Fisiopatologia Sistemica

Ciclo XXI

Identification of differential gene expression profile in circulating lympho-monocytes of apparently healthy young adults in presence of cardiovascular risk factors through whole-genome transcriptomic profile analysis with oligo-RNA microarrays.

Coordinatore:
Chiar.mo Prof. .Ezio Musso

Tutor:
Chiar.mo Prof. .Ivana Zavaroni

Dottorando: .Diego Ardigò

Abstract

Background and Aims: Preliminary evidence indicates that the gene expression profile of circulating white blood cells is significantly affected by the exposure to cardiovascular (CV) risk factors. However the few available studies have several limitations, including small sample size and narrow range of genes tested. The aim of the present thesis is to provide a review of the current literature on this relationship and the presentation of the results of a clinical research program aimed at investigating the gene expression profile at whole-genome level in peripheral blood mononuclear cells (PBMC) of healthy volunteers in presence of CV risk factors and associated biomarkers.

Materials and Methods: In the clinical study, PBMC's expression profile was evaluated using Agilent whole genome oligonucleotide mRNA microarrays in 167 apparently healthy young-adults, all volunteers in a CV risk assessment study. Enrolled subjects were free of diabetes, CV and inflammatory diseases, and did not take pharmacological medications. Data analysis was performed using methods developed to address differential expression for quantitative phenotypes and enrichment of gene sets in ranked lists of genes.

Results. Differential expression profiles were identified in presence of cigarette smoking, high LDL-cholesterol concentrations, increased biomakers associated with CV risk (like IL-6 and TNF-alpha) and in presence of increased carotid intima-media thickness (IMT), a surrogate marker of atherosclerosis. In all these conditions, gene expression profile was characterized by a pro-inflammatory signature. More in detail, cigarette smoking was characterized by an over-expression of groups of genes and molecular pathways involved in the innate immunity, whereas smoking was associated with the over-expression of genes involved in the cell-mediated immunitary response and high IMT was characterized by an over-expression of genes of the mitochondrial respiratory chain. Interestingly, in subjects who smoked and had high LDL-c compared to low LDL-c non-smokers, the transcriptomic profile was characterized by the co-presence of the two bio-signatures and the enrichment of both innate and cell-mediated immunity. The gene expression profile associated with gender, age, and different proportions of leukocyte sub-populations are also presented as collateral results.

Conclusions: In conclusion, in apparently healthy young adults, the presence of several CV risk factors such as cigarette smoking, high LDL-cholesterol and thickened IMT are associated with the over-expression of genes related to the inflammatory response. However, different risk factors are characterized by different transcriptomic "signatures". Presented data and supportive literature suggest that gene expression profile in PBMCs reflects the biological processes involved in the chronic sub-clinical inflammation that is a key patho-physiological element of CV disease.

Table of content

Abbreviations, list of tables and figures

List of abbreviations

Abbreviation	Description
ATP	Adult Treatment Panel
BMI	Body Mass Index
BP	Blood Pressure
CVD	Cardiovascular disease
GO	Gene Ontology
hs-CRP	High-sensitivity C-Reactive Protein
GLM	General Linear Model
KEGG	Kyoto Encyclopedia for Genes and Genomes
NCEP	National Cholesterol Education Program
LDL	Low-Density Lipoprotein
HDL	High-Density Lipoprotein
NIH	National Institute of Health
PBMC	Peripheral Blood Mononuclear Cell
WBC	White Blood Cell

List of tables

List of figures

1 Background and rationale

1.1 Introduction

Atherosclerotic cardiovascular disease (CVD) is the primary cause of morbidity and mortality in the developed world [1,2] and in many countries of the so-called 2nd world, a group of nations in which economic status is improving above the threshold of poverty. CVD comprises several pathologies of the vessels wall, including acute events (such as myocardial infarction, other acute coronary syndromes, and stroke) and chronic disease (myocardial angina, peripheral artery disease / claudicatio, cerebro-vascular insufficiency). The pathophysiological basis of CVD is atherosclerosis, whose basic lesion is the "atherosclerotic plaque", also called "atheroma". The atherosclerotic plaque results from the focal deposit of lipids, especially LDL-cholesterol, in the sub-endothelial space, within the vessel wall.

Despite the chronic nature of the disease, CVD is often undiagnosed in its clinical stages, before the onset of symptoms or complications. The first clinical presentation of CVD in more than half of patients with coronary artery disease is either myocardial infarction or sudden death [3].

The early identification of subjects that are going to experience an acute CVD event has remained a challenge, despite decades of extensive research and that many risk factors for CVD have now been clearly established and are target of specific therapeutic interventions.

Indeed, approximately 40% of deaths from CAD occur in patients with total cholesterol levels lower than average for the general population. These data are confirmed by observations coming from the Framingham population (the first longitudinal observational study on CVD risk factors in the general population) in which 35% of newly diagnosed CVD events occurred in subjects with total serum cholesterol lower than 200 mg/dL [4], a level for long time considered as optimal.

This fact clearly indicates that identification of individuals at high risk for CVD development goes beyond identification of risk based solely on elevated cholesterol levels. Thus, over the past 10 years, the emphasis in CAD prevention has broadened from primary prevention based on single risk factors, such as hyperlipidemia, diabetes or hypertension, to the identification of higher-risk individuals on the basis of global risk assessment. Because many CVD events continue to occur in individuals considered at intermediate risk on the basis of current risk assessment, there is growing interest in the use of novel risk factors and other diagnostic techniques to detect levels of higher risk among this broad population.

However, traditional risk factors and new risk factors, such as hs-CRP or homocysteine, are able to predict the development of cardiovascular disease in no more than 50% of cases.

1.2 Critical appraisal of clinical approach to cardiovascular risk

1.2.1 Definition of CVD risk factor

A cardiovascular risk factor is a condition that is associated with an increased risk of developing cardiovascular disease. The association is statistical, and increases the *probability* of developing a certain type of cardiovascular event over a determined period of time. However the simple association between exposure and event is not enough to define a risk factor.

A "candidate" cardiovascular risk factor must meet several criteria:

1) The statistical association between the factor and cardiovascular disease must be strong. Generally, the presence of the factor should at least double the risk of disease.
2) In addition, the association should be consistent. The risk factor should produce disease regardless of gender, age, or race, and the association should be present in all or most of the studies in which it has been evaluated.
3) The association must make biological sense. A factor may appear to be related statistically to a disease, but unless such a relationship is biologically plausible, the statistical association may have little meaning.

A positive answer to all these points defines the presence of a marker of risk: a factor that is significantly associated to the events and defines the presence of a higher probability of disease independently whether or not the factor is causally related to the disease.

A fourth criterion is mandatory to define a risk factor:

4) A treatment that favorably changes the risk factor should reduce the incidence of disease.

The factor must make an independent contribution to increasing an individual's risk of developing disease and its correction must improve outcome incidence. Therefore, a risk factor is not merely an odds indicator, but usually is a target for pharmacological treatment in prevention strategies.

In fact, clinical risk assessment has two major goals: to identify persons who are at risk for accelerated atherogenesis and acute coronary syndrome development, and to define a set of goals for intervention. Identification of risk factors has both risk assessment and prevention purposes, since, at least so far, all the prevention strategies have been focusing on risk factors reduction.

Longitudinal studies on cardiovascular diseases consistently consider some risk factors as crucial for the identification of reliable and significant risk function for cardiovascular disease: metabolic factors (i.e., total and HDL cholesterol, fasting blood glucose), biological factors (i.e., blood pressure), or others linked to life style (i.e., smoking habit), evaluated together with age and gender.

Other factors, less frequently used in risk functions, have been considered important in the definition of cardiovascular risk: i.e., family history for premature cardiovascular events, body mass index, physical inactivity, genetic features.

Table 1: Traditional and novel CVD risk factors

Non-modifiable risk factors	Classical risk factors	Emerging risk factors
• Age • Male gender • Family history of premature CAD	• Hypertension • Smoking • Diabetes • LDL cholesterol • Overweight/obesity • Sedentary lifestyle • Atherogenic diet	• Triglycerides • HDL cholesterol • Small LDL particles • HDL sub-fractions • Lipoprotein (a) • Apolipoproteins • Homocysteine • ADMA • IFG/IGT • Metabolic syndrome

Subclinical markers of atherosclerosis	Inflammatory markers
• Ankle-brachial pressure index (ABI) • Tests for myocardial ischemia • Flow-mediated dilation (FMD) • Carotid intimal-media thickening (IMT) • Coronary calcium	• High-sensitivity C-reactive protein (hs-CRP) • Fibrinogen • Serum amiid A (SAA) • Adhesion molecules • Cytokines

1.2.2 Clinical effectiveness of assessment of and intervention on CVD risk factors

Current guidelines for the management of individual risk factors are provided by the third Adult Treatment Panel report (ATP III) of the National Cholesterol Education Program (NCEP) ([5]), the seventh report of the Joint National Committee (JNC VII) of the National High Blood Pressure Education Program (6), and the American Diabetes Association (ADA) (7,8). All of these guidelines are currently endorsed or supported by the American Hearth Association (AHA) and the America College of Cardiology (ACC), two of the most important scientific societies implementing guidelines on cardiovascular diseases.

All these guideline reports have a similar approach on cardiovascular risk that is a modified use of the global absolute risk concept: the principle is the adjustment of the intensity of risk factor management to the global risk of the patient. Risk stratification for a single patient starts from the single risk factor which the guideline is addressed to (i.e. plasma cholesterol concentration or blood pressure levels): the risk factor is measured and based on the registered value the patient is classified in a risk rank system. Lower risk usually requires only clinical follow-up and higher risk grades require immediate pharmacological intervention. Only in case of intermediate levels of the risk factor, where no clear strategy has been identify between non-pharmacological and pharmacological interventions, the use of a global risk approach is used to further stratify the patient into sub-groups with different prevention strategies.

In ATP II and JNC VI (the latest version of NCEP and JNC guidelines before the current ones), overall risk was estimated by a simple system of risk assessment that employed counting of categorical risk factors. Treatment goals for LDL cholesterol were set according to the number of risk factors. This system represented a blending of the concepts of relative and absolute risk in an effort to effectively institute prevention and limiting pharmacological intervention only to high risk individuals.

The use of categorical risk factors has undoubtedly the advantage of simplicity but may be lacking in some of the accuracy provided by graded risk factors. In fact, the summation of graded risk factors provides a great advantage over the addition of categorical risk factors and it was already being recommended in risk management guidelines developed by joint European societies in cardiovascular and related fields. Advocates of this approach contend that the increased accuracy provided by the grading of risk factors outweighs the increased complexity of the scoring procedures.

However, the choice of avoiding the use of a total risk estimate based on summation of risk factors that have been graded according to severity was determined mostly on a political basis: the major intervention in NCEP recommendations has been lifestyle changes; LDL-lowering drugs were reserved for persons with categorical elevations of LDL cholesterol who were projected to be at highest risk. After release of ATP II, several major clinical trials reported results showing the efficacy and safety of LDL-lowering drugs for primary prevention. These reports opened the door to wider use of LDL-lowering drugs. In addition, after ATP II there has been a growing view that a more quantitative assessment of short-term risk is required for the selection of persons who will benefit most from intensive risk-reduction intervention.

The ATP III panel was faced the choice between these two approaches and once again the choice was to cope with the need to reconcile the previous method of counting risk factors with the developing field of integrated, “global” risk assessment. There were advantages and disadvantages to each approach: risk factor counting would have provided continuity with previous ATP guidelines; allowed for a history of detected risk factors to be included in risk assessment; included family history of premature CHD; and provided a focus on the individual risk factors, each of which requires clinical intervention. However, risk factor counting alone also has disadvantages: it does not provide a quantitative estimate of absolute risk in the short term; it does not allow for variability in risk factor level or intensity (i.e., it uses only categorical risk factors); and it may underestimate the progressive impact of advancing age on absolute risk in older persons.

The final method chosen was an attempt to capitalize on the advantages of both approaches. Risk factor counting is retained for initial assessment, but Framingham risk scoring, updated for ATP III, is layered over risk factor counting to improve risk estimation for refining decisions about goals, intensity, and types of LDL-lowering therapy in persons with multiple risk factors.

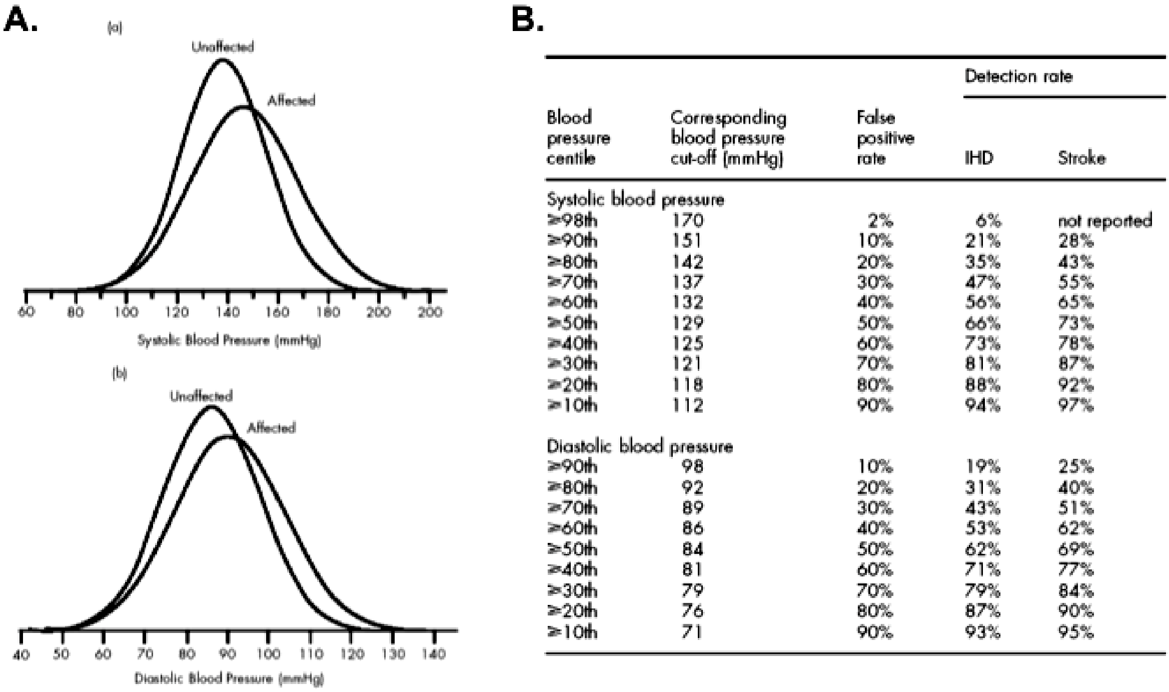
1.2.3 Limitations of the “risk-based” approach

The early identification of subjects that are going to experience an acute CVD event has remained a challenge, despite decades of extensive research and that many risk factors for CVD have now been clearly established. In addition, many of these risk factors can be adequately modulated.

By definition, a risk factor is a clinically relevant parameter that shows an independent association with the future development of the disease in the exposed subjects and whose pharmacological correction decreases the incidence of future events.

High LDL-cholesterol concentrations and hypertensions are two of the best established CVD risk factors. Nonetheless, in the Framingham Heart Study, as many as one third of all coronary heart disease events occurred in individuals with total cholesterol <200 mg/dL [4]. Considering that the average U.S. cholesterol level is approximately 210 to 220 mg/dL, almost half of all cardiac events occur among individuals with below-average lipid levels. In addition, the distribution curves of total serum cholesterol concentrations for subjects that will develop an acute CVD event and subjects free of events at the end of the 26 year follow-up of the Framingham cohort greatly overlap and there is no single value that can discriminate the two groups. The same phenomenon can be observed for hypertension: the curve of subjects with CVD and without CVD are almost completely overlapping and the sensitivity and specificity of using systolic or diastolic blood pressure values to identify subjects at higher risk is extremely poor [9], as shown in the next figure.

Figure 1: Predictive power of blood pressure for the presence of cardiovascular disease



Blood pressure centile	Corresponding blood pressure cut-off (mmHg)	False positive rate	Detection rate	
			IHD	Stroke
Systolic blood pressure				
≥98th	170	2%	6%	not reported
≥90th	151	10%	21%	28%
≥80th	142	20%	35%	43%
≥70th	137	30%	47%	55%
≥60th	132	40%	56%	65%
≥50th	129	50%	66%	73%
≥40th	125	60%	73%	78%
≥30th	121	70%	81%	87%
≥20th	118	80%	88%	92%
≥10th	112	90%	94%	97%
Diastolic blood pressure				
≥90th	98	10%	19%	25%
≥80th	92	20%	31%	40%
≥70th	89	30%	43%	51%
≥60th	86	40%	53%	62%
≥50th	84	50%	62%	69%
≥40th	81	60%	71%	77%
≥30th	79	70%	79%	84%
≥20th	76	80%	87%	90%
≥10th	71	90%	93%	95%

Footnote to figure. Panel A shows the prevalence distribution for systolic and diastolic blood pressure in "Affected" and "Unaffected patients". Panel B shows the detection rate and the corresponding risk of false positives stratified by blood pressure tentiles.

From a certain point of view, it is the same concept of risk factor and its connection to the progression of atherosclerosis that comes with several major limitations. First, the risk factors predicts the course of a process in a group of patients and not of an individual. Furthermore, a low level of a particular risk factor does not necessarily exclude the development of an event. As illustrated for the case of hypercholesterolemia, quite a number of patients with rather low cholesterol levels develop cardiovascular events. This is mainly explained by the fact that although patients with the highest cholesterol have the highest risk, patients with marginally elevated cholesterol comprise the largest proportion of the population and therefore of the patients suffering a cardiovascular event.

This limit appears to be partially overcome by the use of composite weighted risk scores such as Framingham's that start from a population-based assessment of event rate over time and model the relative weight of each risk factor combining them in one single equation model.

Despite the accuracy of these scores in predicting CVD events appears to be greater than what provided by the assessment of a single risk factor, age and gender still explain most of the variability of these scores and therefore of their accuracy.

So taken together, risk factor-based assessment of cardiovascular risk (and therefore risk factor-based treatment) is currently the best and most viable option for prevention at the level of the population and the single individual despite the fact that people are treated that do not need treatment, that people are not recognized as being at risk for cardiovascular disease, and that people are treated without adequately depressing the activity of the disease process.

In fact, the problem is how to define who we should treat and how to verify whether the treatment adequately dampens the systemic disturbance that drives atherosclerosis.

1.3 The role of biomarkers in preventive cardiovascular medicine

1.3.1 From risk factor to biomarker

Since it is clear that risk factors are not connected to disease states on a one-to-one basis, the search for systemic biomarkers of the atherosclerotic process has been progressively increasing over the last decade. The term "biomarker" has been defined in 2001 by the NIH Working Group as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" [10]. Applications of biomarkers as stated by the NIH Working Group include the use as a diagnostic tool for the identification of those patients with a disease or abnormal condition, the use as a tool for staging of disease or classification of the extent of disease, the use as an indicator of disease prognosis and the use for prediction and monitoring of clinical response to an intervention. As such, the concept of a biomarker greatly extends the use of risk factor in cardiovascular disease. With the use of a biomarker comes the application of surrogate end-points, that is a substitute for clinical endpoints [10]. From the description, it is clear that the application of a biomarker to diagnose and follow atherosclerotic disease is preferable to the use of a risk factor or the response of a risk factor to risk factor modulation.

In this view, the lack of representation of the stage and activity of the atherosclerotic disease before clinical manifestations is at least in part due to the lack of biomarkers that accurately identify active atherosclerotic disease before complications occur. Inflammation has been implicated in all stages of CVD and is considered to contribute to the pathophysiological basis of atherogenesis [11,12,13]. Inflammation may therefore serve as a potential source for markers of the disease process itself. In large epidemiological studies, various serum markers of systemic inflammation such as hs-CRP, fibrinogen, and interleukin-6 (IL-6) have been shown to predict cardiovascular events and to correlate with response to therapy [14,15].

More specifically, hs-CRP appeared to be a good marker to predict the development of peripheral artery disease in initially healthy physicians [16]. The reduced rate of progression of atherosclerosis after statin treatment has also been described to be significantly related to the reduction in hs-CRP [17]. Evaluations of hs-CRP have been performed in a variety of other conditions; this has led to an evaluation of markers of inflammation in cardiovascular disease by a Center of Disease Control and American Heart Association workgroup [18]. The report recommends the use of hs-CRP measurements as independent risk, however, identifies several issues that need further research. One of these is that further research is needed to assure optimal measurement of the presence, level, and type of inflammation in an individual patient. The workgroup also identified that the combination of inflammatory factors in the classification of cardiovascular disease risk was not well-explored.

Although potentially useful in risk stratification, the current systemic markers of inflammation lack sufficient disease specificity to be used satisfactory as a screening tool in the diagnosis of CAD [19].

The inaccuracy of current markers may reflect the fact that they are neither derived primarily from the vascular wall nor produced primarily by cells involved in the vascular inflammatory process. Furthermore, they may signal inflammation in a number of different organs and tissues, which may or may not have direct implications for the vasculature.

Recently, a number of studies have examined several other biomarkers on an individual basis as potential novel risk factors for CVD [20,21]. Although these studies demonstrate that some of these inflammatory markers (those known to be expressed in the diseased blood vessel) can predict the onset of clinically significant CVD, none of the markers provide clinically meaningful incremental value over traditional risk factors in predicting CVD complications.

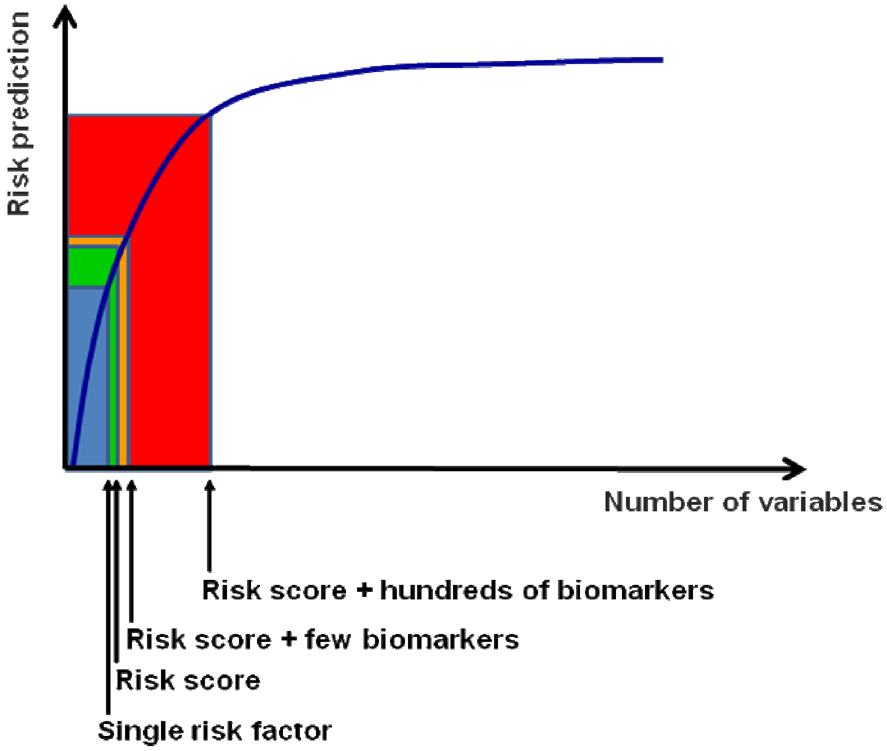
Taken together, these data point at an interesting weakness of the use of CRP (or any other single biomarker) as a proxy indicator for cardiovascular disease: shortcoming in reflecting a complex disease state in an individual. In other words, it is highly likely that, due to the heterogeneity of the disease phenotype in the population at risk, a single marker may not provide sufficient biological information for an accurate assessment of vascular damage in the coronary circulation.

This shortcoming has been approached by *multimarker scores* in which a composite of biomarkers are measured in parallel to assess risk. There are several considerations attached to the use of the multimarker approach, such as the potential interdependence of the markers. Another issue is that the course of event that starts with the presence of a risk factor to the occurrence of a cardiovascular risk in pathophysiological terms is a continuously changing complex condition. This is nicely illustrated by the large number of biomarkers that have been described for the various stages of atherosclerosis, in particular increased arterial vulnerability. It may well be overly optimistic to try to find one, or a small set, of biomarkers that will form an adequate reflection of the entire disease process. In other words, *the stage and activity of a very complex disease likely can not be captured in a single parameter or a small set of parameters and thus a complex state may need a complex representation.*

In addition, there are also statistical considerations that need to be put into the discussion. Even in presence of apparently independent risk factors or markers that serve as variables of a multi-variate model, part of the biological variability of the modeled phenomenon (in this case CVD occurrence) is shared among the independent variables. For this reason, each variable that enters the model after the first carries less and less information since part of its variability is already explained by the rest of the model. For this reason, one single parameter explain a certain quote of the outcome variability, whereas a few parameters account for significantly more variability explained but much less than the sum of they can explain singly. This is the case of the composite risk scores that show a better predictive power that a single risk factor, but this increase in accuracy is not tremendous. Adding one single or a few further parameters at this point does not change the overall accuracy of the model, almost independently from the predictive power of the parameter itself. This is what happens adding hs-CRP or a composite score of few biomarkers to the risk score.

The concept is clearly illustrated in the following figure.

Figure 2: Risk prediction as a function of the variables put into the model


Risk prediction
Number of variables
Risk score + hundreds of biomarkers
Risk score + few biomarkers
Risk score
Single risk factor

In this context, the only possibility to significantly improve risk prediction is to add hundreds or thousands of biomarkers to the model in order to overcome the issue of the co-information.

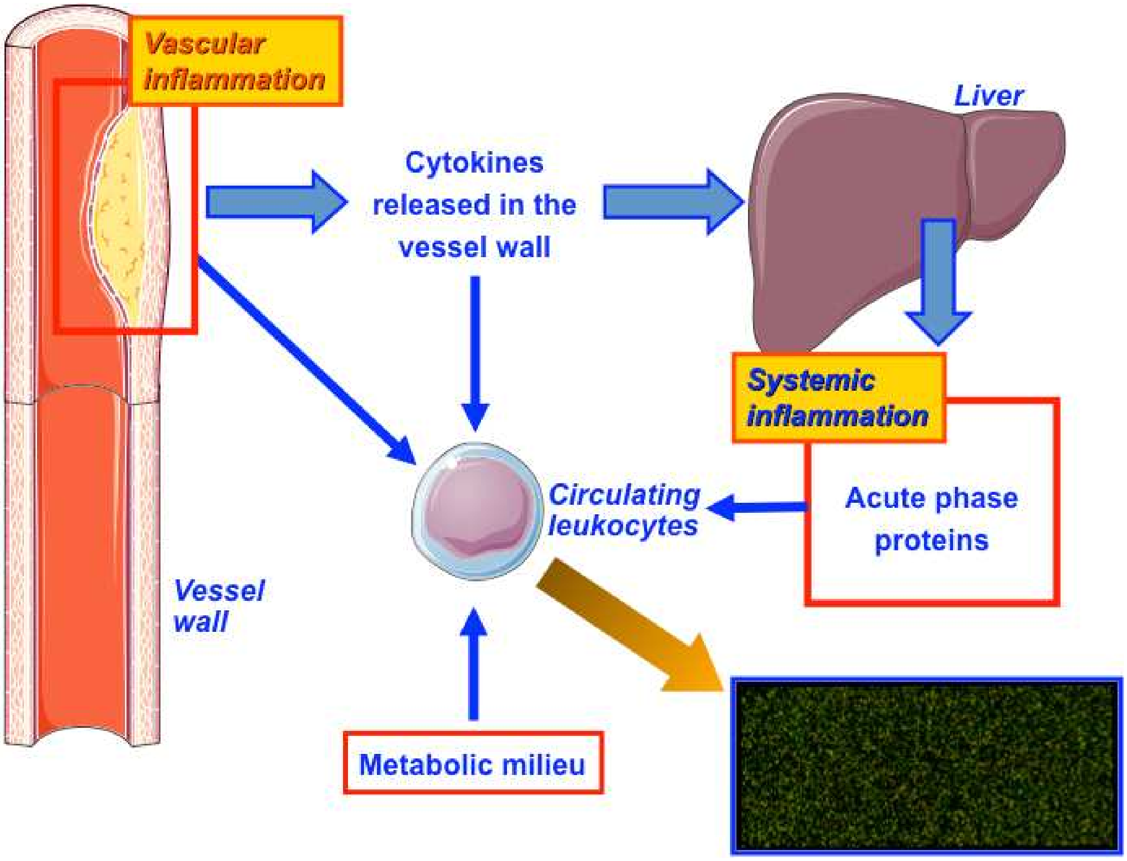
The identification of tens or hundreds of biomarkers need a discovery search among thousands or tens of thousands. Currently only high-throughput –omic technologies can provide such a tremendous amount of data of a single biological phenomenon in parallel.

1.3.2 Biomarker discovery from circulating cells

Much of the focus in finding new biomarkers for atherosclerosis has been on circulating proteins and mediators related to the inflammatory response. Very little attention has focused on the cellular component of the blood, which is remarkable, at least in view of the role of the monocyte population and other cell types in the initiation and progression of atherosclerosis. Interestingly, all cells in the blood stream are subjected to the same stresses as the endothelial cell layer, and will translate these stresses into responses such as inflammation and oxidative excess.

From this standpoint it is important to underline that white blood cells, especially monocytes and lymphocytes, are indeed exposed to several and variate stimuli known to be related to the atherosclerotic process (see figure below): these cells are known to re-circulate into the vessel wall and the atherosclerotic plaques, where they are exposed to the local pro-inflammatory mediators. In addition, circulating into the bloodstream, they are also exposed to systemic inflammatory mediators such as CRP and other acute phase proteins. Finally, these cells are also directly exposed to several cardiovascular risk factors that are present into the blood or condition the blood environment such as high cholesterol or blood glucose concentrations, or arterial hypertension.

Figure 3: Circulating WBCs are exposed to local and systemic inflammation



In addition, it must be noted that leukocytes are not only exposed to CVD risk factors and involved in the pathogenesis and progression of atherosclerosis, but also that these basically are inflammatory cells, programmed to respond in a pro- or anti-inflammatory way to any environmental stimulus to which they are sensitive.

We propose that the presence of one or more risk factors, together with normal physiological stimuli, induce changes in leukocyte gene transcriptions with patho-physiological responses. It should also be remarked that obtaining circulating cells is very easy compared to any biopsy technique currently available and, in fact, all widely applied diagnostic tests are derived from blood.

A large amount of data has been collected in circulating leukocytes to test in-vitro or in-vivo the relationship between single biomarkers and CVD traits. However, the analysis of single candidate biomarkers can be as much deceiving in circulating cells than it is for circulating biomarkers.

We therefore focused our attention on the massive screening of thousands of biomarkers from circulating white blood cells in order to accurately represent their physiology and pahto-physiology.

Among the different, available "omics" that can be used to represent cellular phenotype, transcriptomic and proteomic appear to be the most suitable to be applied in this context. Transcriptomic is the investigation at high-throughput level of the transcribed RNA with a quantitative estimation of the concentrations of all the known cellular transcripts. Proteomic is a molecule-discrimination technique that allows separating hundreds of molecules mixed in a biological matrix. Although the molecular complex that has the highest correlation with the cellular phenotype is the protein, the proteomic approach still has several technical limitations. First, the amount of proteic molecules (including peptides) is enormous and therefore the proteomic assay should be preceded by purification steps to exclude all the proteins outside of a narrow molecular weight range. Therefore it is usually not the entire proteome to be tested but just a part of it. In addition, once separated the molecules and identified spot that are differentially represented between different groups, there is no automatic information describing which proteins are represented by the identified spots. Therefore a subsequent effort should be put in characterizing the molecules and their function. Finally, we are very far form having a comprehensive representation of the proteome.

Instead, high-throughput gene expression analysis is currently essentially based upon the use of synthetic probes of already known genes or sequences. After the completion of the genome project, the almost complete totality of genes has been identified and a large quote of them has also been classified in functional ontology groups. Therefore, at the end of a gene expression analysis, it is clear which genes or transcripts are up- or down-regulated and what is the inter-relation among them. Since the totality of the known genes (and therefore the almost totality of the existing human genes) can be tested in parallel, gene expression provide a comprehensive holographic representation of the cellular phenotype. In presence of chronic exposure to environmental stimuli (like in the case of CVD risk factors) gene expression is significantly affected and reflects the long-term changes induced on the cellular metabolism.

The rational for using high-throughput techniques (and especially gene expression profiling) is further explained in details in the next chapter.

1.4 Rationale and use for high-throughput technologies for biomarker discovery

1.4.1 The concept of phenotype from a cellular standpoint

A phenotype can be generically defined as “The observable physical or biochemical characteristics of an organism, as determined by both genetic makeup and environmental influences” [American Heritage Dictionary; Houghton Mifflin Company, USA]. From biological or medical perspective, the phenotype of an individual organism is either its total physical appearance, constitution, and macroscopic clinical state or a specific manifestation of a trait (such as size, eye color), cellular response or behavior that varies between individuals or populations.

More strictly pertinent to cellular biology, the phenotype can be defined as the expression of a specific trait or molecular pattern. In a simplified view we can define as phenotype, all the changes in morphology and function that occurs in any single cell as consequence of its internal predetermined program (genotype) and the external environment.

Phenotype is therefore determined to some extent by genotype, or by the identity of the alleles that an individual carries at one or more positions on the chromosomes. In case of some purely genetic traits, such as classic monogenic diseases (hemophilia, sickle cells anemia, etc.), gene mutations are entirely responsible for the final phenotype, whereas in many other, phenotypes are determined by multiple genes and influenced by environmental factors. Thus, the identity of one or a few known alleles does not always enable prediction of the phenotype.

The interaction between genotype and phenotype has often been described using a simple equation:

genotype + environment → phenotype

However, the current scientific view is that neither genetics nor environment are solely responsible for producing individual variation, and that virtually all traits show a gene-environment interaction.

genotype + environment + gene-environment interaction → phenotype

Gene-environment interaction is a term used to describe any phenotypic effects that are due to interactions between the environment and genes. The difference between the two formulas is not of marginal importance, since genes and environment can interact in several either additive or non-linear ways, making any modeling assumption partially ineffective. Moreover, genes interact each other, as well as environmental situations do, making the system extremely complex. Finally random variations in gene expression occur in presence of a similar environment.

Therefore:

genotype + *environment* + *gene-gene interaction* +

gene-environment interaction + *random-variation* → *phenotype*

Other definitions of phenotype:

Definitions from http://ghr.nlm.nih.gov/ghr/glossary/phenotype

"The observable physical and/or biochemical characteristics of the expression of a gene; the clinical presentation of an individual with a particular genotype"

Definition from: GeneTestsThis from the University of Washington and Children's Health System, Seattle

"The observable traits or characteristics of an organism, for example hair color, weight, or the presence or absence of a disease. Phenotypic traits are not necessarily genetic."

Definition from: National Human Genome Research Institute at the National Institutes of Health

"Observable characteristics of an organism produced by the organism's genotype interacting with the environment"

In conclusion, phenotype can be defined as any detectable characteristic of an organism (i.e., structural, biochemical, physiological and behavioral) determined by an interaction between its genotype and environment. In other words "genotype" can be regarded as the genetic fingerprint of a particular cell, whereas "phenotype" is the outward manifestation of the genotype, conditioned by environmental stimula.

1.4.2 The use of gene expression microarrays for cellular phenotypization

To maintain homeostasis and function robustly in the face of a changing environment, a cell must sense changes in its surroundings and alter its physiology in a compensatory manner. To accomplish this task, a number of receptors on the cell surface and within the cell act as sensors for different environmental factors (eg, heat shock, hormonal stimulation, inflammation). When these sensors are triggered by environmental changes, signals propagate (typically through a network of protein kinases and release of second messengers) to the nucleus, initiating DNA transcription in RNA (the operative code for protein synthesis).

This conditional response of the cell leads at the end to changes in its proteins either in terms of function or synthesis/degradation. Therefore, changes in environment determine a differential post-translational modification of proteins, as well as a differential gene expression and protein synthesis. If investigating gene polymorphism and aplotypes means exploring the genotype, performing a gene or protein expression profiling means ultimately investigating the cellular phenotype.

Recent complementary advances in genomics knowledge and technology miniaturization are beginning to overcome the intricacies of exploring complex diseases and allow to perform a genome-wide exploration of the cellular phenotype.

Advances in robotics, miniaturization, and the development of robust reagents of molecular biology have allowed the development of microarrays, which are platforms that allow massively parallel screening of RNA or protein abundance with sufficient sensitivity and specificity for each single gene transcript. Microarray quantifies the transcriptome or the proteome of a sample of cells as compared to a reference sample, and therefore provides a functional measurement of gene expression.

Microarrays are small, solid supports onto which thousands of different elements (oligonucleotides, proteins, cells, cell extracts, and tissues) are immobilized, at fixed locations. The support surfaces are usually glass microscope slides. The power of microarray analysis lies in its miniature platform features, which facilitates relatively rapid global exploration of the systems studied

By now gene-expression microarrays can measure around 40.000 different RNA molecules at the same time covering the whole genome in a single assay. This technology relies on the 1-to-1 correspondence of the Watson-Crick base pair (adenosine always pairs with thymine and cytosine with guanine). Single-stranded DNA or oligo-fragments of RNA bases (oligomers) of a unique sequence within each gene are used as a probe for the presence of that gene within a sample. Direct hybridization of a marked sample (or concurrent, competitive hybridization of a marked sample and a reference) with the cDNA or oligomer probe determine a relative proportion between abundance of sample RNA bound to the surface of the array and the absolute concentration of the transcript in the sample. Quantitative fluorescent labeling of samples from control and experimental groups allows the determination of the relative abundance of any gene in a genome.

Gene expression profiling by microarrays is an especially powerful analytical tool for complex diseases because of the multi etiology characterized by a combination of genes involved in development of complex diseases. By profiling the global RNA content of any cell, it is possible to monitor in parallel all expressed genes to explore any phenotypical abnormalities in diseases.

Using this technique, researchers have been uncovering rapidly new genes and pathways involved in complex diseases and discovering novel treatment targets. Moreover, the microarray technique has been shown to be useful for patient management as a precise molecular device for diagnosis, prognosis and, personalized treatment, especially in oncology.

1.5 Methodological principles of gene expression profiling by microarrays

A microarray 'chip' consists of thousands of nucleic acid sequences considered to be highly specific for target transcripts and working as "probes", spotted in discrete quantities (each resulting spot is called "feature") on the surface of a microscope glass slide, thus forming an array of spots.

The basic principles of microarray experiments are the following:

1) The hybridization of complementary nucleic acids
2) The spatial discrimination as the way for multiplexing
3) A probe selection procedure representative of target transcripts
4) The selective fluorescent hybridization for relative abundance quantization

1.5.1 Hybridization principle

Hybridization reactions are based on the property of nucleic acid chains to reconstruct a double-strand structure when single-stranded: at the right temperature and pH, a strand of nucleic acid will base-pair (hybridize) with a strand made of the complementary (or anti-parallel) sequence. This is due to hydrogen bonds that form between specific pairs of nucleotides from the opposing strands (adenine with thymine and cytosine with guanine). The hydrogen bonds can be disrupted, causing the two strands to separate, by heating (a DNA double-strand separates at 95°C). If the double-strand is exposed to a temperature sufficient to break hydrogen bonds without disrupting the covalent phosphodiester bonds that link adjacent nucleotides together, the DNA sequence is not altered and two single-strand complimentary molecules are formed. This process is called "denaturation". If one were to denature a large number of specific DNA segments and then carefully lower the temperature, because of the tendency of bases to pair specifically (AT and GC), each strand would hybridize only with its complementary partner, perfectly reforming all of the original duplex structures. Thus nucleic acid hybridization is sequence-specific and exhibits near-perfect fidelity.

1.5.2 Spatial discrimination

The physical location of each spot on the array is pre-determined during the design phase and gives the correspondence between the measured local fluorescence and the annotation data of the probe-sequence spotted. Using a laser scan at the proper wavelength it is therefore possible to excite the fluorescent labeled sample hybridized to the surface of the array for each individual spot and photograph the fluorescent emission. A computer software, called feature extractor, identifies each spot because of the discrete intensity of fluorescence compared to local background, measures the distribution of the intensity based on a grey-scale (proportional to mRNA abundance in the sample), and match the abundance data with the array annotation information.

1.5.3 Probe selection

Two major techniques for gene-expression microarray probe selection and identification are now on the market:

- cDNA arrays
- oligo-arrays

Both of them can now be used for two-color gene-expression arrays.

The basis of probe selection is to identify relatively small sequences which could be considered as uniquely representative of the target transcript. This aim can be achieved using a single-strand cDNA of the target mRNA or by synthesizing directly on the array a short (usually 60 bases) sequence known to be specific of the target mRNA. The first approach requests the creation of cDNA clone libraries, whereas the second needs the identification of unambiguous sequences.

cDNA clones are prepared individually by amplifying genes of interest from an organism's genome. The genes are maintained in a bacterial vector system until the chip is ready to be fabricated. To assemble the microarray, the gene inserts are isolated from the bacterial clones and set at designated coordinates on the slide, which serves a solid support platform for the hybridization reaction. The many probes included in an array can be chosen individually from a database used to store sequenced genes and expressed sequence tags, or from commercially available cDNA clone sets. A robotic arrayer, designed to manipulate small volumes of solution along three axes with extreme speed and precision, is used to assemble each chip. The arrayer places a nanoliter quantity of each probe onto the predetermined grid coordinate on the slide (spotting). The probes are immobilized onto the slide after each sample has been spotted onto its coordinate.

Oligo-arrays are prepared by in-situ synthesis, via photolithography, of a short (usually 50-70 bases) sequence. Unlike cDNA clones, oligonucleotide probes can be synthesized from sequence data alone, so bacterial clone sets do not need to be maintained. Therefore, the major advantages are: the freedom of changing sequence as wanted to improve accuracy of hybridization, and the possibility of including into the array also genes for what the cDNA is not available in the reference library. Sequences for oligos are selected on the basis of public repositories of gene sequences.

Oligonucleotide arrays are the ‘highest density’ platform of microarrays; an incredibly large number of genes can be represented on a single chip using oligos instead of the full length cDNAs. The high density of oligo-based arrays allows for the representations of the whole genome and the inclusion of redundant probes on the slide, as well as a large number of positive and negative control features.

1.5.4 Concurrent selective fluorescent hybridization

Microarray measures gene expression as compared to a control sample of reference RNA.

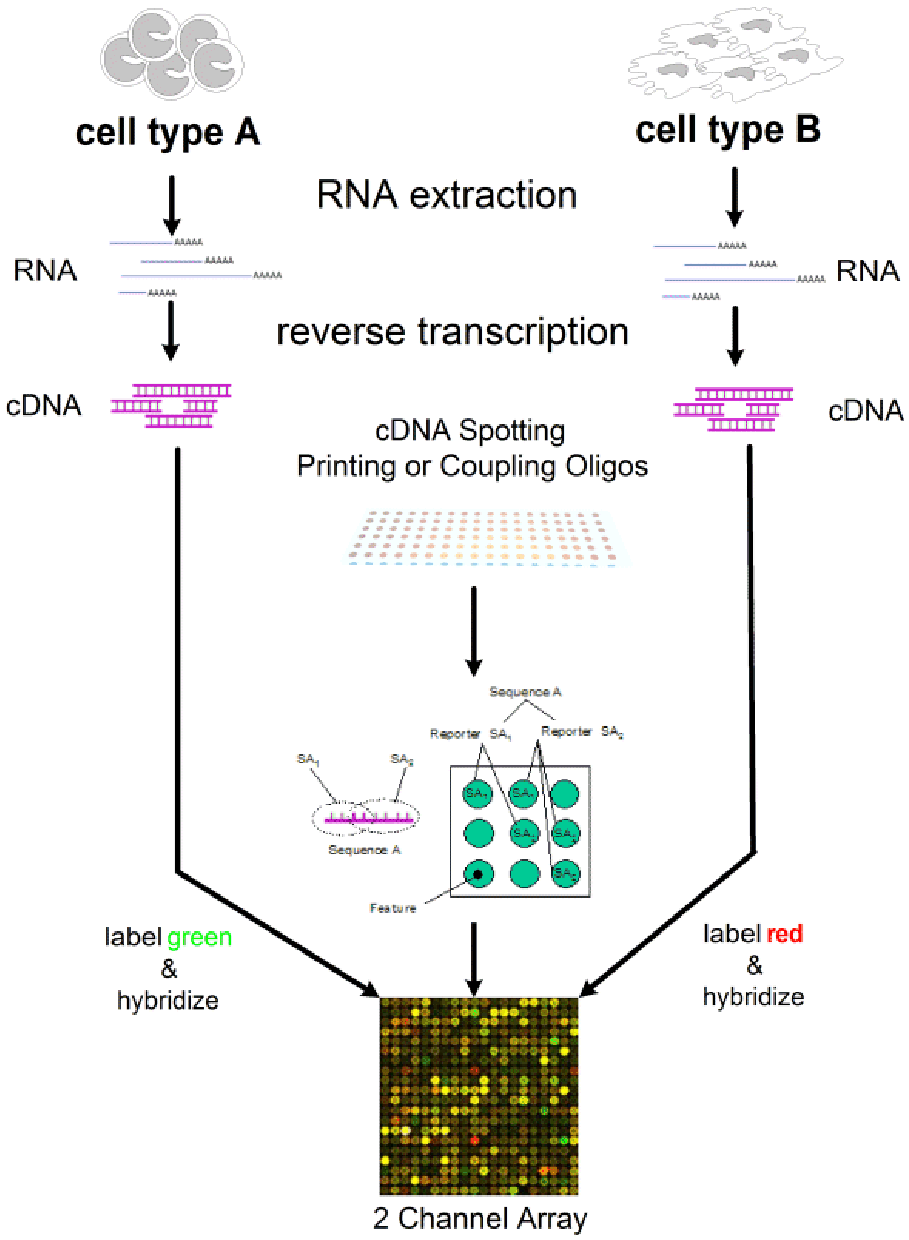
Total RNA is extracted from a sample of cells of interest, and from the appropriate control sample. Then sample and reference RNA can be processed directly or after one or more cycles of linear amplification. A cDNA library is constructed from each sample using the RT-PCR reaction with random primers to obtain a complete cDNA representation of the original RNA. During reverse transcription each cDNA library is separately labelled with a fluorescent molecule. Usually, reference cDNA is tagged with the green fluorescent molecule cyanine 3 (Cy3), while sample library is labeled fluor-red with cyanine 5 (Cy5). An equal amount of each solution of targets (both experimental and control cDNA) are pooled together, denatured, and hybridized to a single chip. Because both libraries are incubated together with a single chip, identical targets which are present in both libraries must compete for binding sites to their corresponding probe. This is known as competitive hybridization.

The sample that expresses greater amounts of a given gene (and thus contains a greater number of fluorescent targets in the reaction mixture) will competitively inhibit the hybridization of the other target. Thus at each coordinate on the slide, red or green fluorescence intensity is directly proportional to the amount of target present in one sample compared to the other. The slide, after being washed clean of unhybridized reaction substrates, is analyzed with a confocal laser scanner.

The scanner resolves fluorescent feedback from each coordinate, segregating it into red and green channels. The expression of individual genes is then quantified as a relative value, measured as the ratio of red-to-green ratio.

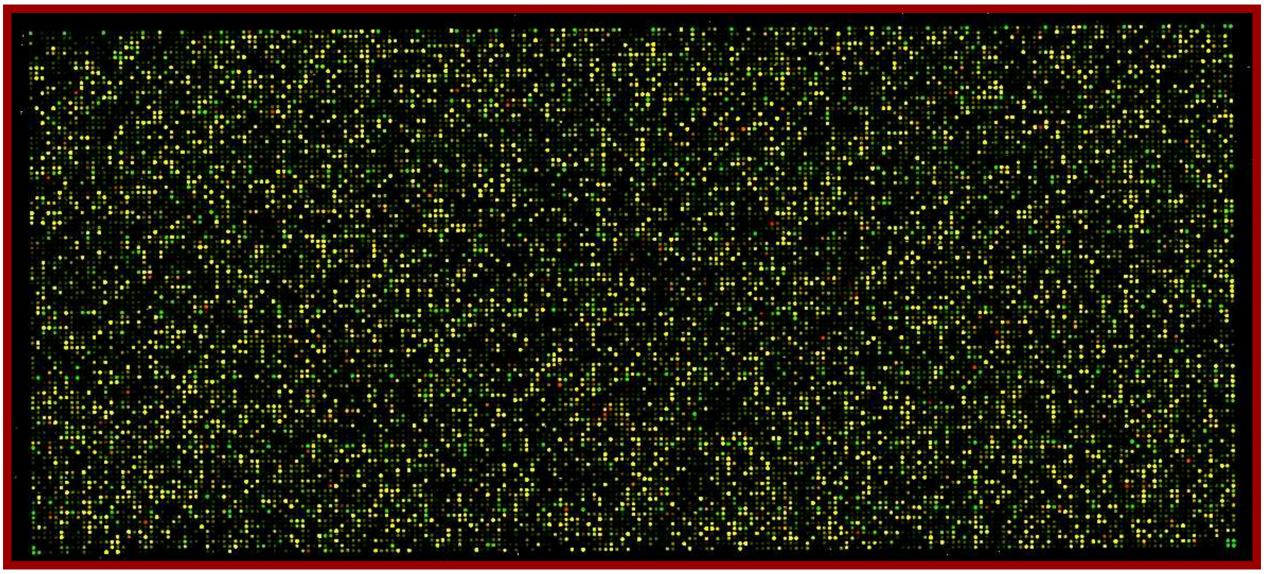
Following picture depicts the main steps of a typical dual-color microarray experiment.

Figure 4: Typical experiment workflow for dual-color gene-expression microarray



After hybridization, microarray slides are washed, dried, and then scanned. Arrays are scanned with a confocal laser scanner that generates beams at the two wavelengths specific for Cy5 and Cy3. The fluorescent emission of the slide is converted into two matched digital pictures (one for each fluorescent dye) whose degree of grey for each pixel on a 256 levels grey-scale is proportional to the fluorescent emission from the area of the array covered by that pixel. Thus at each coordinate, red or green fluorescence intensity is directly proportional to the amount of target present in one sample compared to the other. After scanning, a typical microarray slide picture is typically represented with false colors, with a three-color scale from bright green (the gene represented in the spot is highly over-expressed in the reference compared to sample) to yellow (the gene is approximately equally expressed in sample and reference) to bright red (the gene is highly over-expressed in the sample than in the reference). [See following figure]

Figure 5: Dual-color microarray slide represented by false colors



The average value of intensity within the same feature, compared to local background signal, is used as measure of relative RNA concentration for the transcript represented in that feature. The expression level for each gene in the array is described in comparison to the reference sample using the log base 2 scale, or log2 ratio.

For example, a gene with a log2 ratio equal to 1 indicates a 2-fold increase in expression (i.e. the mRNA transcribed from the gene of interest is present in the experimental sample at twice the levels as compared to the control). A log2 ratio equal to -1 indicates a 2-fold decrease in expression, and log2 ratio equal to zero corresponds to a perfectly yellow point on the slide and thus identical levels of expression between samples.

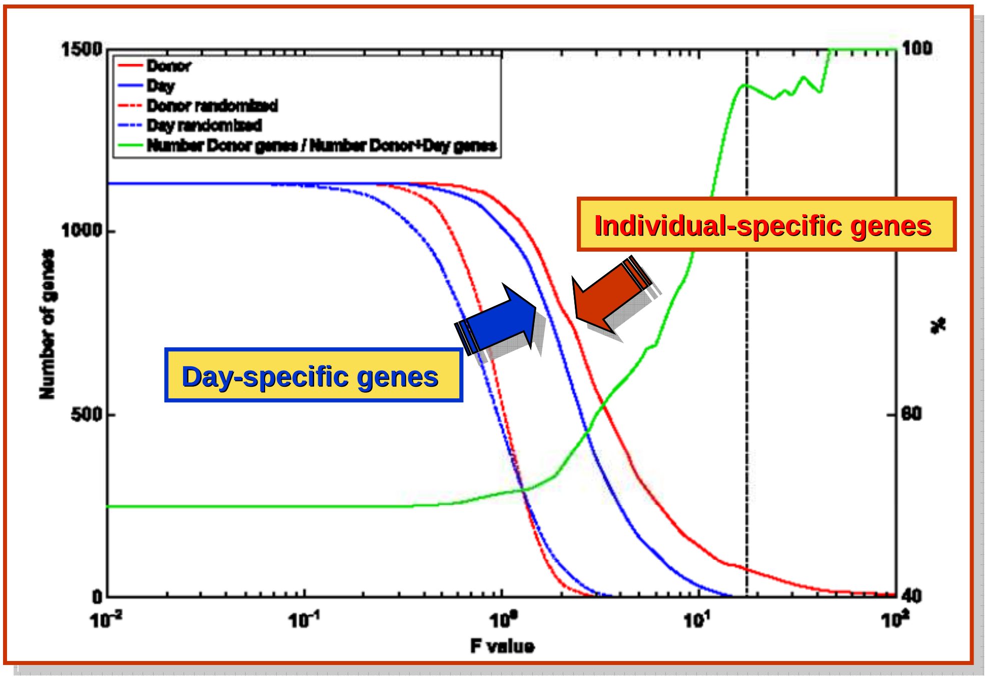
1.6 Methodological issues related to the use of gene expression profile of circulating cells in the study of atherosclerosis (and other chronic complex diseases)

When narrowing down which cell type is optimal to be used as biosensor for cardiovascular disease, several standpoints can be taken. One is that one uses the whole leukocyte population. Cardiovascular risk factors could well shift the balance between the various populations of cells, however, one could support the view that it is not so much the cell type per se, but the expression level of all cells in the blood resembles the influence of the atherosclerotic process. In other words, if, for example, the uremic state causes oxidative stress which can induce an inflammatory response in a variety of cell types, it may not be so relevant to know which cell type exactly contributes to this inflammatory response, merely it matters that the expression of particular cytokines is modulated. As such, the entire cell population can serve as biosensor. A trade-off may be that modulation of gene expression in a small, relevant subset of the cells may be 'diluted' by a large number of other populations of which the gene expression remained unaltered. One could also take the point of view that, to obtain more precise insight in the process of atherosclerosis, specific information is needed from subtypes of cells which form the effector cells for the process. Monocytes that are heavily involved in initiation of arteriole wall damage, which first becomes visible as a fatty streak, may be the cell type of choice to study very early atherosclerosis. Lymphocytes, and perhaps also granulocytes, may be more relevant cells to study once atherosclerotic lesions are already formed. Furthermore, one could envision that selecting a circulating cell subpopulation which displays phenotypical changes heralding invasion of the arterial wall, such as increased surface expression of adhesion molecules could yet further improve the specificity of the information supplied by gene expression analysis. However, based on a trade off between relevance for CVD pathogenesis and relative simplicity of specimen processing, peripheral white blood cells or mononuclear cells (lympho-monocytes or PBMCs) appear to be the most suitable targets.

1.6.1 Reproducibility of WBC/ PBMC gene expression analysis by microarrays

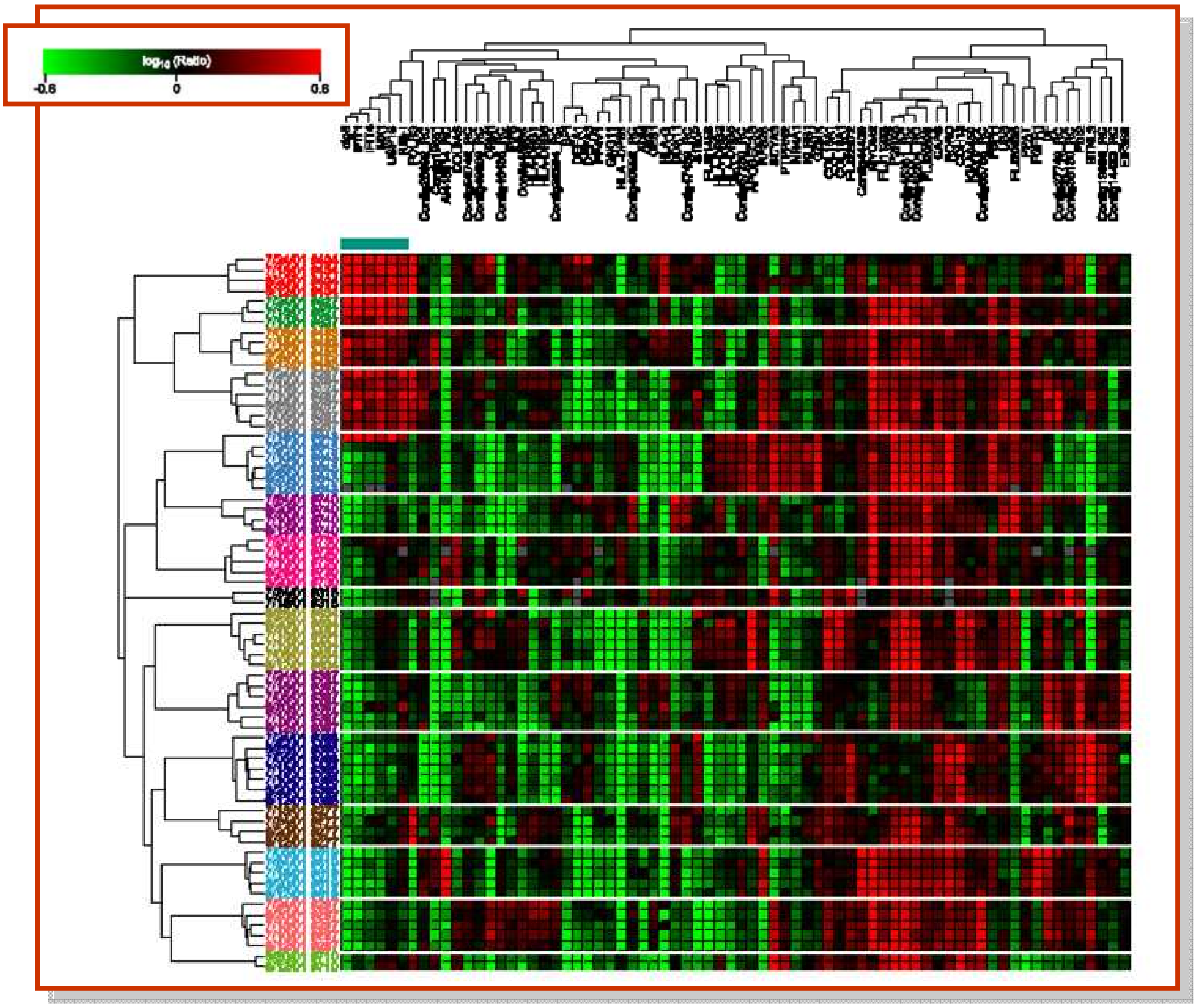
An obvious question that has developed is whether gene expression in circulating cells, specifically leukocytes and peripheral blood mononuclear cells, is stable. Three studies have been performed, with the same overall conclusion. One study assessed temporal changes in individual gene expression patterns in whole blood using microarrays and in peripheral blood mononuclear cells, and concluded that inter-individual gene expression modulation is much more prominent than intra-individual gene expression [22]. More in detail, the authors used a two-way ANOVA approache to analyze inter-individual and intra-individual gene expression differences. The following graph shows the distribution curve of the number of differentially-expressed genes for each given ANOVA F-value.

Figure 6: Individual-specific variation of gene expression in peripheral blood leukocytes



As it appears from this representation, the use of a sufficiently restrictive definition of "differential expression" (in this case an F-value approximately above 20) identifies a group of genes that are different between groups without false positive genes related to intra-individual daily variations in gene expression. In addition, the following picture shows the results of a 2D-Hierarchical Clustering analysis operated on these data, performed to assess whether the main determinant in variability of gene expression data was the intra-individual or inter-individual expression patterns.

Figure 7: 2D-Hierarchical Clustering of intra-individual VS inter-individual gen e expression profiles



Footnote to figure. Two-dimension hierarchical clustering of gene expression data in Radich et al. experiment. Each row represents a sample and each column a gene. Different specimens from the same subject are collected in different days and are represented in different rows using the same color. Each color corresponds uniquely to a subject.

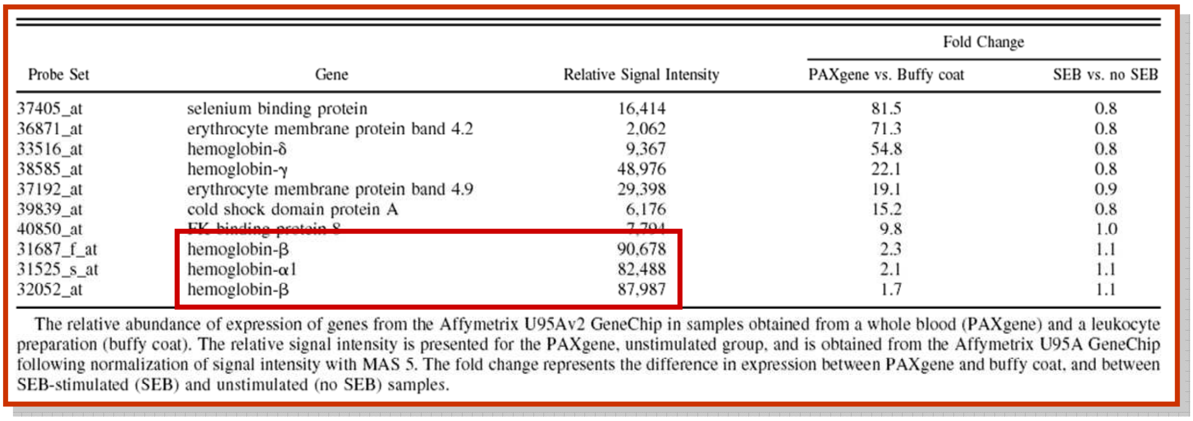
As it clearly appears from the picture, different samples form the same subject perfectly cluster together indicating that it is the inter-individual variability that drives the composition of the clusters. This is of high relevance as the clustering technique is by definition unsupervised and therefore no comparison or pre-analytical assumption can be made to force the identification of the clusters. It is the distance matrix that discriminates the similarities and differences among samples that start all from a single large cluster as H0 hypothesis.

Another study in peripheral blood leukocytes confirmed these findings and also analyzed whether immediate processing of the samples was necessary; indeed the individual profiles slowly disappeared through time, indicating that immediate processing of the samples is necessary to obtain these stable results [23]. Finally, a study assessed gene expression variation in peripheral blood mononuclear cells solely and also concluded that intra-individual variation is not very pronounced [24]. Shared between two of the studies is that several of the genes that appear to have the greatest intrinsic individual variation. PBMCs are also known to have highly polymorphic sequences; this supports that variation and expression of these genes reflects an underlying allelic variation in their regulatory sequences [23]. Taking a look at these studies it is very encouraging for the use of leukocytes as biosensors. Nevertheless, the methodological issue will remain in that in a population where gene expression patterns are derived from people with disease states, a suitable reference pool has to be available.

1.6.2 Collection and processing of circulating cells

It is self evident that circulating cells are easy to access, and therefore isolation and manipulation of specimens is apparently free of problems. However, peripheral blood contains several populations of mature white cells with an uncountable number of sub-populations each one requiring a different isolation protocol. In addition, there is clear evidence that also differentiated red blood cells and platelets hold significant amount of total and messenger RNA [25]. Erythrocyte RNA is extremely abundant in whole blood samples, and globin A1 is the most expressed gene, as shown in the next figure [26].

Figure 8: Top expressed genes in whole blood samples



			Fold Change	
Probe Set	Gene	Relative Signal Intensity	PAXgene vs. Buffy coat	SEB vs. no SEB
37405_at	selenium binding protein	16,414	81.5	0.8
36871_at	erythrocyte membrane protein band 4.2	2,062	71.3	0.8
33516_at	hemoglobin-δ	9,367	54.8	0.8
38585_at	hemoglobin-γ	48,976	22.1	0.8
37192_at	erythrocyte membrane protein band 4.9	29,398	19.1	0.9
39839_at	cold shock domain protein A	6,176	15.2	0.8
40850_at	FK binding protein 8	7,794	9.8	1.0
31687_f_at	hemoglobin-β	90,678	2.3	1.1
31525_s_at	hemoglobin-α1	82,488	2.1	1.1
32052_at	hemoglobin-β	87,987	1.7	1.1

The relative abundance of expression of genes from the Affymetrix U95Av2 GeneChip in samples obtained from a whole blood (PAXgene) and a leukocyte preparation (buffy coat). The relative signal intensity is presented for the PAXgene, unstimulated group, and is obtained from the Affymetrix U95A GeneChip following normalization of signal intensity with MAS 5. The fold change represents the difference in expression between PAXgene and buffy coat, and between SEB-stimulated (SEB) and unstimulated (no SEB) samples.

In the context of this complexity, two opposite tendencies in RNA isolation for microarray studies are emerging: on one side, there is the attempt to use whole blood specimens, on the other the effort to identify the population or sub-population most representative of the disease process.

Protocols for RNA extraction from whole blood aim at reducing sample manipulation. This approach has the double advantage of decreasing the delay between specimen extraction and RNA stabilization/

isolation, and of decreasing the complexity of the pre-analytical procedures to a level suitable for hospital laboratories and clinical practice. This latter is of major importance in multi-center clinical trials and large-scale population studies, where the effort required to isolate subpopulations from peripheral blood and to stabilize RNA right after blood withdrawal can be overwhelming and often unrealistic, but there is clear evidence showing that adequate sample handling is of paramount importance to prevent ex vivo transcriptional changes [27,28].

Whole blood RNA collection systems are commercially available as specialized Vacutainer tubes containing a proprietary reagent that stabilizes intracellular RNA for days at room temperature and weeks at 4° C, reducing the need of immediate processing and/or freezing. The PAXgene system (PreAnalytiX, Hombrechtikon, Switzerland) has been shown to stabilize RNA [29], prevent its degradation [30] and provide high-quality sample for RT-PCR [31] applications. However, when compared to white blood cells, total RNA from stabilized whole blood specimens shows much lower signal-to-noise ratio (as demonstrated by reduced overall present call rates in single dye arrays [27]) and inferior data quality (as assessed by increased intragroup variance [32]). The high abundance of globin mRNA has been suggested to account for these limitations [26] and RNase H digestion protocols or erythrocyte-depletion systems have been showed to greatly improve reproducibility, variance, and signal-to-noise ratios in whole blood specimens [33,34]. However these protocols are complex and require an extra work, currently not easy to standardize.

Based on this information, the isolation of total white blood cells provides a more robust sample with better array performance. In fact, when systematically compared [26], leukocytes and whole blood samples show a similar reproducibility (Pearson's correlation coefficient for replicates higher than 0.985 for both), but a quite different ability to detect changes in gene expression. Whole blood RNA discriminates between normal and pathological samples (trauma patients) with less accuracy, and provides a lower differential signal after stimulation with Staphylococcus enterotoxin B compared to buffy-coat RNA. As mentioned above, these findings can be partially explained by the extreme abundance of erythrocyte/ reticulocyte-specific transcripts (such as globin genes) that do not change in response to the external environment and impair the performance of discrimination and clustering analyses.

Looking at published data in general, it is evident that both whole blood and leukocyte isolation protocols generate high-quality RNA, but their impact on transcriptome results is significantly different, raising major concerns for the possible use tout-court of blood-based systems in clinical research. However, despite a quite clear superiority of white blood cell RNA as biological matrix for high-throughput microarray studies, it should be kept in mind that the technical procedures to obtain an uncontaminated buffy-coat sample (including centrifugation, erythrocytes and platelets lyses, washing, and re-suspension) are not of routinely use in clinical research.

In addition, white blood cells are a complex population comprising a variety of cell types (T-cells, B-cells, monocytes, NK cells, and granulocytes), each of which can be further subdivided (Th1, Th2, neutrophils, eosinophils, etc.). It is without dispute that the relative proportion of these subpopulations displays variations upon physiological stimuli, upon disease processes, and in response to treatments (i.e. corticosteroids). Their involvement in the pathophysiology of the disease is also not homogeneous: granulocytes, for instance, are the most abundant white blood cells subpopulation (about 50 to 70% in healthy adults), but their role in vessel atherosclerosis is minimal compared to monocytes or lymphocytes. Therefore, granulocytes can be either non-representative of the disease process or influenced by other sources of inflammation. For this reason, granulocytes are often removed from buffy-coat and mononuclear cells selected for RNA extraction. Peripheral blood mononuclear cells (PBMC) are an enriched preparation of monocytes and lymphocytes. Their relationship with CVD pathophysiology is more stringent than total WBC and therefore they appear to be a highly representative sample, but the isolation procedure is complex and delicate because of the large sample processing request after blood drawn.

In addition neither PBMCs are a single population and their composition includes several highly polarized cell types whose proportion is also highly variable: monocytes range from 2 to 10% of PBMCs, T- and B-lymphocytes vary from 61-85% and 7-23% respectively, and also the ratio between CD4+ and CD8+ T-cells ranges from 1.0 to 2.0 [35]. This difference and variability in the relative proportions of WBC subpopulations is not trivial and could potentially affect the composite gene expression profiles of whole blood or unfractionated PBMCs. Palmer and colleagues [36] performed a microarray analysis of purified subpopulations of peripheral blood cells, and were able to identify highly specific gene expression signatures for B-cells (427 genes), T-cells (222 genes), CD8+ T-cells (23 genes), granulocytes (411 genes), and lymphocytes (67 genes), indicating a high level of heterogeneity among PBMC sub-populations. Moreover, these differences are more prominent than the differences in gene expression profile between individuals in unsupervised clustering analyses, showing that subpopulation-specific genes have a greater impact on transcriptome than person-specific genes. In other words, changes in the proportion of sub-populations could affect statistical analysis more than changes in the composition of the study population. In addition major differences in gene expression could also be present in polarized subgroups of cells within the same sub-population. For instance, T cells subpopulations share a large number of commonly expressed genes, but a comparison within CD4+ cells shows that more than 100 genes are differentially expressed between Th1 and Th2, and this list includes not only genes associated with cell polarization, but also cytokines, transcription factors, molecules involved in cell migration, and genes with unknown function, not previously associated with differences in T cells subpopulations in pre-array studies [37].

Finally, the role of a cell population can vary overtime in the pathophysiology of a disease: if it is clear that lymphocytes and monocytes are highly involved in atherosclerosis development, there is also evidence that during the first 24 hours after an acute cerebrovascular event gene expression in polymorphonuclear leukocytes changes to a higher extent than in monocytes, being perhaps more representative of the acute response to injury [38].

1.6.3 Reference selection for two-color arrays

Together with the choice of an adequate biological sample, representative of CVD risk, the use of two-color arrays (microarrays that use two samples, a test and a reference, labeled with two separate dyes) raises the fundamental question of the selection of an adequate reference sample. This choice is not trivial and influences the possibility to make comparisons across laboratories, platforms, and populations. The definition of a common reference for two-color array studies would favor the integration of different datasets and would provide a set of basic values to compare raw data with one-color arrays. However, so far no clear standard for referencing has emerged and two opposite strategies have been proposed: the use of a “tissue-based reference” extracted from the same tissue of the samples, and the development of a “universal reference” suitable for any tissue. However, these two choices imply two diverging philosophies of array data interpretation and pose different issues to the statistical analysis. The “tissue-based” strategy identifies a reference RNA as close as possible to the sample but from “normal” tissue, prepared either from cell lines or mixing RNA samples from tissues of normal donors. This approach implies the possibility to unequivocally define a “normal” sample (which can be not easy in studies oriented to cardiovascular risk), and the risk to end up working with very small differences between sample and reference, and therefore with log-ratios too close to 0. However, this strategy makes log ratio values directly interpretable without the need to compare the expression data with other samples to obtain meaningful information on differential expression.

On the other hand, the “universal-reference approach” is based on the principle that a perfect reference should be able to produce the maximum yield of positive spots throughout the array, and therefore the choice of the reference is array-based, instead of sample based. In this context, the perfect reference for whole-genome human arrays would hold significant amount of mRNA for any human gene. Since there is no single tissue able to express the whole genome, pooled RNA from multiple cell lines are usually employed [39]. This approach makes expression data not directly interpretable because the intensity values of the sample are normalized with a completely unrelated reference RNA, but allows to handle greater log ratio values for the statistical analysis, and virtually to directly compare data from different tissues.

In conclusion, there is no consolidated standard or guideline to help with the choice of the best sample and the best reference in blood-based transcriptomic studies, and further data are needed to better define: 1) the blood cell population most representative of any manifestation of CVD, 2) the best protocols for RNA isolation, and 3) the most advantageous reference strategy for inter-array and inter-sample data comparison.

1.7 Disease profiling by gene expression analysis in circulating cells

Based on what discussed in the previous sections, gene expression profiling in circulating white blood cells has a clear rationale as multi-marker strategy in personalized CVD risk assessment, and it is time to move to test this approach in a clinical research context. As functional biological marker, the high-throughput approach to CVD risk has the potential to provide not only a measure of the exposure to CVD risk, but to reflect the degree of physiological and biochemical response to the exposure. Clearly, at this stage, there are no definitive evidences pro or against this vision, but in the last years several observations have pointed to support this general hypothesis. In addition, a similar approach has been implemented in hemato-oncologic disorders, and also other diseases without a circulating cellular component. Characteristic blood transcriptomic patterns have been identified for infective, auto-immune, neurological and psychiatric diseases, and for environmental exposure to toxic substances.

The following table shows a revision of the published literature in this field.

Table 2: Studies demonstrating a correlation between clinical features of a disease and gene expression profile in circulating cells

Disease	Reference	Transcriptome correlates with clinical phenotype*	Diagnostic proof-of-concept †	Prediction of outcome or response to treatment ‡
Neurological diseases				
Migraine headache	40	X		
Tourette syndrome	41	X		X
Multiple sclerosis	42	X		X
Neurofibromatosis type I, Tuberous sclerosis type II, Down syndrome	43	X	X	
Huntington chorea	44	X	X	
Alzheimer	45	X		
Psychiatric diseases				
Post-traumatic distress	46	X	X	
Psychological stress	47	X		
Solid neoplasms				
Renal cell carcinoma	48,49	X	X	X
Toxic Exposure				
Benzene	50	X		
Lung diseases				
Asthma patients	51	X		X
Pulmonary hypertension	52	X	X	
Infective diseases and vaccination				
SARS	53	X		
Influenza immunization	54	X		
HIV	55	X	X	
Allergic diseases				
Contact dermatitis	56	X	X	
Allergic rhinitis	57	X	X	X

Disease	Referen ce	Transcriptome correlates with clinical phenotype*	Diagnostic proof-of-concept †	Prediction of outcome or response to treatment ‡
Autoimmune diseases				
Rheumatoid arthritis	58,59,60	X	X	
Systemic sclerosis	61	X	X	
Systemic lupus erythematosus	62,63,64	X		X
Inflammatory bowel disease	65	X		
Other systemic diseases				
Chronic fatigue syndrome	66	X	X	
Sarcoidosis	67	X		
Aldosteronism	68	X		
Sickle cell disease	69	X		
Trauma	19	X	X	

*Footnote to table. The table contains bibliographic references for non-cardiovascular diseases whose peripheral blood transcriptomic profile has already been investigated with high-throughput microarrays. Table also indicates with a "X" sign for each study whether: * a differential gene expression profile for the disease has been identified; † the study reports a classification analysis showing that gene expression profile can accurately identify the disease, and ‡ pre- and post- or with and without treatment groups with the same disease have been compared or a clinical outcome recurrence has been verified.*

The presence of highly specific gene expression patterns in circulating cells representative of diseases of different organs suggests that WBC behave as biosensors whose gene expression is influenced by circulating molecules produced in other organs. As such, gene expression of blood cells might "represent" pathological processes developing in any body system, giving valuable diagnostic and prognostic information. Preliminary data show an interesting degree of concordance in gene expression between WBC and certain tissues (i.e. central nervous system), based on sharing similarity in gene expression network of receptors and mechanisms of transduction [70,71].

To demonstrate that transcriptome profiling has the potential to describe the systemic cellular phenotype and to provide personalized, clinically-relevant information, several principles must be proved: first, it should be demonstrated that genomic data re-confirm consolidated notions on CVD risk. This point implies showing a different transcription fingerprint in presence of the disease or different phases of the disease, and after exposure to known risk factors. That means that transcription profiling should serve as diagnostic tool (case patient identification), as well as risk stratification strategy (high-risk individuals identification). In addition, a transcription profile is expected to change dynamically in presence of acute CVD events, to be predictive of future clinically relevant events, and to be modulated in presence of treatment. We need, therefore, to reproduce first already known information and to observe already expected changes in risk after treatment. These descriptive results should also be consistently related to the pathophysiological background of each one of these scenarios, and therefore observed gene expression modulation should fit in seamlessly.

After this preliminary observations, it needs to be demonstrated that gene expression profiling provide further (and more personalized) information than what carried by classical risk factors and mono-marker strategies. In other words, a transcription profile of circulating cells should improve the coverage of knowledge beyond the limits intrinsic to the clinical and current singular biomarker approach. In this context, deviations from the information given by clinical risk factors are expected, and personalized profiling should effect a quantum leap in CVD prediction/ prognosis when compared to current risk stratification algorithms. The gene expression profile is also expected to partially change during treatment (markers of drug response), but also to be in part unaffected providing an accurate quantification of the residual risk.

Finally, we expect that gene expression can accurately identify the fingerprint of a disease, also among other conditions with similar pathophysiological mechanisms and therefore be able to distinguish CVD from other pro-inflammatory states such as autoimmune or infective diseases. Similar discrimination power should be provided among different causal factor for a common clinical phenotype. For instance, a fingerprint for high cholesterol-related CVD should be distinguishable from the diabetes or smoking-associated CVD signatures.

So far, preliminary evidences are available as proof-of-concept that individuals with acute CVD events are different in leukocyte transcriptomic profile compared to healthy controls [72,73,74,75], that exposure to risk factors is associated with abnormal gene expression [76], and that treatment can partially revert these alterations [77,78,79,80]. Classification and diagnostic evidences have also been provided [81]. However, no proof studies are available on prognostic significance of microarrays in preventive cardiology, and the published literature is still addressed to the average behavior of genomic profiles, instead of working with single-patient unique phenotype.

The following sections describe the available proof-of-concept data, and clearly suggest that blood leukocyte fingerprint can be used in several clinically relevant situations as surrogate for other tissues, providing markers of disease/ treatment and new insights into the disease pathophysiology and drug mechanism of action.

1.7.1 Gene expression profiling can portray acute cardiovascular events

Preliminary animal studies suggested that acute cerebral injuries of different origin (ischemic stroke, intracerebral hemorrhage, status epilepticus, and insulin-induced hypoglycemia) are portrayed by a different gene expression profile in circulating cells compared to control animals [73], and that the genomic signature univocally identifies the type of injury [74].

Although only little information is available in humans, published data clearly show a different transcriptomic profile in presence of acute stroke compared to control subjects [72,73]. Moore and colleagues evaluated the peripheral blood mononuclear cell (PBMC) gene expression profile during the acute phase of 19 patients with CT confirmed stroke and 19 controls to investigate whether a systemic genomic signature could be demonstrated as fingerprint of acute cerebro-vascular event [75].

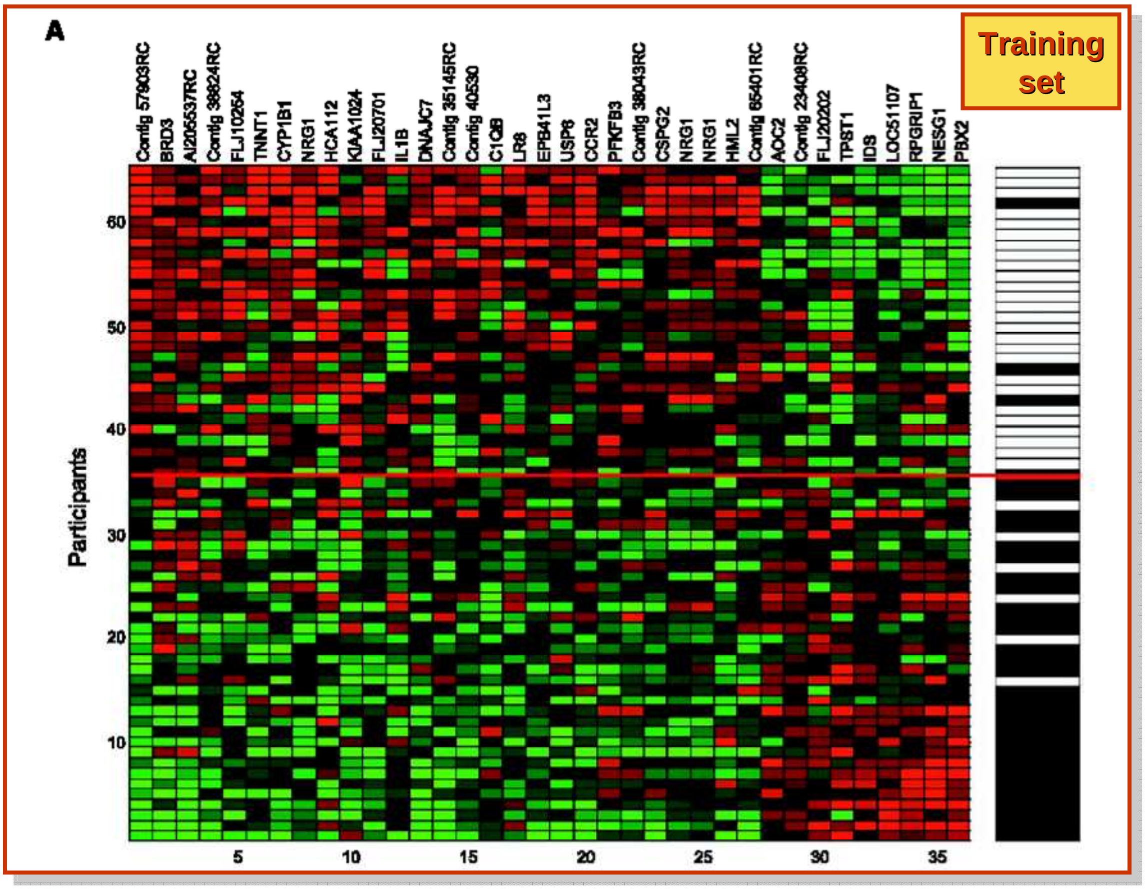
About 200 genes resulted differentially expressed in stroke patients, even though a quite conservative approach for comparison was used, including multiple comparison corrections, false discovery rate conservative setting, and permutation analysis. Moreover, the selection of a highly representative set of 22 genes by Prediction Analysis for Microarrays (PAM) [82], a supervised classification algorithm based on shrunken centroids, allowed to distinguish between cases and controls with high accuracy (Sensitivity 89%, Specificity 95%). When the small PAM-selected list of genes was retested in an independent cohort (9 cases and 10 controls), the classification performance still resulted in 78% sensitivity and 80% specificity.

In addition to this proof of concept of the possible role of transcriptomic signatures for diagnostic purposes, it is of interest to note that, looking at the results at large, several broad classes of genes differentially expressed in stroke patients were related to leukocytes activation and differentiation, response to hypoxia, vascular repair, and included genes potentially associated with an altered cerebral microenvironment. Clearly PBMCs were neither hypoxic themselves nor in a hypoxic environment, but in any case they were able to “sense” and represent the ongoing ischemic insult to the brain tissue (hypoxia reaction genes), suggesting the presence of a local or systemic signaling. It is even more interesting to note that stroke vascular risk measured by the Framingham score accounted for about 28% of the variance of the gene expression profile in the group of acute ischemic stroke patients as to emphasize that transcriptomics may provide an accurate snapshot of an acute CVD event, and this representation has only partial overlap with the one portrayed by classical risk factors.

1.7.2 Gene expression profiling in presence of environmental exposure to smoking

Smoking is a clear and strong environmental factor with the highest probability of yielding a detectable and repeatable signature in white blood cell transcriptome among the classic risk factors for CVD. In a recently published paper, authors [76] identified 42 smokers and 43 non-smokers based on self-reported questionnaire and confirmed by plasma concentration of cotinine, a stable metabolite of nicotine. The two groups of subjects were fairly well matched for clinical profile of CVD risk factors and demographics, and more than 800 genes where identified as signature of smoking exposure. Differentially expressed genes were then ranked in descending order by the absolute value of Pearson's correlation to plasma cotinine concentration with the aim of identifying a small predicting set to classify smokers and non-smokers (as shown in the next figure).

Figure 9: Heatmap of differentially expressed genes between smokers and non-smokers

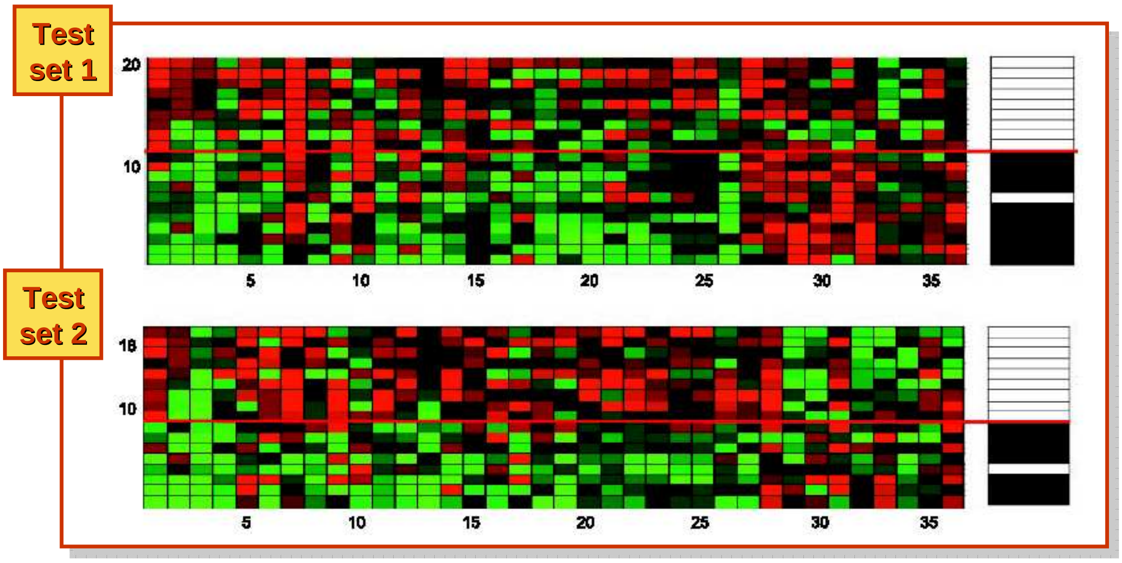


Footnote to figure. The heatmap shows the relative expression value of most differentially expressed genes in smokers compared with non smokers. Each column represents a gen and each row a subject. Smokers are represented in white (open boxes) and non-smokers in black (filled boxes).

The final set of 36 top candidate genes (27 directly and 9 inversely correlated to cotinine levels) was defined based on total misclassification error minimization algorithms and performed with 81% sensitivity and 88% specificity.

A test set of 10 smokers and 10 nonsmokers, both cotinine confirmed, was used for model validation in two different time points yielding an impressingly high accuracy (about 90% sensitivity and 100% specificity in both occasions) and showing great consistency and stability over the time, as shown in the next figure.

Figure 10: Accuracy of gene expression markers in the identification of smokers



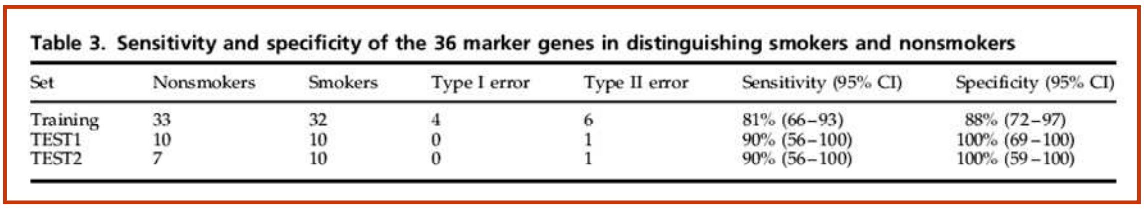


Table 3. Sensitivity and specificity of the 36 marker genes in distinguishing smokers and nonsmokers

Set	Nonsmokers	Smokers	Type I error	Type II error	Sensitivity (95% CI)	Specificity (95% CI)
Training	33	32	4	6	81% (66–93)	88% (72–97)
TEST1	10	10	0	1	90% (56–100)	100% (69–100)
TEST2	7	10	0	1	90% (56–100)	100% (59–100)

Footnote to figure. The heatmaps show the relative expression value of most differentially expressed genes in smokers compared with non smokers in the training set, when tested in two test sets. Each column represents a gen and each row a subject. Smokers are represented in white (open boxes) and non-smokers in black (filled boxes).

This study clearly demonstrates that the presence of a highly-defined clinically-relevant phenotype (chronic exposure to cigarette smoking) can be accurately identified on the basis of mRNA expression in peripheral leukocytes, which are neither the toxicological target for nicotine nor directly exposed to cigarette smokes.

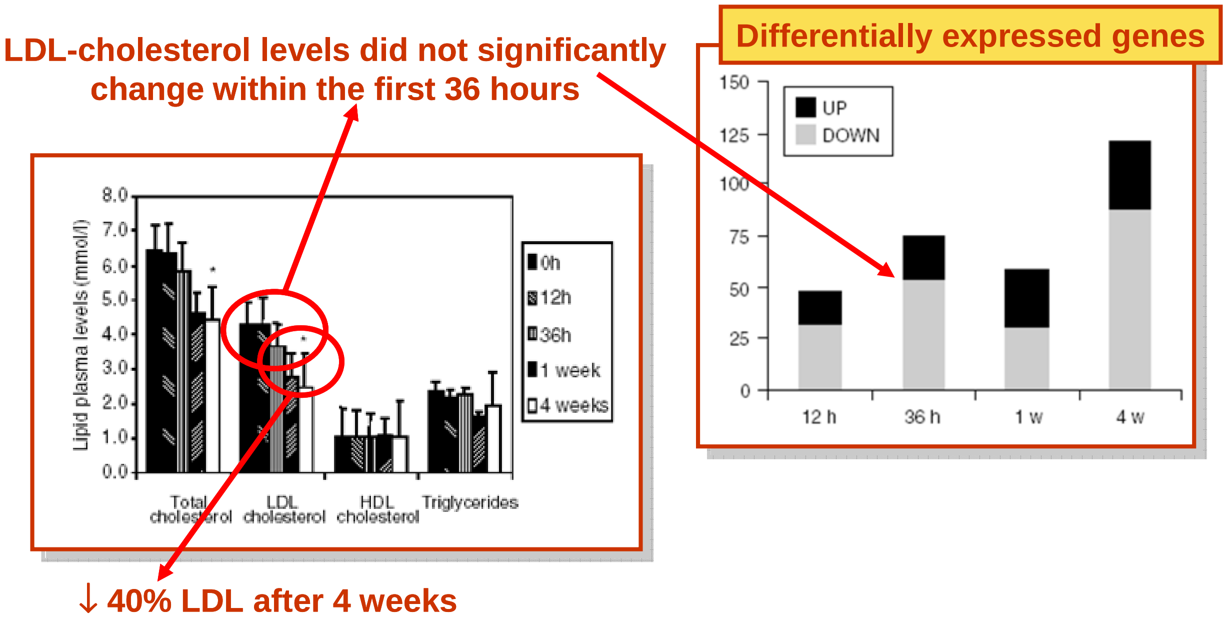
1.7.3 Gene expression profiling after treatment

Biologically relevant gene expression signatures for pharmacological treatment have also been demonstrated with high-throughput techniques. Expression profiling in hypertension research has risen high expectations on the one hand [83], and has evoked criticism on the other hand [84]. The optimistic view envisioned physiological genomics as a tool to unravel pathways in hypertension, and to guide therapy (i.e. pharmacogenomics); critics argued that the likelihood that gene profiling experiments would lead to the discovery of new, relevant pathways was low and that the value of time course experiments in experimental hypertension was limited [84]. In a small pilot study gene expression profiles of leukocytes of patients with essential hypertension were compared with profiles obtained from leukocytes of carefully-matched healthy control subjects [78]. Subjects were matched for age and body weight; smoking subjects were excluded from the experiments. Cholesterol, glucose and creatinine levels were not different. The leukocytes of the essential hypertensive patients revealed 606-differentially-expressed genes. Genes of interest included a group of genes involved in chemotaxis, a number of cytokine receptors, and a number of genes involved in transmission of cytokine signals. Also, interestingly, the AT1 receptor displayed an increased expression level, which was confirmed by conventional RT-PCR analysis. Most remarkably, the vast majority of changes observed in the untreated group were not any more visible in a separate group of treated patients. It should be remarked that in this pilot study RNA within groups was pooled to obtain sufficient material for microarray analysis and no conclusions can be drawn about individual expression levels.

In a relatively recent study [79], 11 male patients with primary hyperlipidemia were treated with Atorvastatin, 20 mg daily for 4 weeks, yielding an average decrease in LDL cholesterol of 40%. Blood specimens were collected at baseline, after 12 and 36 hours from the first drug administration, and after 1 and 4 weeks of treatment. PBMC gene expression was evaluated at all time-points with the aim of investigating early and late effects of statins. A total of 240 genes were significantly regulated by atorvastatin during the 4 weeks of study, and the differential expression signature for anti-dyslipidemic treatment included several genes involved in hemostasis, inflammation and other processes critical to atherosclerosis.

As expected, LDL-cholesterol levels did not significantly change within the first 36 hours of treatment, but 48 genes were already differentially regulated after 12 h and 75 after 36 h, as shown in the next figure.

Figure 11: Effect of Atorvastatin treatment on gene expression profile in WBC


LDL-cholesterol levels did not significantly change within the first 36 hours

Differentially expressed genes

↓ 40% LDL after 4 weeks

Since no significant effect on plasma lipid levels was detectable at these time-points, it is likely that the atorvastatin-induced modulation of gene expression was independent of lipid-lowering. Then, 58 additional genes resulted differentially regulated after 1 week and 121 after 4 weeks of treatment, when plasma lipid levels were significantly decreased, and therefore these changes in gene expression may be secondary to lipid lowering. The presence of different patterns of response to atorvastatin over time suggests that its impact on gene expression, and therefore cell biology, implies both lipid-dependent and independent effects. Since it is debated whether the cardioprotective effect of statins goes beyond its action on cholesterol synthesis and lipid homeostasis [80], microarray studies with adequate time course design can provide original, descriptive and, most important, in-vivo insights into cell pathophysiology and advocated pleiotropic effects.

Further evidence that circulating cell gene expression profile can serve as functional biological markers of drug exposure with a clear link to disease pathophysiology is provided by the results of an unsupervised hierarchical clustering performed on this repeated-measures study to identify possible functional/ temporal associations. Clustering approach identified a group of genes associated with proinflammatory effects that resulted upregulated soon after the start of treatment and then progressively down-regulated over-time. These genes were all functionally inter-related and involved in INF-alpha response pathway. This apparently surprising observation is supported by in vitro studies showing that lipophilic statins (such as atorvastatin) induce a transient proinflammatory response in monocytes [81].

2 Study design and aims

2.1 Aim of the project

Based on the scientific background provided in the previous chapters, we decided to provide preliminary data to assess the presence of a relationship between gene expression profile in circulating PBMCs and the presence of cardiovascular risk factors in apparently healthy young-adults without significant diseases or chronic pharmacological treatments.

As already outlined, most of the available data were reported by single research groups, often in underpowered samples and almost never at a whole genome level. For this reasons, we intended to verify whether the gene expression profile of circulating PBMCs had pro-inflammatory features in subjects at low risk with moderate exposure to CVD risk factors, assessing the expression profile at whole genome level in a relevant cohort of subjects.

2.2 Study design and phases

The conducted study is a monocentric, observational, cross-sectional study. The study activities were carried out in two different time periods.

In the first instance, a small pilot was conducted in 50 healthy subjects to pilot the expression profile results and decide whether to collect data from a larger cohort. Then, this pilot experience was followed by a larger, confirmatory recruitment effort.

The overall size of the enrolled population is 167 individuals of both genders. Data here reported are from all enrolled subjects.

3 Population and selection criteria

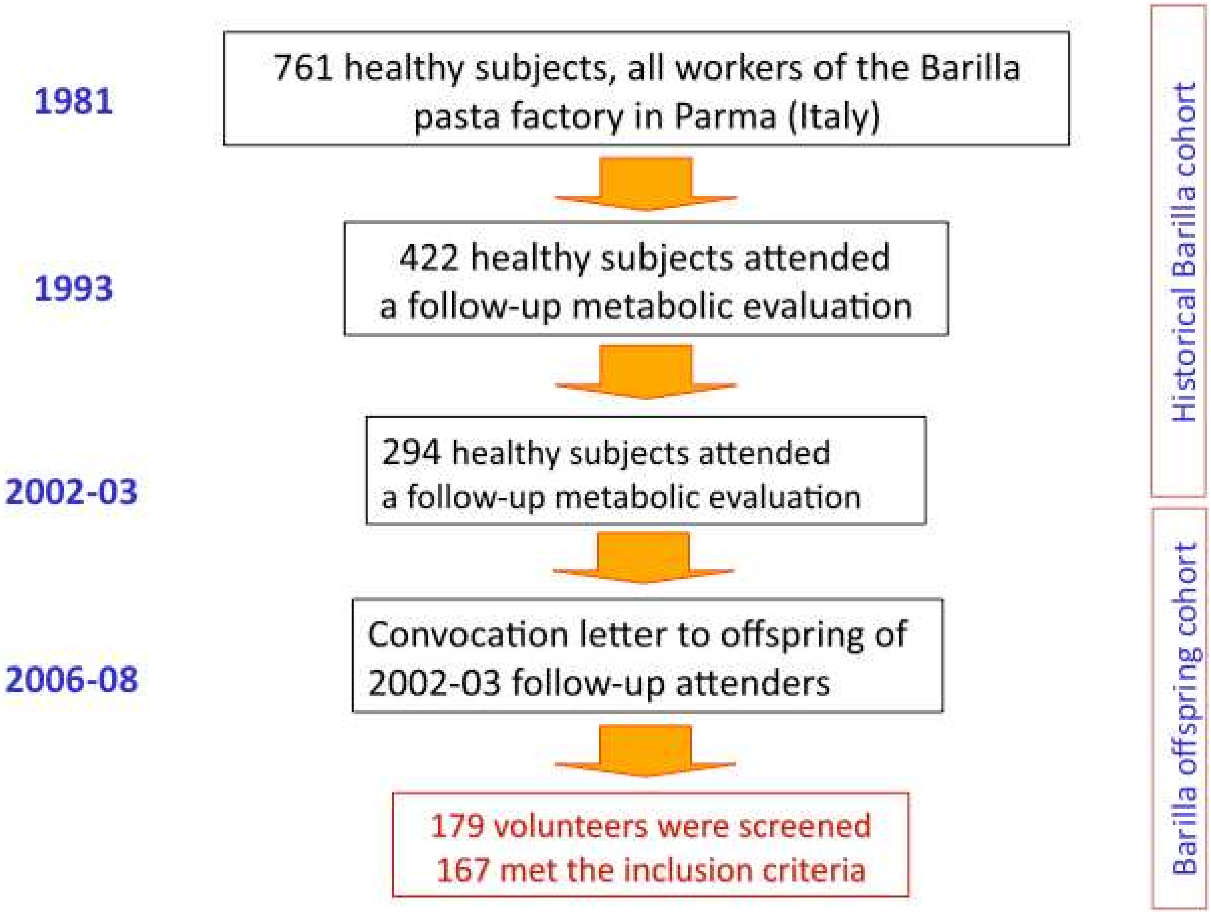
The study population included a total of 167 apparently healthy volunteers of both genders, all young adults (18-65 years) without medical history or clinical findings of major chronic diseases and not undergoing any pharmacological treatment.

3.1 Population screening

Participants were selected as offspring of volunteers attending the Barilla study, an epidemiological survey started in 1981 to investigate several aspects of the impact of insulin resistance on human disease [85,86]. To recruit the "Barilla Offspring" population, a letter was sent to all the 422 parents who attended the last follow-up visits [87] asking them to inquire the interest of their offspring to participate into the study. A total of 179 offspring accepted the invitation and were screened.

The following flow-chart depicts the Barilla offspring cohort.

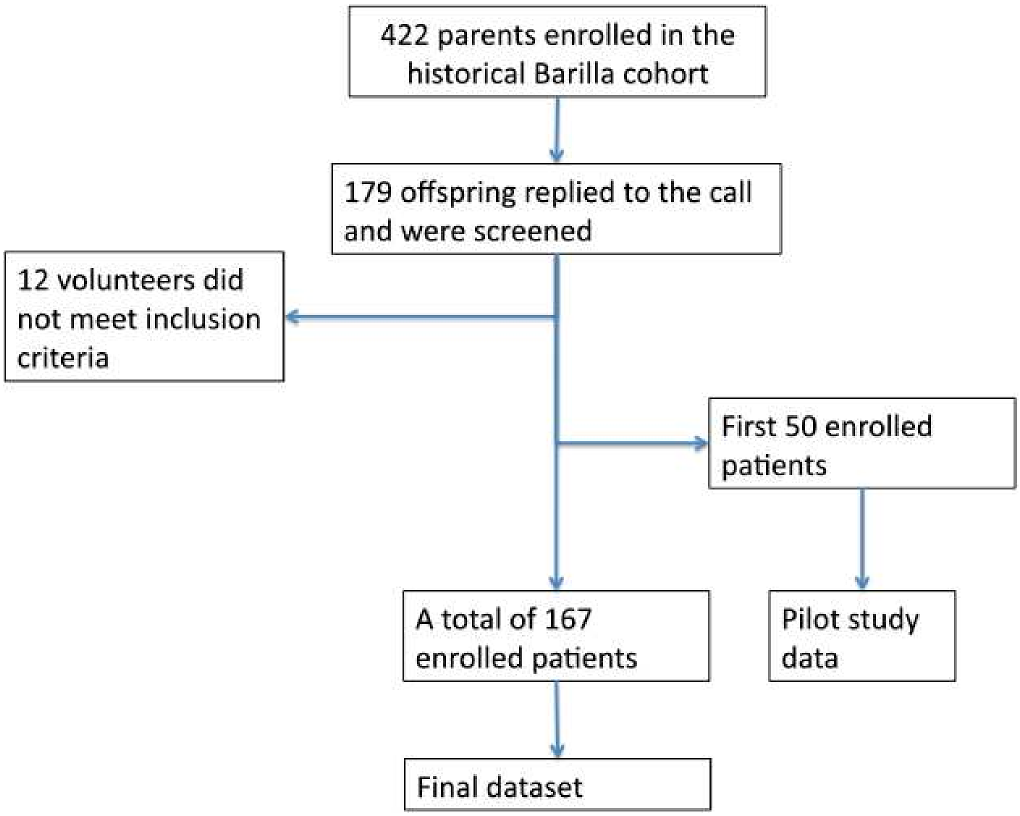
Figure 12: Barilla Offspring cohort flowchart


1981
761 healthy subjects, all workers of the Barilla pasta factory in Parma (Italy)
1993
422 healthy subjects attended a follow-up metabolic evaluation
2002-03
294 healthy subjects attended a follow-up metabolic evaluation
2006-08
Convocation letter to offspring of 2002-03 follow-up attenders
179 volunteers were screened
167 met the inclusion criteria
Historical Barilla cohort
Barilla offspring cohort

Out of the 179 screened subjects, 12 did not meet the inclusion criteria for presence of concomitant diseases. The first 50 subjects meeting the inclusion criteria were included in a pilot dataset whose gene expression profiling was performed before completing the enrollment of the overall cohort.

The following flow-chart shows patient's disposition in the study.

Figure 13: Barilla Offspring cohort disposition


422 parents enrolled in the historical Barilla cohort
179 offspring replied to the call and were screened
12 volunteers did not meet inclusion criteria
First 50 enrolled patients
A total of 167 enrolled patients
Pilot study data
Final dataset

Subjects have been fully informed on the aims of the study, experimental design and study procedures in order to sign the consent form. The study was approved by the local institutional ethic committee. All screened subjects signed the consent form before participation.

3.2 Selection criteria

During the screening visit, the following inclusion and exclusion criteria were checked. Only subjects meeting all the inclusion and none of the exclusion criteria were enrolled.

3.2.1 Inclusion Criteria

Patients with the following inclusion criteria entered the study:

- Both genders;
- Age >= 18 or < 65 years;
- For women: being in the fertile age (as defined by presence of menses in the last 6 months);
- Absence of major diseases or other medical exclusion criteria enlisted below;
- Daily alcohol intake below 80 gr
- Written informed consent to participate into the study

3.2.2 Exclusion Criteria

Patients with at least one of the following exclusion criteria entered the study:

- Age < 18 or > 65 yrs
- Postmenopausal women
- Pregnancy
- Previous cardiovascular events
- Type 1 or type 2 diabetes
- Medical history of malign cancer, or ongoing cancer disease with the exception of skin cancer (excluding melanoma) and thyroid cancer (excluding medullary and anaplastic histotypes)
- Acute illness
- Autoimmune or inflammatory chronic illness
- Organ failure or severe systemic illness
- Psychiatric illness or diseases possibly causing a poor compliance to the study protocol
- Drug addiction or alcohol intake> 80 gr/day

4 Study visits

After direct phone interview with the volunteers to rule out evident screening failures, subjects were invited for a first meeting during which investigators provided exhaustive information on the study. If the subject was interested in participating into the study, the informed consent was signed at this time and the volunteers entered the screening procedure. If all inclusion and none of the exclusion criteria were met, the subject was enrolled and two further appointments were scheduled (visit 1 and visit 2).

The following table shows the content of the three study visits.

Table 3: Content of the study visit

Visit	Tasks	Type of parameters recorded
Screening	▪ Subject's information ▪ Signing of the informed consent ▪ Unique ID assignment	
Metabolic visit (V1)	▪ Personal and medical data gathering ▪ Physical examination ▪ Fasting blood drawn ▪ Urine collection and storage ▪ Oral Glucose Tolerance Test (OGTT) with multiple blood drawn ▪ PBMCs isolation from peripheral blood and storage ▪ RNA extraction from a PBMC aliquote	▪ Personal and family medical history ▪ Clinical data ▪ Laboratory data ▪ PBMC samples ▪ RNA for array hybridization
Cardiovascular visit (v2)	▪ Carotid ultrasound	▪ Intima-Media Thickness (IMT)

The content of the three visits is here described in more detail:

4.1 Screening visit (V0)

Study information

Subjects were personally and exhaustively informed by the investigators about the aims, procedures, possible risks and timeline of the study. The informed consent was red and commented together with the investigators.

Signing of the informed consent and ID assignment

The informed consent was signed in duplicate both by the participant and the investigator in the presence one of the other. One copy was given to the participant and one copy stored in the study records. An unambiguous consecutive ID number was assigned to each single participant.

4.2 Metabolic visit (V1)

Participants were scheduled at 8 o'clock in the morning, fasting, without taking any medication. The metabolic evaluation was performed at Parma University, Department of Internal Medicine and Biochemical Sciences.

Fasting blood sample collection

Twelve hours fasting venous blood samples were collected and subjects asked to refrain from smoking and drink coffee the morning of the visit.

The following biochemical evaluations were performed:

- Hemocytometric analysis and full blood count
- Glycemia
- Lipid profile

  - Total cholesterol

  - HDL cholesterol

- Triglycerides

- Liver enzymes (AST, ALT, gammaGT, ALP)
- Uric acid

The concentrations of glucose, insulin, total cholesterol, high-density lipoprotein-cholesterol (HDL-C) cholesterol, triglycerides (TG), AST and ALT were quantified as previously described [88]. Low-density lipoprotein-cholesterol (LDL-C) was calculated using Friedewald's Formula.

Plasma / urine storage

Plasma and 24 urine collection samples were stored at -80°C in order to create a "sample bank" for post-hoc biochemical analysis.

The following parameters were measured on stored samples:

- Plasma insulin concentration
- Nitrates/nitrites (as index of overall nitric oxide plasma concentration)
- Adhesion molecules (such as ICAM-1, VCAM-1, E-Selectin)
- High-sensitivity PCR (Hs-CRP)
- Urine
- Microalbuminuria, urine proteins
- Urine creatinine

Oral Glucose Tolerance Test (OGTT)

A baseline, fasting blood sample was drawn. Subjects were then given a glucose solution (75 gr) to be drunk within 5 minutes. Blood samples were drawn at prefixed intervals (30, 60, 120 minutes) after glucose ingestion to measure glucose and insulin levels.

Personal and medical history data collection

A complete medical history and physical examination was conducted as previously described [89].

Anamnestic data were collected with a questionnaire using close or semi-close questions and the investigator entered data in an electronic CRF.

Personal data collected to be analyzed included:

Personal medical history:

1) Synoptic variables concerning the main cardiovascular diseases, risk factors, and metabolic disturbances (this list of question served as double check on the inclusion criteria delineate above)
2) Systematic information of any past or present disease of any kind using a validated international classification code (International Classification of Diseases v9, ICD-9). For any disease or pathological condition, starting and (if applicable) ending dates were also collected

Family history with close question aiming to prevalence of major cardiovascular diseases and risk factors in first degree relatives, including

- Diabetes and glucose intolerance
- Dyslipidemia and hypertension
- Heart, brain and peripheral vascular diseases

Personal habits and lifestyle, including:

1) Smoking. The following data were recorded: smoker/ non smoker/ past smoker, cigarette/ cigar/ pipe smoking, years of smoking, number of cigarette/ cigar smoked per day, packs/ year of smoking, years since quitting (if applicable)
2) Alcohol intake. The following data were recorded: daily alcohol/hard liquor consumption with gram calculation. Data from this questionnaire were afterward matched with more accurate estimation of alcohol consumption from 7-day food intake records and overall diet food frequency questionnaires on macronutrients and major micronutrients intake. Alcohol intake was estimated by frequency assumption questionnaire and expressed as grams per week. We considered 12 g of alcohol from a glass of wine, and 13 g from a can of beer or from a small glass of spirits.
3) Physical exercise. The following data were recorded: weekly physical activity estimated using a validated questionnaire, taking into account both work-related and leisure physical exercise

Physical examination

A complete physical examination was performed, including heart, vessels, chest, abdomen, and lymph-node exploration. Anthropometrics and blood pressure were recorded for statistical analysis.

The following anthropometric parameters were measured

1) Body height, to the nearest 0.5 cm using a stadiometer with subjects in the erect position without shoes
2) Body weight, using an electronic balance
3) Waist circumference, to the nearest 0.5 cm using a 1 cm-wide measuring tape following the NHANES III protocol. The waist circumference was measured in standing subjects placing a measuring tape in a horizontal plane around the abdomen, half-way between the iliac crest and the rib edge on the mid right axillary line. Before reading the tape measure, the tape must be secure, but not too tight and parallel to the floor and the reading should be taken at the end of expiration.
4) Systolic and Diastolic blood pressure, and heart rate, using a Mercury Sphygmomanometer (Riva-Rocci) from the non-dominant arm. The first and fifth Korotkoff sounds were taken to identify systolic and diastolic values respectively. Heart rate will be measured according to the palpatory method (30 seconds) after recoding the blood pressure. Blood pressure measurements were performed every 2 minutes by the same investigator until two measurements will differ no more than ±5 mm Hg for both systolic and diastolic value. The average of three measurements was considered for the study.

Plasma/urine storage

Plasma and urine samples will be stored in 200uL plasma aliquots and in 5ml urine aliquots respectively at – 80°C following standard procedure.

4.3 Cardiovascular visit (V2)

The cardiovascular visit will be performed at the Metabolic Diseases outpatient clinic of the Department of Internal Medicine and Biochemical Sciences (Parma University). Subjects will be required to be fasting and refraining from tea or caffeine intake since the previous day. Subjects will be asked not to take any medication before the visit but to carry medication with them to be taken after the visit is concluded.

Carotid intima-medial thickness (IMT)

A Doppler ultrasound carotid intima-medial thickness evaluation were performed in all subjects participating in the study.

Measurement of the carotid intima-media thickness (IMT) is a non-invasive, ultrasound based test to quantify subclinical vascular disease which has been shown to be an independent predictor of myocardial infarction, stroke and death from coronary artery disease [90]. IMT assessment, together with carotid plaque evaluation, is a useful tool to redefine risk beyond traditional risk factors [91].

Subjects were examined in a supine position, comfortably position after removing clothing and jewelers in the area to be examined.

A trained operator, blind to the patient general characteristics, performed all tests using a linear-phased multi-frequency (from 7 to 10 MHz) transducer on a HP SONOS 5500 ultrasound device.

The examiner was seated at the patient's head. An ultrasound-specific gel was applied on the neck to help the transducer making a secure contact with the skin.

A continuous ECG registration was performed during the test and ultrasound image frames recorded on a digital support in the telediastolic phase of the cardiac cycle (deflection from the isoelectric line measured from the beginning of the QRS complex will be used as a reference), in order to avoid IMT variations related to carotid pulsatility.

The imaging protocol involves obtaining three longitudinal sections with lateral, and anterior oblique views of the distal 10 mm of the right and left common carotid arteries, the carotid bifurcation, and the internal carotid artery.

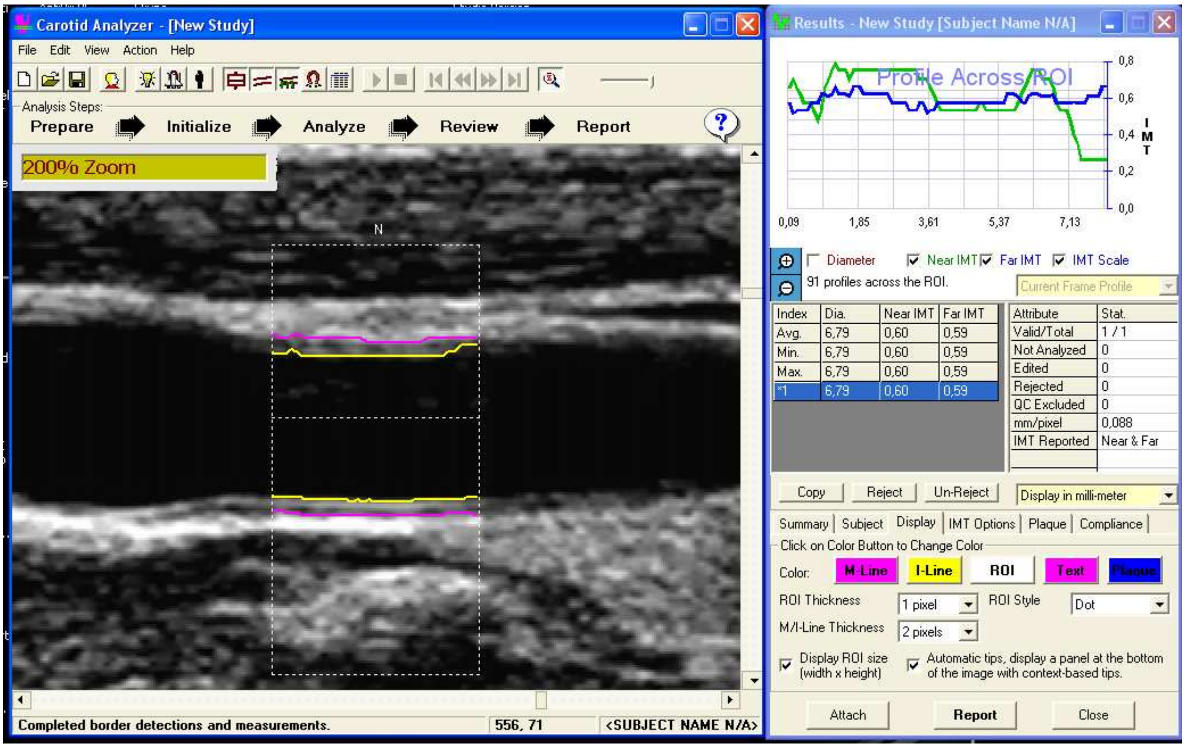
For each main carotid segment (common carotid arteries, the carotid bifurcation, and the internal carotid artery) longitudinal sections are recorded in loops of specific images in order to better visualize both the near and the far walls.

For each projection, an electrocardiographic-triggered loop of at least 4 heart beats was recorded and subsequently analyzed by a single operator, using a semi-automated edge-detection and measurement system (Carotid Analyzer; Medical Imaging Applications LLC, Iowa, US).

The maximum and average IMT was calculated performing measurements in 36 different points (2 each side X 3 projections X 2 walls).

The following figure shows the identification of near and far wall using a edge-detection software

Figure 14: Snapshot of IMT identification during carotid ultrasound examination



□ Diameter ☑ Near IMT ☑ Far IMT ☑ IMT Scale

91 profiles across the ROI.

Index	Dia.	Near IMT	Far IMT
Avg.	6,79	0,60	0,59
Min.	6,79	0,60	0,59
Max.	6,79	0,60	0,59
*1	6,79	0,60	0,59

Attribute	Stat.
Valid/Total	1 / 1
Not Analyzed	0
Edited	0
Rejected	0
QC Excluded	0
mm/pixel	0,088
IMT Reported	Near & Far

Display in milli-meter

Click on Color Button to Change Color

Color:

ROI Thickness: 1 pixel

ROI Style: Dot

M/I-Line Thickness: 2 pixels

☑ Display ROI size (width x height)

☑ Automatic tips, display a panel at the bottom of the image with context-based tips.

Completed border detections and measurements. 556, 71 <SUBJECT NAME N/A>

5 Gene expression profiling

5.1 Peripheral blood mononuclear cells (PBMCs) isolation and processing

Approximately 50 mL of peripheral blood was collected in fasting conditions from each volunteer in a BD Vacutainer tube containing EDTA as anticoagulant (Becton, Dickinson and Company, Franklin Lakes, NJ, US). PBMCs isolation was carried out in sterile condition under a laminar flow cabinet. Briefly, blood was aliquoted (10ml) and diluted 1:1 with Phosphate Saline Buffer (PBS), mixed by vortexing and centrifuged at 1400 rpm for 10 minutes. Serum supernatant containing platelets was removed leaving a 3-4 ml layer of leucocytes lying over a red blood cells pellet. A total of 5 ml of Lymphoprep (Amersham Pharmacia, UK) was added in each tube and samples were centrifuged at 2200 rpm for 20 minutes at room temperature in order to isolate PBMCs following a centrifugation gradient.

PBMCs (a layer between serum and Lymphoprep) were collected with a maxi pipette at the interface. Approximately 4 ml of PBMCs was obtained, of which 1ml was used for RNA extraction and the remaining 3ml were further processed to be stored in liquid nitrogen. The 3ml PBMC layer was diluted 1:3 with HBSS (Hanks buffered saline solution with calcium and magnesium) (Sigma), samples were centrifuged at 1800 rpm for 10 minutes at 4 °C and each pellet was re-suspended in an equal volume of RPMI 1640 medium 20% Foetal Calf Serum (FCS). A RPMI 20% dimethyl sulfoxide (DMSO) solution was added drop by drop to the cell solution every 3-5 minutes. Cell suspension was aliquoted in 1ml criovials and placed at -80°C for 12 hours and subsequently placed in liquid nitrogen for final storage.

5.2 RNA extraction and quality control

One PBMC aliquot (1ml) was used for total RNA extraction for gene expression analysis.

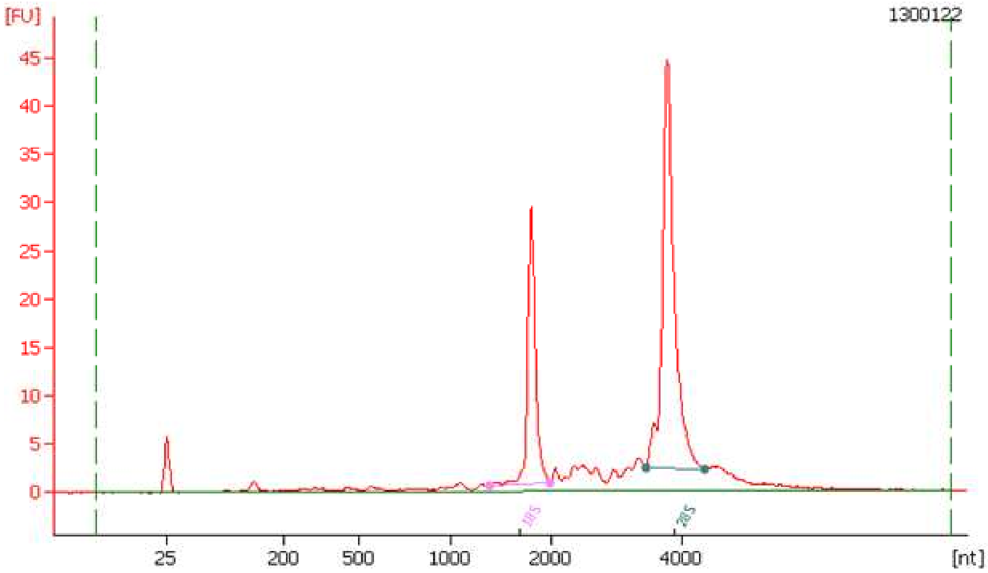
RNA extraction was performed using a commercial Kit (Qiagen RNeasy Mini kit) following manufacturer's instructions[92]. Isolated RNA was also digested with RNase-Free DNase to remove possible DNA contamination. Samples were run onto 0.8% agarose RNase-free TAE agarose gel containing 0.5ug/ml Ethidium Bromide in standard RNase free conditions to check for RNA quality (degradation or DNA contamination). RNA was then eluted with RNase-free water into a new microcentrifuge tube and stored at -80°C until hybridization.

Extensive quality control testing of extracted RNA was done before hybridization to the array.

Total RNA yield and absence of DNA contamination were checked using a NanoDrop ND-1000 UV-Vis Spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA). RNA quality was assessed measuring the 28S/18S rRNA and the RNA Integrity Number (RIN) with a Bioanalyzer 2100, using RNA 6000 Nano Chips (Agilent Technologies, Santa Clara, CA, USA).

The following picture shows the typical pherogram observed with Agilent's Bioanalyzer for PBMC's RNA.

Figure 15: PBMC's RNA spectral analysis



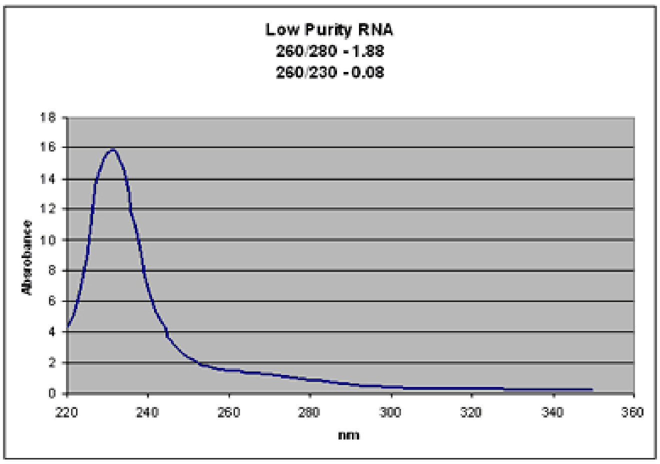
Assessing the purity and the quality of extracted RNA is fundamental in a microarray experiment.

The gold standard methods used in this field are the Nanodrop spectrophotometer and the Bioanalyzer 2100. Nanodrop calculates the 260/280 and 260/230 adsorbance ratio, besides the RNA concentration.

First ratio gives an estimation of protein contamination, and it must be included from 1.8 and 2. The second one reflect carbohydrates contamination, and it must be grater than 1.5.

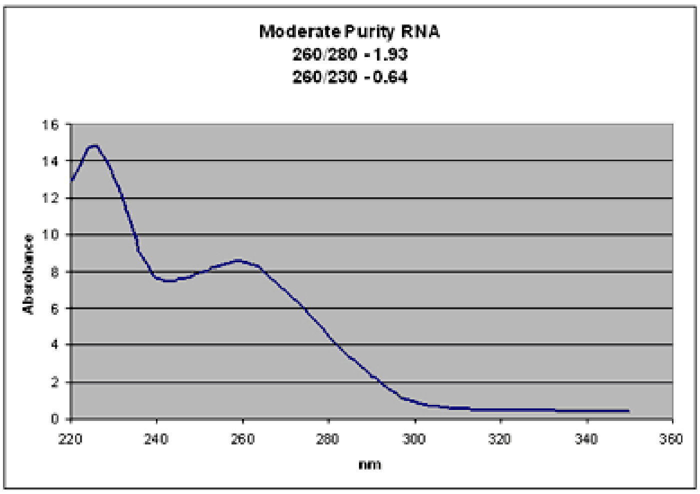
The following graph shows a low purity RNA that can't be used for microarray experiment. A low quality RNA sample was discharged from further processing.

Figure 16: Low purity RNA sample graph



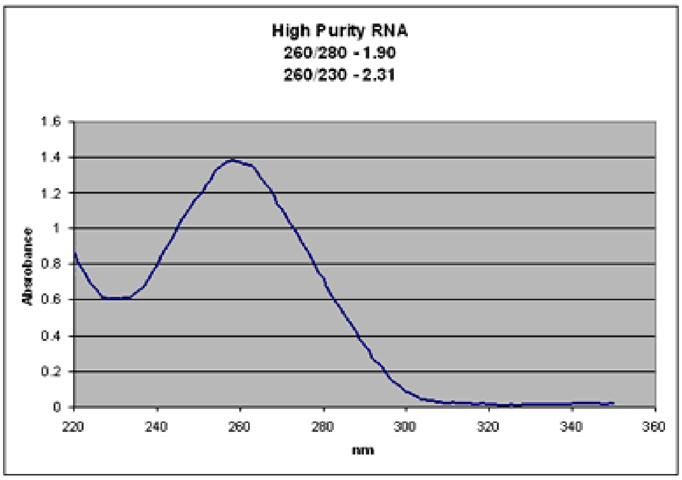
This graph shows a RNA sample with medium purity that has to be purified before experiment. This type of specimen quality was considered for hybridization only if the amount of extracted RNA allowed to perform an adequate purification process.

Figure 17: Medium purity RNA sample graph



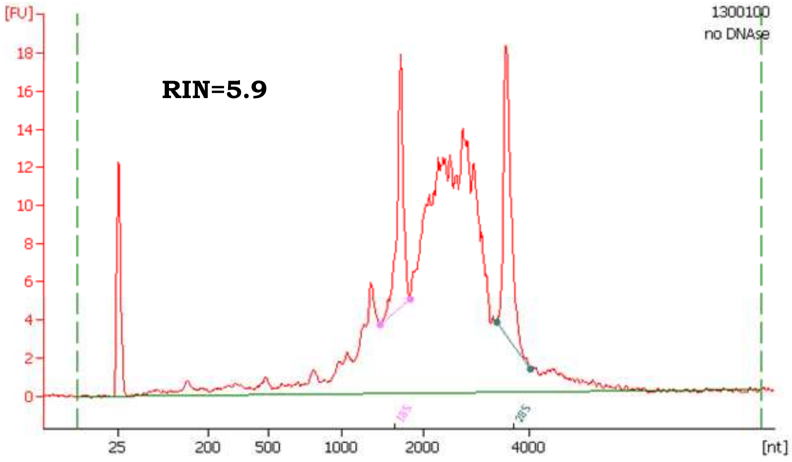
The following graph shows the NanoDrop result for a good quality RNA sample. This type of sample was amplified, labeled and hybridized without further pre-processing.

Figure 18: NanoDrop representation of good quality RNA sample

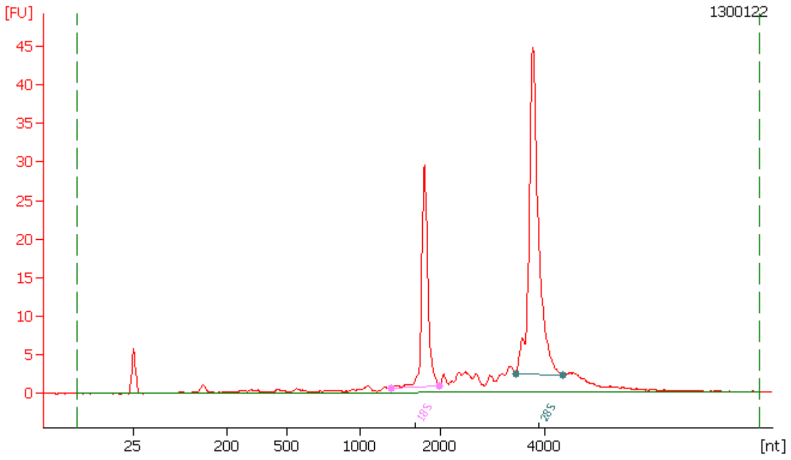


Quality control was performed with Bioanalyzer 2100. This instrument gives as results electropherograms showing the degree of integrity of extracted RNA, with two distinct peaks representing the 18S and 28S ribosomal RNA and giving the RNA Integrity Number, that indicates the degree of integrity of RNA.

1) Partially degraded RNA



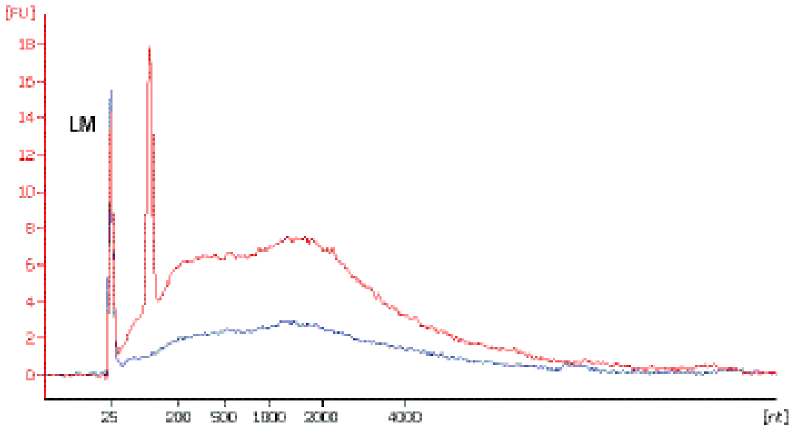
2) Good quality RNA



After cRNA amplification and dye incorporation it is necessary to perform another control with Bioanalyzer. The resulting electropherogram should have a broad band. The majority of signal for amplified sample should fall into the size range from 200 to 2000 nucleotides.

Red, Cy5 labeled cRNA; Blue, Cy3 labeled cRNA show the same size distribution.

Figure 19: Distribution of a good RNA electropherogram after cRNA amplification



5.3 Microarray hybridization and quality control

As mentioned above the dual-color protocol allows a competitive hybridization among RNA sample and RNA reference, so that the RNA that expresses greater amounts of a given gene will competitively inhibit the hybridization of the other target. Thus at each coordinate on the slide, red or green fluorescence intensity is directly proportional to the amount of target present in one RNA sample compared to the other.

A total of 500ng of total RNA for each sample (and each related reference) was reverse transcribed in cDNA in parallel with a given amount of spike-in RNA (Agilent Technologies, US) as internal control, followed by linear amplification and transcription in cRNA with the incorporation of a fixed proportion of nucleotides conjugated with fluorescent dyes. Reference RNA was labelled with green fluorescent molecule cyanine 3 (Cy3) while sample RNA was tagged by red fluorescent Cyanine-5 (CY5). After labeling, conjugated-cRNA were purified through adsorbance columns (Qiagen, US) and re-tested for quality and integrity. Finally, sample and reference cRNA were mixed together in equal concentration and hybridizated together on 4X44 Whole Human Genome Agilent Microarray slides (Agilent Technologies, ref. G4112F). After 17 hours of incubation, slides were washed to discard unbounded cRNA and then scanned with an Agilent's G2565AA Microarray Scanner System. From the acquired high-resolution 16-bit tiff image, raw intensity data were extracted in numerical values using a dedicated software (Agilent's Feature Extraction, 9.5 V), which provides the values of log ratio of signal intensity. Dye-normalized, background-subtracted log-ratios of sample to reference expression were calculated.

After scanning, row data are analyzed with the Feature Extraction software which gives as output the expression value and a quality report. This file shows if the array was hybridized correctly, if the signal intensity is consistent, and calculate the variability of probes replications. The simultaneous amplification of Spike In RNA is used as internal control.

For further detail on the hybridization protocol, refer to the producer's website [93]

6 Statistical analysis

To perform joint analysis of both clinical and gene expression data, different types of statistical approaches were used. This section describes the analysis plan, dividing the approach on the basis of the primary type of statistics used. Statistical approach to analyze clinical data will be described first, and then gene expression profile data handling will be discussed.

The main aim of data analysis has been to verify the presence of associations between clinical factors (considered either continuously or after partition in classes) and gene expression data using the whole gene expression data set, without pre-filtering or data reduction strategies.

Gene expression data were analyzed using two different levels of statistics: a first level statistic to allow the ranking of the expression values in relation to the clinical variable of interest, and then a second level statistic to provide results of enriched groups or pathways of genes accordingly to the rank.

6.1 Clinical statistics

Descriptive statistics

Distributions of clinical parameters were tested for normality using Kolmogorov-Smirnov test. Non-gaussian variables were rescaled to base 10 logarithm to produce normality. Descriptive statistics are reported in tables as mean ± SD, median and inter-quartile range (IQR), or count and percent, according to their distribution.

Inferential statistics

Between-group comparisons were performed on normal/ normalized variables using Student's t-test and on non-continuous variables using Chi-square test. Presence of trends among more than two groups (i.e. LDL tertiles) was assessed using Analysis of Variance (ANOVA) test with linear polynomial contrast.

The probability (p) value to reject null-hypothesis for clinical parameters was set at 0.05 with a two-tail distribution. No p-value correction for multiple comparisons was used for to compare clinical variables in order to ensure a high level of sensitivity.

Multivariate models for variable adjustment

To identify a set of clinical predictors of a dependent clinical variable, Pearson's correlation analysis was used. Clinical parameters correlated with the dependent variable with a p-value lower than 0.1 were considered as possible independent variables to be entrerd into the multivariate model.

Multiple linear regression (MLR) analysis was performed to identify the independent predictors and estimate their relative impact. Age, gender, and all experimental measurements significantly associated with the dependent variable as described before, were entrered into the model as independent variables. In presence of several independent variables identified, analysis was performed with multi-step backward elimination (entry probability=0.05; removal= 0.15). Based on the final MLR model, values of the dependent variables were adjusted for the independet predictors in a subsequent MLR model.

Fulfillment of the assumptions form MLR was checked by residual analysis and distribution/ correlation graphs.

Statistical analysis for the clinical parameters was performed using software package SPSS 17.0 for Windows (SPSS Inc., Chicago, IL, US).

6.2 Gene expression statistics

Filtering and normalization of microarray data

Probes with >50% missing values were excluded from the analysis. Each experiment was standardized to have mean 0 and standard deviation of 1. Finally, probes annotated as targeting the same gene were combined by taking their median value of expression in each experiment.

Differential expression

One of the basic tasks in gene expression data analysis is finding differentially expressed genes between 2 classes (such as tumor vs. normal or diabetics vs. non diabetics). A variety of methods were developed to address this task, such as TNoM (Ben-Dor et al., 2001A), SAM (Tusher et al., 2001) and others (see review of Cui and Churchill, 2003). In common practice bioinformaticians typically use categorical information on the samples to derive partitions. This study includes the investigation of the differential gene expression associated quantitative clinical phenotypes. We use two mechanisms for partitioning the set of samples (Supp Figure DA1): Parametric - Setting two values as thresholds (t and u) – all samples with phenotype value below t will be in class A, and all samples whose chosen phenotype value is above u will be in class B. The rest of the samples (between t and u) will be ignored. Non-parametric - Setting two percentile values as thresholds. Once the partition is determined one can use any measure of differential expression. In this study we used TNoM and Student t-test approaches. Differential expression in quantitative phenotypes is further discussed in 16.

Enrichment analysis, mHG

In our analyses we use methods to detect enrichment of a specific group of set at the top of a ranked list of genes. More formally, consider a universe gene set G = $\{g_i\}_{i=1..N}$, and a set of genes all participating in the same biological process, which we shall denote by T. Consider a candidate binary partition Q=(A, B) of the mRNA expression data and assume that for every transcript g we computed a differential expression score, reflecting under/over expression in A as compared to B, denoted d(g). Rank the transcripts according to d(g), where the most significant transcripts are at the top of the list. We assign a statistical score to the enrichment of T at the top of this list. This enrichment score, e(Q,T,G), associates an activity of T with Q. e(Q,T,G) is computed using the mHG statistics (Eden et al., 2007 and 17). The enrichment procedure can be used with GO terms, TF cohorts, KEGG classes, miRNA target sets (as inferred from databases such as TARGETSCAN), genomic intervals and sets of genes derived from other studies. We use the same methods to statistically assess enrichment in lists ranked according to correlations, as is the case for the number of cigarettes smoked.

7 Results

At the end of the enrollment period, a total of 164 volunteers were identified and data collected. The final study population includes 91 males and 73 females, with a mean age of 37 years. Also in the overall population, enrolled subjects showed a low cardiovascular risk. The following table shows the characteristics of the enrolled population stratified by gender.

Table 4: Clinical characteristics of the study overall population stratified by gender

Variables	Males (N=91)	Females (N=73)
Age [yy] *	37 ± 9	37 ± 7
Smoking habit [current/former/never (%)]	(53.8/20.9/25.3)	(71.2/11.0/17.8)
BMI [kg/m2]	22,7 ± 3,3	25,8 ± 4,0
Waist [cm]	83 ± 8	92 ± 11
Systolic BP [mmHg]	109 ± 13	122 ± 14
Diastolic BP [mmHg]	70 ± 13	80 ± 9
Fasting plasma glucose [mg/dL]	84 ± 6	90 ± 10
2-hour plasma glucose [mg/dL]	95 ± 23	94 ± 31
Total-Cholesterol [mg/dL]	200 ± 29	205 ± 37
HDL-Cholesterol [mg/dL]	67 ± 14	53 ± 13
Triglycerides [mg/dL] *	54 (28)	73 (70)
LDL-Cholesterol [mg/dL]	121 ± 27	134 ± 33
WBC [Count*10^3/mm3]	5,863 ± 1,275	5,977 ± 1,374
hsCRP [mg/L] *	0,74 (1,00)	0,65 (0,82)

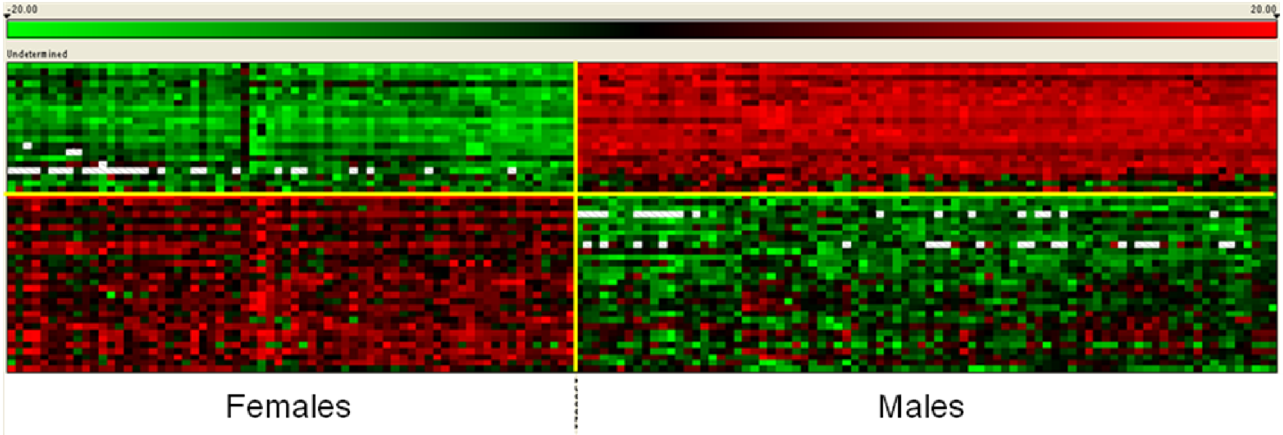
Footnote to the table: Continuous variables are expressed as mean ± SD or median (IQR), if not normally distributed (*); nominal variables (**) are presented as number (percent). BMI = Body Mass Index, BP = Blood Pressure, HDL = High-Density Lipoprotein, LDL = Low-Density Lipoprotein, WBC = White Blood Cells, hsCRP = high sensitivity C-Reactive Protein.

As first analysis in the pivotal dataset, we explored the variability of gene expression profile as a function of the main demographic characteristics: gender, age and adiposity.

7.1 Gender-specific gene expression profile

Subjects were compared for expression values stratified by gender through differential expression analysis using TNoM scoring. As depicted in the following figure, we observed a substantially different expression profile between the two genders. Almost all the top 50 differentially expressed genes were involved in sexual differentiantion or were genes lying on the sexual chromosomes.

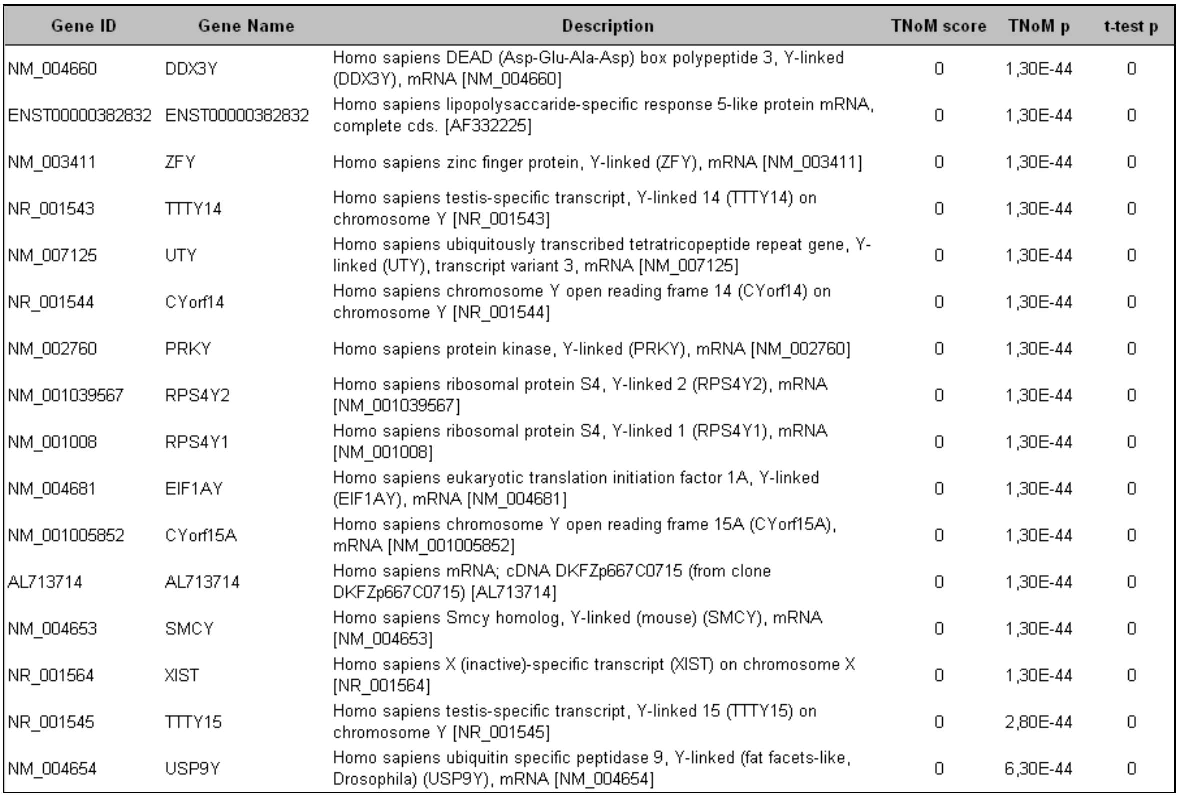
Figure 20: Gene expression profile stratified by gender



Footnote to figure. Heatmap of the 50 most differentially expressed genes between males and females. Each column corresponds to a patient and each row to a gene. Color represents log2 ratio of sample compared to reference. For illustration purposes, only genes with less than 25% of missing values have been represented.

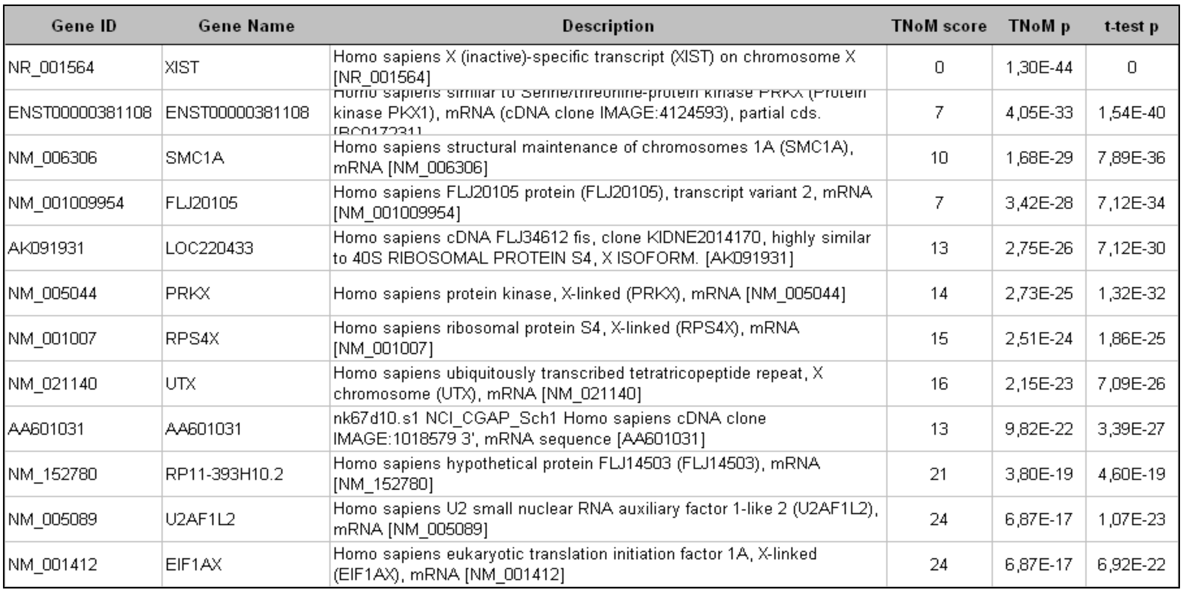
More in details, male subjects displayed –as predictable- the expression of genes located on the Y chromosome that are not present in females, whereas women showed a relevant over-expression of genes located on the X chromosome compared to men. The two following tables depicts the top differentially expressed genes between males and females.

Figure 21: Most over-expressed genes in males compared to females



Gene ID	Gene Name	Description	TNoM score	TNoM p	t-test p
NM_004660	DDX3Y	Homo sapiens DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, Y-linked (DDX3Y), mRNA [NM_004660]	0	1,30E-44	0
ENST00000382832	ENST00000382832	Homo sapiens lipopolysaccaride-specific response 5-like protein mRNA, complete cds. [AF332225]	0	1,30E-44	0
NM_003411	ZFY	Homo sapiens zinc finger protein, Y-linked (ZFY), mRNA [NM_003411]	0	1,30E-44	0
NR_001543	TTTY14	Homo sapiens testis-specific transcript, Y-linked 14 (TTTY14) on chromosome Y [NR_001543]	0	1,30E-44	0
NM_007125	UTY	Homo sapiens ubiquitously transcribed tetratricopeptide repeat gene, Y-linked (UTY), transcript variant 3, mRNA [NM_007125]	0	1,30E-44	0
NR_001544	CYorf14	Homo sapiens chromosome Y open reading frame 14 (CYorf14) on chromosome Y [NR_001544]	0	1,30E-44	0
NM_002760	PRKY	Homo sapiens protein kinase, Y-linked (PRKY), mRNA [NM_002760]	0	1,30E-44	0
NM_001039567	RPS4Y2	Homo sapiens ribosomal protein S4, Y-linked 2 (RPS4Y2), mRNA [NM_001039567]	0	1,30E-44	0
NM_001008	RPS4Y1	Homo sapiens ribosomal protein S4, Y-linked 1 (RPS4Y1), mRNA [NM_001008]	0	1,30E-44	0
NM_004681	EIF1AY	Homo sapiens eukaryotic translation initiation factor 1A, Y-linked (EIF1AY), mRNA [NM_004681]	0	1,30E-44	0
NM_001005852	CYorf15A	Homo sapiens chromosome Y open reading frame 15A (CYorf15A), mRNA [NM_001005852]	0	1,30E-44	0
AL713714	AL713714	Homo sapiens mRNA; cDNA DKFZp667C0715 (from clone DKFZp667C0715) [AL713714]	0	1,30E-44	0
NM_004653	SMCY	Homo sapiens Smcy homolog, Y-linked (mouse) (SMCY), mRNA [NM_004653]	0	1,30E-44	0
NR_001564	XIST	Homo sapiens X (inactive)-specific transcript (XIST) on chromosome X [NR_001564]	0	1,30E-44	0
NR_001545	TTTY15	Homo sapiens testis-specific transcript, Y-linked 15 (TTTY15) on chromosome Y [NR_001545]	0	2,80E-44	0
NM_004654	USP9Y	Homo sapiens ubiquitin specific peptidase 9, Y-linked (fat facets-like, Drosophila) (USP9Y), mRNA [NM_004654]	0	6,30E-44	0

Figure 22: Most over-expressed genes in females compared to males

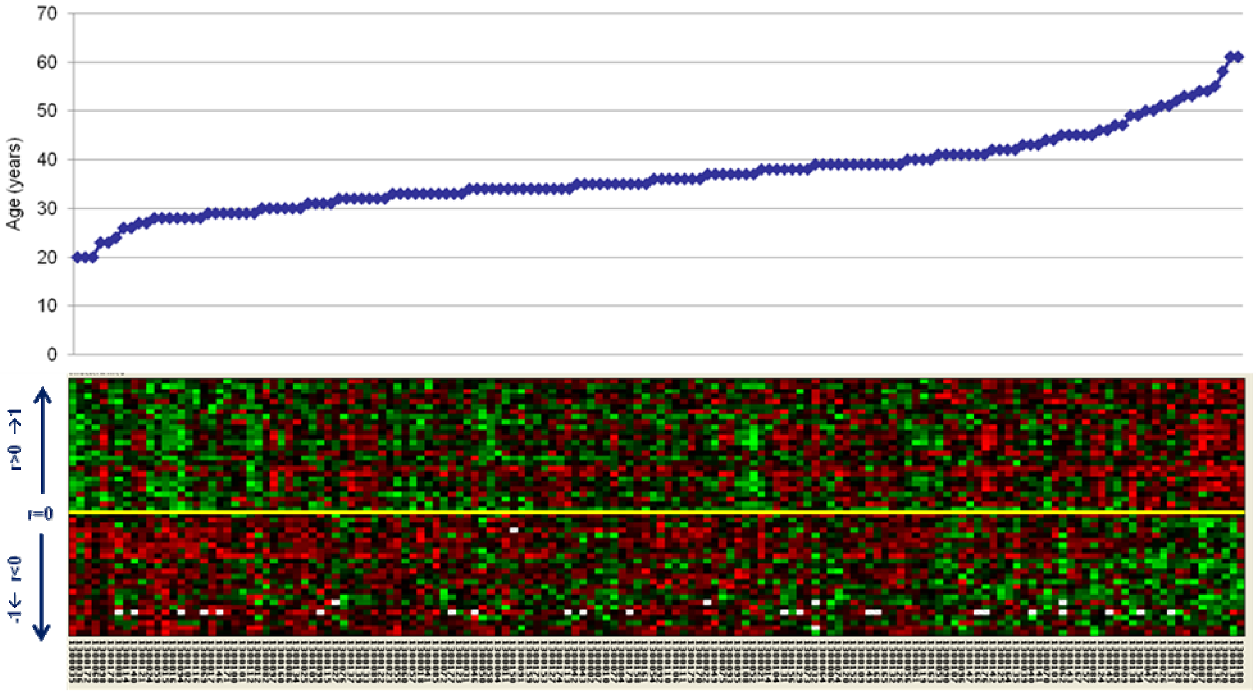


Gene ID	Gene Name	Description	TNoM score	TNoM p	t-test p
NR_001564	XIST	Homo sapiens X (inactive)-specific transcript (XIST) on chromosome X [NR_001564]	0	1,30E-44	0
ENST00000381108	ENST00000381108	Homo sapiens similar to Serine/threonine-protein kinase PRKX (Protein kinase PKX1), mRNA (cDNA clone IMAGE:4124593), partial cds. [BC017231]	7	4,05E-33	1,54E-40
NM_006306	SMC1A	Homo sapiens structural maintenance of chromosomes 1A (SMC1A), mRNA [NM_006306]	10	1,68E-29	7,89E-36
NM_001009954	FLJ20105	Homo sapiens FLJ20105 protein (FLJ20105), transcript variant 2, mRNA [NM_001009954]	7	3,42E-28	7,12E-34
AK091931	LOC220433	Homo sapiens cDNA FLJ34612 fis, clone KIDNE2014170, highly similar to 40S RIBOSOMAL PROTEIN S4, X ISOFORM. [AK091931]	13	2,75E-26	7,12E-30
NM_005044	PRKX	Homo sapiens protein kinase, X-linked (PRKX), mRNA [NM_005044]	14	2,73E-25	1,32E-32
NM_001007	RPS4X	Homo sapiens ribosomal protein S4, X-linked (RPS4X), mRNA [NM_001007]	15	2,51E-24	1,86E-25
NM_021140	UTX	Homo sapiens ubiquitously transcribed tetratricopeptide repeat, X chromosome (UTX), mRNA [NM_021140]	16	2,15E-23	7,09E-26
AA601031	AA601031	nk67d10.s1 NCI_CGAP_Sch1 Homo sapiens cDNA clone IMAGE:1018579 3', mRNA sequence [AA601031]	13	9,82E-22	3,39E-27
NM_152780	RP11-393H10.2	Homo sapiens hypothetical protein FLJ14503 (FLJ14503), mRNA [NM_152780]	21	3,80E-19	4,60E-19
NM_005089	U2AF1L2	Homo sapiens U2 small nuclear RNA auxiliary factor 1-like 2 (U2AF1L2), mRNA [NM_005089]	24	6,87E-17	1,07E-23
NM_001412	EIF1AX	Homo sapiens eukaryotic translation initiation factor 1A, X-linked (EIF1AX), mRNA [NM_001412]	24	6,87E-17	6,92E-22

7.2 Age correlated gene expression profile

To investigate whether age was associated with a specific expression profile, we performed a Pearson's correlation analysis between age and gene expression values. As reported in the figure below, age resulted correlated with a specific gene expression profile.

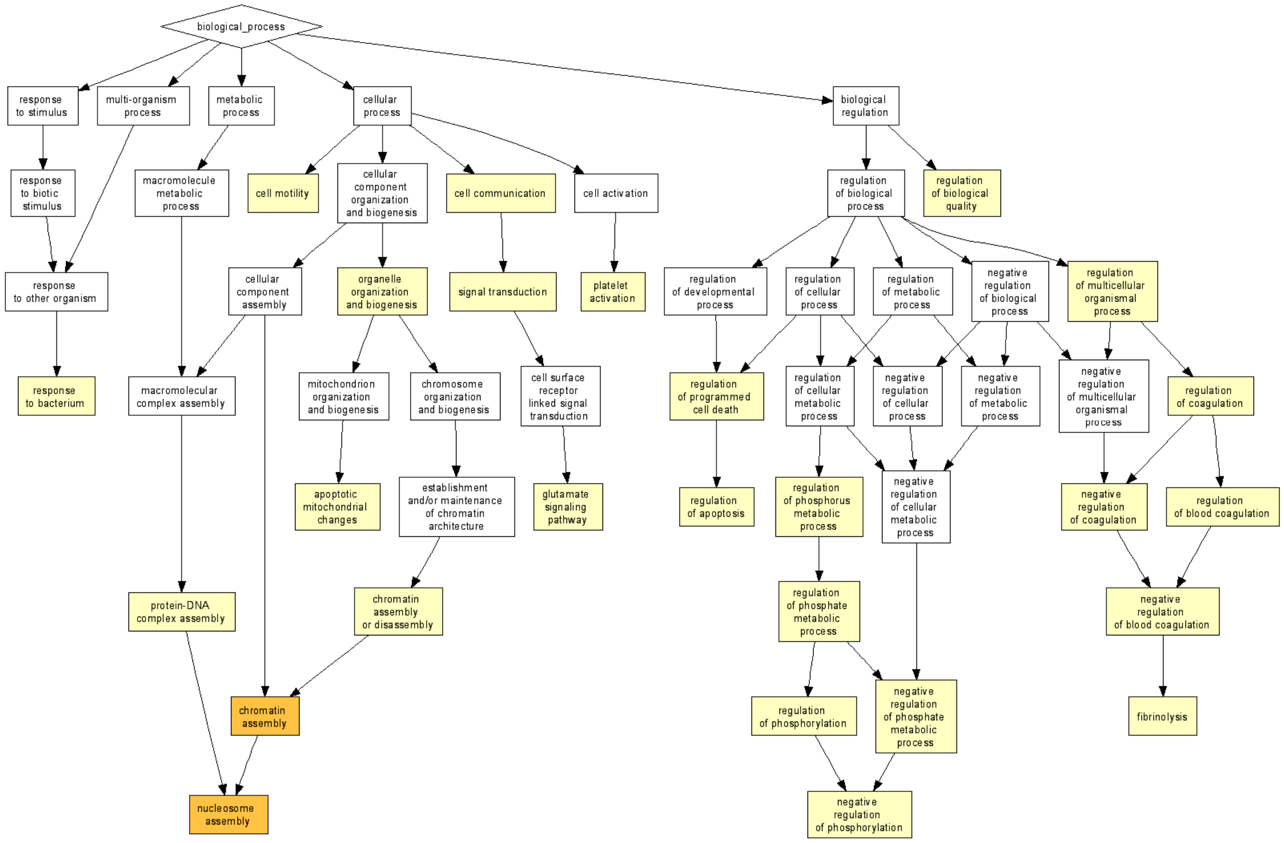
Figure 23: Gene expression profile associated with age



Footnote to figure. Heatmap of the 50 genes most correlated with age. In the upper panel, the age value for each enrolled patient is displayed sorted on the x-axis for age value, from left to right. In the lower panel, each column corresponds to a patient and each row to a gene. Color represents log2 ratio of sample compared to reference. Each patient with his/ her age value reported in the upper panel corresponds to a column in the lower. For illustration purposes, only genes with less than 25% of missing values have been represented.

The age of the subject resulted significantly and positively correlated with the expression of several groups of genes, based on GO analysis. More in detail, the assembly of chromosome and chromatin packaging resulted the most prominent signature over-expressed with increasing age. In addition, older age was associated with a pro-apoptotic signal and to a significant change in expression of groups of genes related to coagulation and fibrinolysis. However, the level of significance of this profile resulted relatively weak, with a p-value around 10^{-4}. The only groups of genes that resulted significantly correlated with increasing age were related to chromosome packing, as shown in the following figure.

Figure 24: Groups of genes significantly over-expressed with increasing age


biological_process
response to stimulus
multi-organism process
metabolic process
cellular process
biological regulation
response to biotic stimulus
macromolecule metabolic process
cell motility
cellular component organization and biogenesis
cell communication
cell activation
regulation of biological process
regulation of biological quality
response to other organism
cellular component assembly
organelle organization and biogenesis
signal transduction
platelet activation
regulation of developmental process
regulation of cellular process
regulation of metabolic process
negative regulation of biological process
regulation of multicellular organismal process
response to bacterium
macromolecular complex assembly
mitochondrion organization and biogenesis
chromosome organization and biogenesis
cell surface receptor linked signal transduction
regulation of programmed cell death
regulation of cellular metabolic process
negative regulation of cellular process
negative regulation of metabolic process
negative regulation of multicellular organismal process
regulation of coagulation
apoptotic mitochondrial changes
establishment and/or maintenance of chromatin architecture
glutamate signaling pathway
regulation of apoptosis
regulation of phosphorus metabolic process
negative regulation of cellular metabolic process
negative regulation of coagulation
regulation of blood coagulation
protein-DNA complex assembly
chromatin assembly or disassembly
regulation of phosphate metabolic process
negative regulation of blood coagulation
chromatin assembly
regulation of phosphorylation
negative regulation of phosphate metabolic process
fibrinolysis
nucleosome assembly
negative regulation of phosphorylation

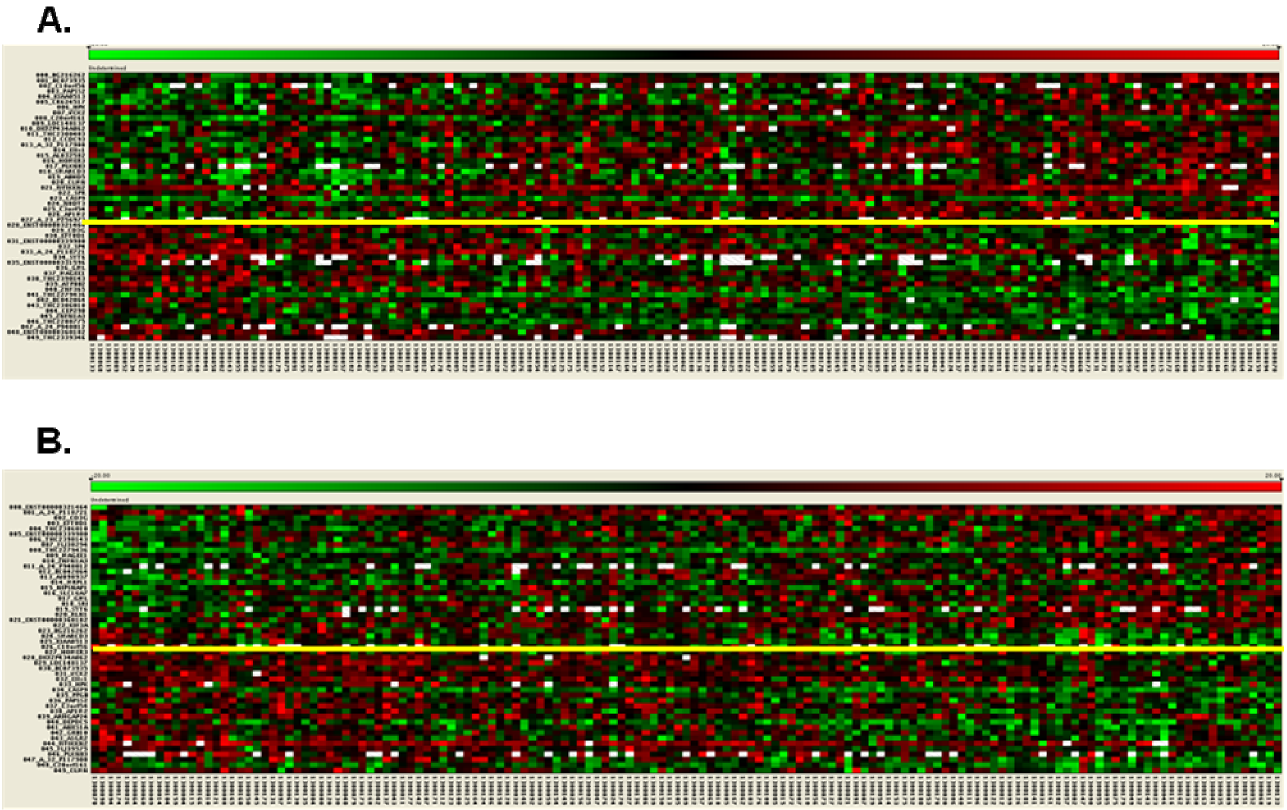
Footnote to figure. Figure shows the groups of genes significantly over-expressed by GO analysis. The panel depicts a graphical representation of the portion of the GO hierarchy that includes the terms significantly over-represented in the list of genes over-expressed in current smokers. Color is a function of the p-value of GO term enrichment by mHG analysis.

7.3 Gene expression profile associated with white blood cell populations

To verify whether the expression profile in circulating PBMCs could reflect the overall inflammatory state as determined by white blood cell sub-population count, we investigate the relationship between gene expression in PBMCs and the percentage of both neutrophiles and lymphomonocytes in the peripheral blood. Since gene expression microarrays are performed with a fixed starting amount of PBMC's RNA, the presence of genes up- or down-regulated as a function of the relative concentration of leukocyte sub-populations appears to be dependent only upon the systemic environment determined by this proportion and not by the confounding presence of a different quantity of RNA from PBMCs.

As shown in the following picture, a gene expression profile was identified both in relation to the relative concentration of neutrophiles and PBMCs.

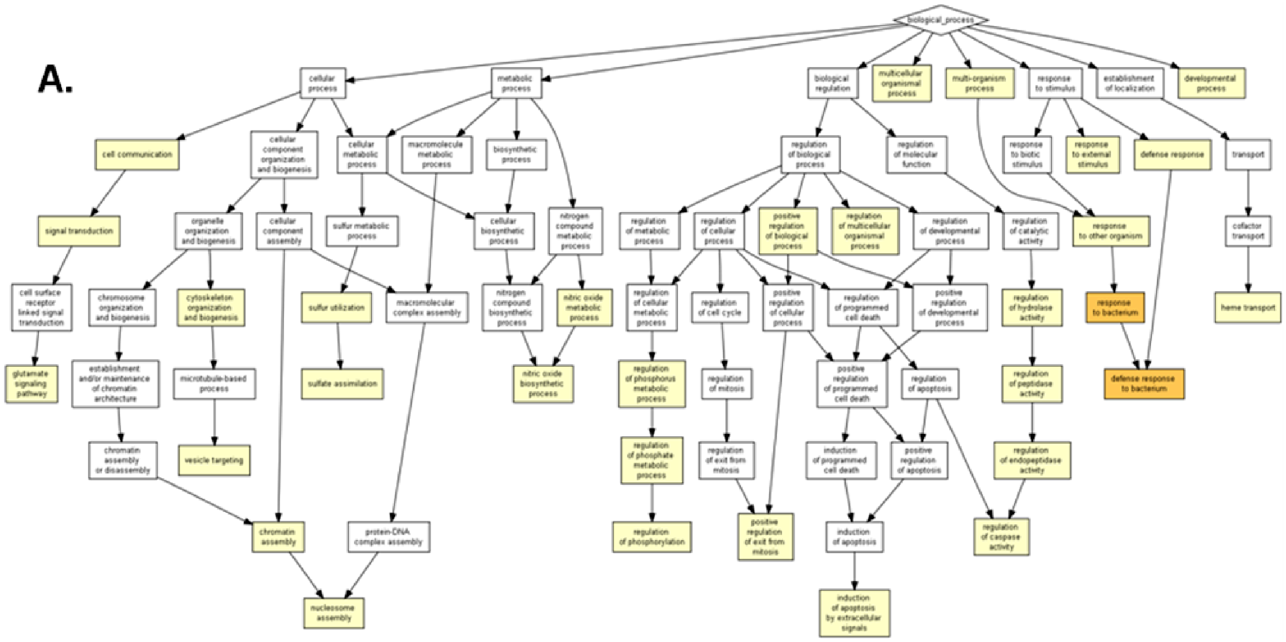
Figure 25: Gene expression profile associated with white blood cell sub-populations



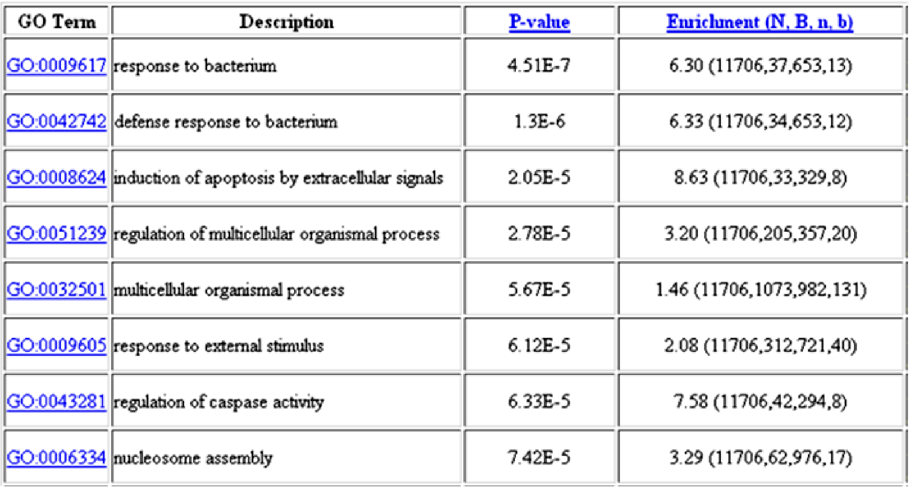
Footnote to figure. Heatmap of the 50 genes most correlated with leukocytes sub-populations. Each column corresponds to a patient and each row to a gene. The relative proprotion of the population is displayed sorted on the x-axis, increasing from left to right. Color represents log2 ratio of sample compared to reference. For illustration purposes, only genes with less than 25% of missing values have been represented. Panel A shows gene expression profile in relation to neutrophyle count, whereas panel B shows the gene expression profile assiciated with mononuclear cell cpount.

As shown in the following picture, the presence of a relatively higher neutrophile count (in percentage of total white blood cells as well as in absolute count values) is associated with the over-expression of genes involved in the response to bacteric infections such as “response to bacterium” and “defense response to bacterium” (p-value of $4.5*10^{-7}$ and $1.3*10^{-6}$ respectively). In addition, genes more associated with the response to virus and parasites such as genes involved in the MHC class I and II defense mechanisms resulted down-regulated in subjects with relatively higher neutrophile concentraion [data not shown].

Figure 26: Groups of genes signficantly over-expressed in case of increasing neutrophile count


A.
biological_process
cellular process
metabolic process
biological regulation
multicellular organismal process
multi-organism process
response to stimulus
establishment of localization
developmental process
cell communication
cellular component organization and biogenesis
cellular metabolic process
macromolecule metabolic process
biosynthetic process
regulation of biological process
regulation of molecular function
response to biotic stimulus
response to external stimulus
defense response
transport
signal transduction
organelle organization and biogenesis
cellular component assembly
sulfur metabolic process
cellular biosynthetic process
nitrogen compound metabolic process
regulation of metabolic process
regulation of cellular process
positive regulation of biological process
regulation of multicellular organismal process
regulation of developmental process
regulation of catalytic activity
response to other organism
cofactor transport
cell surface receptor linked signal transduction
chromosome organization and biogenesis
cytoskeleton organization and biogenesis
sulfur utilization
macromolecular complex assembly
nitrogen compound biosynthetic process
nitric oxide metabolic process
regulation of cellular metabolic process
regulation of cell cycle
positive regulation of cellular process
regulation of programmed cell death
positive regulation of developmental process
regulation of hydrolase activity
response to bacterium
heme transport
glutamate signaling pathway
establishment and/or maintenance of chromatin architecture
microtubule-based process
sulfate assimilation
nitric oxide biosynthetic process
regulation of phosphorus metabolic process
regulation of mitosis
positive regulation of programmed cell death
regulation of apoptosis
regulation of peptidase activity
defense response to bacterium
chromatin assembly or disassembly
vesicle targeting
regulation of phosphate metabolic process
regulation of exit from mitosis
induction of programmed cell death
positive regulation of apoptosis
regulation of endopeptidase activity
chromatin assembly
protein-DNA complex assembly
regulation of phosphorylation
positive regulation of exit from mitosis
induction of apoptosis
regulation of caspase activity
nucleosome assembly
induction of apoptosis by extracellular signals

B.

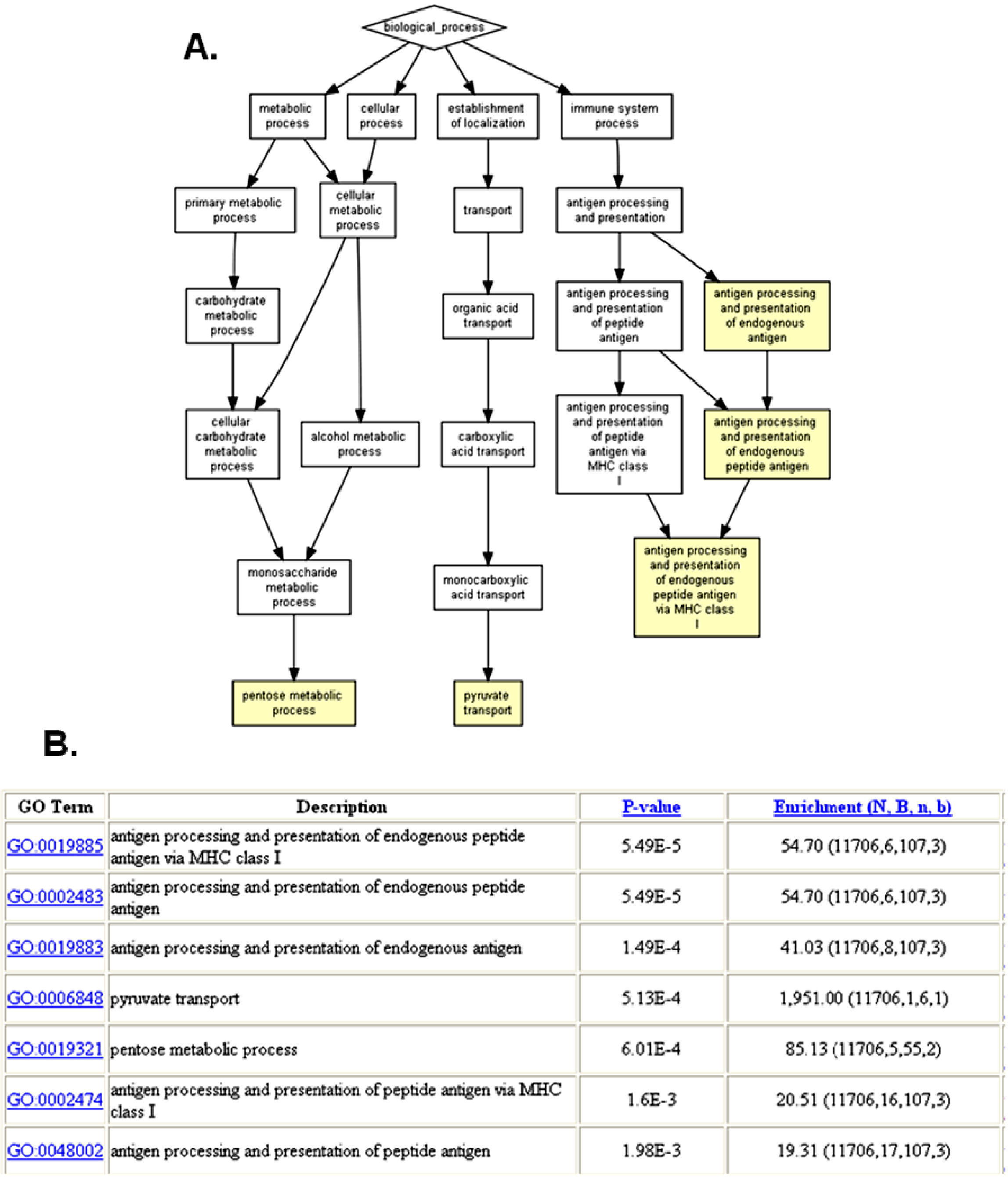


GO Term	Description	P-value	Enrichment (N, B, n, b)
GO:0009617	response to bacterium	4.51E-7	6.30 (11706,37,653,13)
GO:0042742	defense response to bacterium	1.3E-6	6.33 (11706,34,653,12)
GO:0008624	induction of apoptosis by extracellular signals	2.05E-5	8.63 (11706,33,329,8)
GO:0051239	regulation of multicellular organismal process	2.78E-5	3.20 (11706,205,357,20)
GO:0032501	multicellular organismal process	5.67E-5	1.46 (11706,1073,982,131)
GO:0009605	response to external stimulus	6.12E-5	2.08 (11706,312,721,40)
GO:0043281	regulation of caspase activity	6.33E-5	7.58 (11706,42,294,8)
GO:0006334	nucleosome assembly	7.42E-5	3.29 (11706,62,976,17)

Footnote to figure. Figure shows the groups of genes significantly over-expressed by GO analysis. The upper panel depicts a graphical representation of the portion of the GO hierarchy that includes the terms significantly over-represented in the list of genes over-expressed in current smokers. Color is a function of the p-value of GO term enrichment by mHG analysis and is coded following the scale depicted in the previous figures. Lower panel: GO terms significantly enriched in the list of genes over-expressed in current smokers. Enrichment is defined as (b/n) / (B/N), where N: total number of genes, B: total number of genes associated with a specific GO term, n: number of genes in the 'target set', and b: number of genes in the 'target set' associated with a specific GO term. The reported p-values are not corrected for multiple testing.

Conversely, the presence of relatively higher proportion of lymphomonocytes in the circuylating blood compared to neutrophiles, as assessed by leukocyte sub-population count, resulted associated with an increased expression of genes more associated to viral response and anatigen presentation, like MHC class I genes (following picture), whereas genes related to the response to bacterial infections resulted decreased [data not shown].

Figure 27: Groups of genes signficantly over-expressed in case of increasing lymphomonocyte count


A.
biological_process
metabolic process
cellular process
establishment of localization
immune system process
primary metabolic process
cellular metabolic process
transport
antigen processing and presentation
carbohydrate metabolic process
organic acid transport
antigen processing and presentation of peptide antigen
antigen processing and presentation of endogenous antigen
cellular carbohydrate metabolic process
alcohol metabolic process
carboxylic acid transport
antigen processing and presentation of peptide antigen via MHC class I
antigen processing and presentation of endogenous peptide antigen
monosaccharide metabolic process
monocarboxylic acid transport
antigen processing and presentation of endogenous peptide antigen via MHC class I
pentose metabolic process
pyruvate transport
B.

GO Term	Description	P-value	Enrichment (N, B, n, b)
GO:0019885	antigen processing and presentation of endogenous peptide antigen via MHC class I	5.49E-5	54.70 (11706,6,107,3)
GO:0002483	antigen processing and presentation of endogenous peptide antigen	5.49E-5	54.70 (11706,6,107,3)
GO:0019883	antigen processing and presentation of endogenous antigen	1.49E-4	41.03 (11706,8,107,3)
GO:0006848	pyruvate transport	5.13E-4	1,951.00 (11706,1,6,1)
GO:0019321	pentose metabolic process	6.01E-4	85.13 (11706,5,55,2)
GO:0002474	antigen processing and presentation of peptide antigen via MHC class I	1.6E-3	20.51 (11706,16,107,3)
GO:0048002	antigen processing and presentation of peptide antigen	1.98E-3	19.31 (11706,17,107,3)

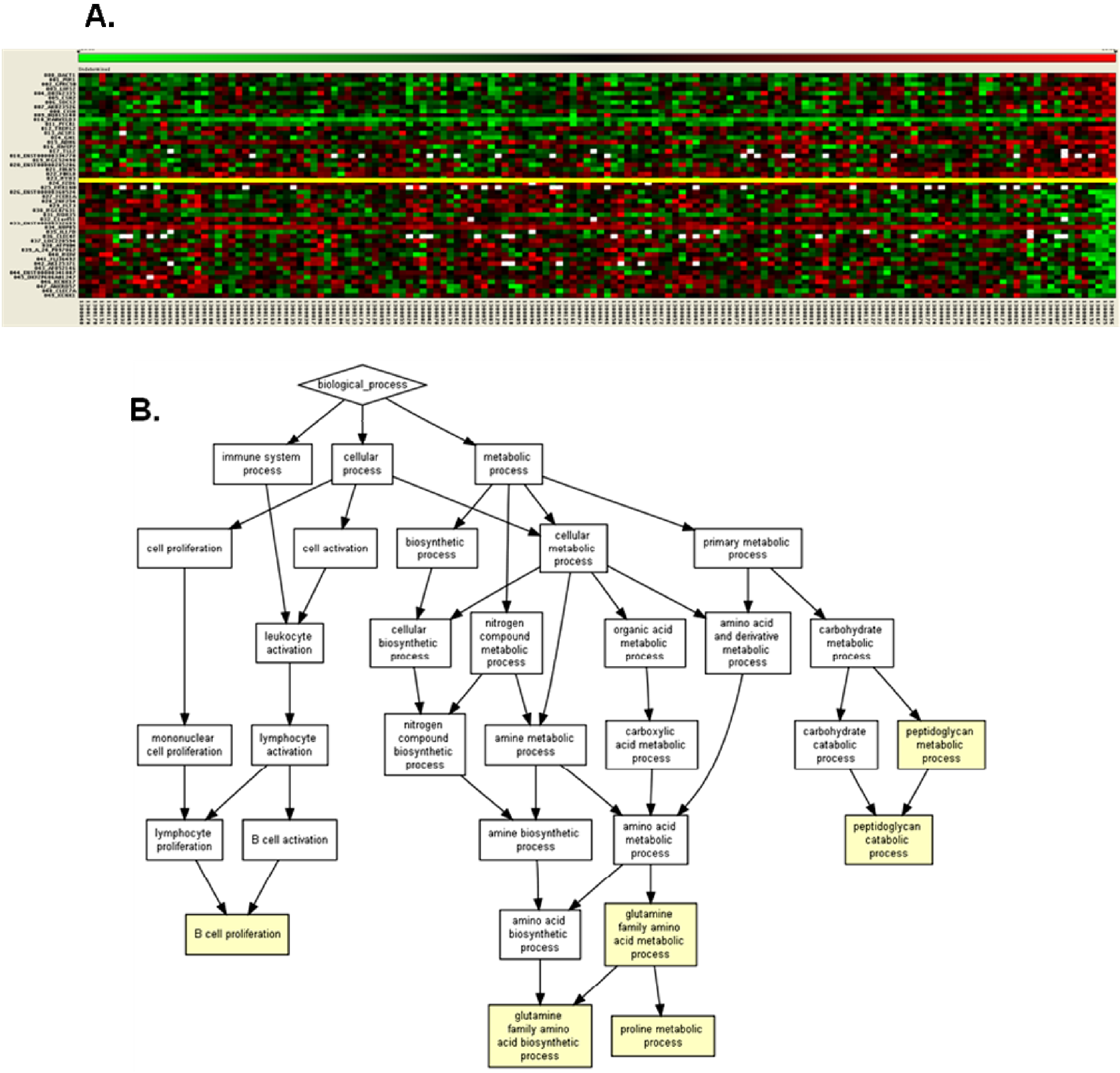
Footnote to figure. Figure shows the groups of genes significantly over-expressed by GO analysis. The upper panel depicts a graphical representation of the portion of the GO hierarchy that includes the terms significantly over-represented in the list of genes over-expressed in current smokers. Color is a function of the p-value of GO term enrichment by mHG analysis and is coded following the scale depicted in the previous figures. Lower panel: GO terms significantly enriched in *the list of genes over-expressed in current smokers. Enrichment is defined as (b/n) / (B/N), where N: total number of genes, B: total number of genes associated with a specific GO term, n: number of genes in the 'target set', and b: number of genes in the 'target set' associated with a specific GO term. The reported p-values are not corrected for multiple testing.*

7.4 Gene expression profile associated with emerging inflammatory markers

The correlation analysis of gene expression profile with inflammatory cells and CVD risk factors illustrated so far support the hypothesis that expression profiling in circulating PBMCs is a suitable technique to identify the presence of the sub-clinical inflammation associated with CV disease. Currently, several candidate plasma markers are investigated with a similar approach. We therefore decided to investigate whether emerging inflammatory markers whose increase has been associated with CV risk are significantly correlated with apro-inflammatory gene expression profile in circulating PBMCs. To pursue this aim, we performed a Pearson correlation analysis of gene expression values with circulating levels of hsCRP and TNF-alpha.

The following picture shows the results of the correlation analysis between hsCRP and gene expression in PBMCs. As it appears, higher hsCRP concentrations resulted associated with a significant increase in several group of genes, some of which related to some extent with leucocyte inflammatory response. However, the biological signal appears to be weak (maximum p-value level around 10^{-4}) and without a strong internal consistency among gene groups. In addition, no specific KEGG pathway resulted significantly associated with hsCRP levels.

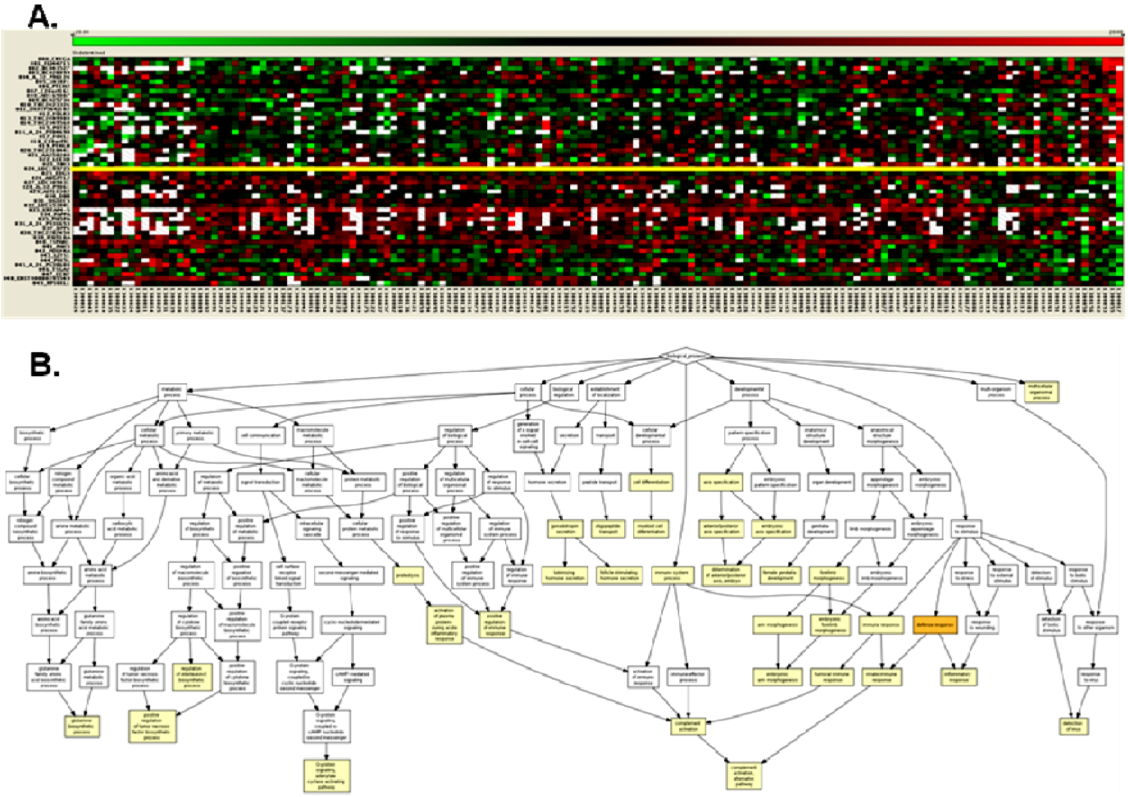
Figure 28: Gene expression profile associated with plasma hsCRP concentrations

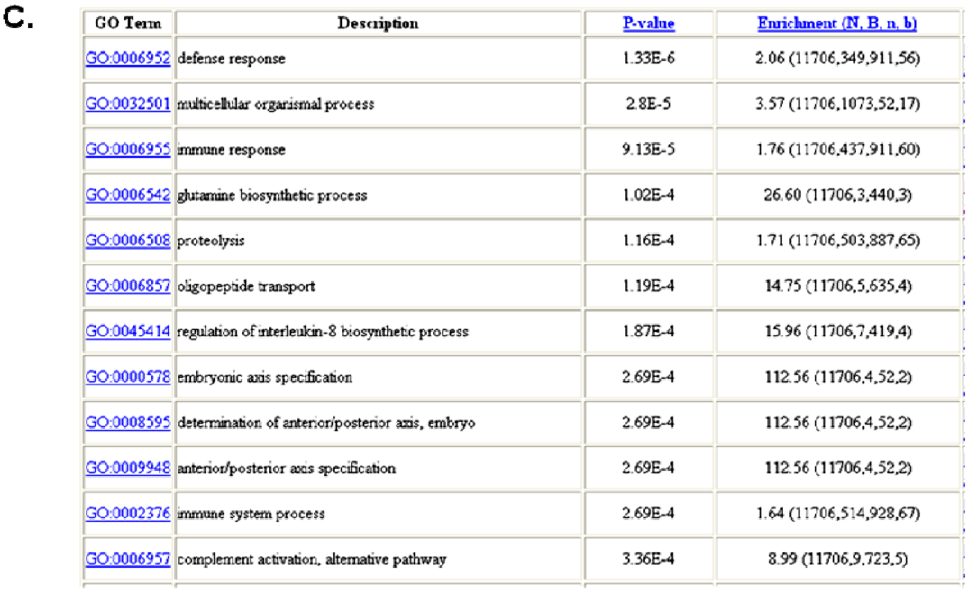


Footnote to figure. Panel A shows the heatmap of the 50 genes most correlated with circulating hsCRP. Each column corresponds to a patient and each row to a gene. Color represents log2 ratio of sample compared to reference. For illustration purposes, only genes with less than 25% of missing values have been represented. Panel B shows a graphical representation of the portion of the GO hierarchy that includes the terms significantly over-represented in the list of genes over-expressed in correlation with hsCRP. Color is a function of the p-value of GO term enrichment by mHG analysis and is coded following the scale depicted in the previous figures.

On the contrary, the circulating concentration of TNF-alpha resulted significantly associated with a gene expression profile characterized by a pro-inflammatory signature, as demonstrated by the significant over-expression of gene groups such as “Defense response” (p= $1.33*10^{-6}$), “Immune response" (p= $9.13*10^{-5}$), and "Complement activation" (p= $3.36*10^{-4}$) in subjects with relatively high TNF-alpha levels.The following picture shows the gene expression profile and the groups of genes significantly correlated with circulating TNF-alpha concentrations.

Figure 29: Gene expression profile associated with circulating TNF-alpha concentrations





GO Term	Description	P-value	Enrichment (N, B, n, b)
GO:0006952	defense response	1.33E-6	2.06 (11706,349,911,56)
GO:0032501	multicellular organismal process	2.8E-5	3.57 (11706,1073,52,17)
GO:0006955	immune response	9.13E-5	1.76 (11706,437,911,60)
GO:0006542	glutamine biosynthetic process	1.02E-4	26.60 (11706,3,440,3)
GO:0006508	proteolysis	1.16E-4	1.71 (11706,503,887,65)
GO:0006857	oligopeptide transport	1.19E-4	14.75 (11706,5,635,4)
GO:0045414	regulation of interleukin-8 biosynthetic process	1.87E-4	15.96 (11706,7,419,4)
GO:0000578	embryonic axis specification	2.69E-4	112.56 (11706,4,52,2)
GO:0008595	determination of anterior/posterior axis, embryo	2.69E-4	112.56 (11706,4,52,2)
GO:0009948	anterior/posterior axis specification	2.69E-4	112.56 (11706,4,52,2)
GO:0002376	immune system process	2.69E-4	1.64 (11706,514,928,67)
GO:0006957	complement activation, alternative pathway	3.36E-4	8.99 (11706,9,723,5)

Footnote to figure. Panel A shows the heatmap of the 50 genes most correlated with circulating TNF-alpha. Each column corresponds to a patient and each row to a gene. Color represents log2 ratio of sample compared to reference. For illustration purposes, only genes with less than 25% of missing values have been represented. Panel B shows a graphical *representation of the portion of the GO hierarchy that includes the terms significantly over-represented in the list of genes over-expressed in correlation with TNF-alpha. Color is a function of the p-value of GO term enrichment by mHG analysis and is coded following the scale depicted in the previous figures. Panel C enlists in a tabular form the GO terms significantly enriched in the list of genes over-expressed in presence of higher TNF-alpha concentrations. Enrichment is defined as (b/n) / (B/N), where N: total number of genes, B: total number of genes associated with a specific GO term, n: number of genes in the 'target set', and b: number of genes in the 'target set' associated with a specific GO term. The reported p-values are not corrected for multiple testing.*

7.5 Gene expression profile in relation to classical CVD risk factors

As already detailed, we first conducted a proof-of-concept study on the first 50 enrolled patients to identify whether classical CVD risk factors that have already been associated with changes in gene expression profile in previous preliminary studies form other groups were correlated with relevant changes in gene expression also in our population. Out of the 50 patients enrolled, the RNA extracted from two PBMCs specimens resulted of insufficient quality to be hybridized, and were consequently excluded. A total of 48 patients were therefore included in this analysis. The presence of a correlation between gene expression and clinical characteristics of the population was checked using Pearson correlation analysis. The following clinical parameters were included into this bivariate exploratory analysis: smoking status, number of smoked cigarettes per week, LDL-cholesterol concentrations, systolic blood pressure, diastolic blood pressure, and fasting plasma glucose concentrations. Out of this analysis we identified relevant gene expression profiles in association with smoking status (and amount of cigarettes smoked) and LDL-cholesterol concentrations.

7.5.1 General description of the pilot population

As shown in the next table, the 48 selected volunteers included both males (N=20) and females (N=28), aged between 20 and 64 years (median 35, IQR 7). All subjects had a low cardiovascular risk as estimated by Framingham risk score[94], namely lower than 10% at 10 years. In addition, incidence of classical CVD risk factors in the cohort was fairly low: only 8 subjects had blood pressure above 140/90 mmHg, none had impaired fasting glucose (as defined by fasting plasma glucose concentration of 100 mg/dL or more), and only 3 were glucose intolerant during OGTT (maximum 2-hour glucose value was 153 mg/dL). In addition, only 5 volunteers had LDL-cholesterol concentrations above 160 mg/dL and only 8 were current cigarette smokers with a median weekly consumption of 65 cigarettes, whereas 28 had never smoked neither continuously nor occasionally.

Table 5: Clinical characteristics of the pilot population stratified by gender

Variables	Males (N=20)	Females (N=28)	P
Age [yy] *	34 (4)	36 (10)	NS
Smoking habit [never/former/current (%)] **	12/5/2 (63.2/26.3/10.5)	19/3/6 (67.9/10.7/21.4)	NS
BMI [kg/m2]	25.5 ± 2.7	22.5 ± 3.0	0.001
Waist [cm]	90 ± 9	82 ± 8	0.001
Systolic BP [mmHg]	124 ± 11	106 ± 12	< 0.001
Diastolic BP [mmHg]	82 ± 11	69 ± 10	< 0.001
Fasting plasma glucose [mg/dL]	90 ± 5	85 ± 6	0.003
2-hour plasma glucose [mg/dL]	100 ± 27	93 ± 20	NS
Total-Cholesterol [mg/dL]	208 ± 43	199 ± 31	NS
HDL-Cholesterol [mg/dL]	56 ± 15	71 ± 11	< 0.001
Triglycerides [mg/dL] *	65 (120)	60 (35)	NS
LDL-Cholesterol [mg/dL]	130 ± 36	114 ± 31	0.036
WBC [Count*10^3/mm3]	5.883 ± 0.997	6.009 ± 1.403	NS
hsCRP [mg/L] *	0.65 (1.47)	0.67 (0.99)	NS

Footnote to the table: Continuous variables are expressed as mean ± SD or median (IQR), if not normally distributed (*); nominal variables (**) are presented as number (percent). Significance of difference between-groups has been evaluated using t-test or Chi-square (**) as appropriate. Not normally distributed variables (*) were log-transformed before t-test. BMI = Body Mass Index, BP = Blood Pressure, HDL = High-Density Lipoprotein, LDL = Low-Density Lipoprotein, WBC = White Blood Cells, hsCRP = high sensitivity C-Reactive Protein.

7.5.2 Pro inflammatory expression profile in PBMCs in smokers

The next table compares clinical variables between current smokers and subjects who never smoked. As it appears, the two groups did not differ in terms of age, gender distribution or cardiovascular risk factors, including inflammatory markers such as leukocyte count and hs-CRP.

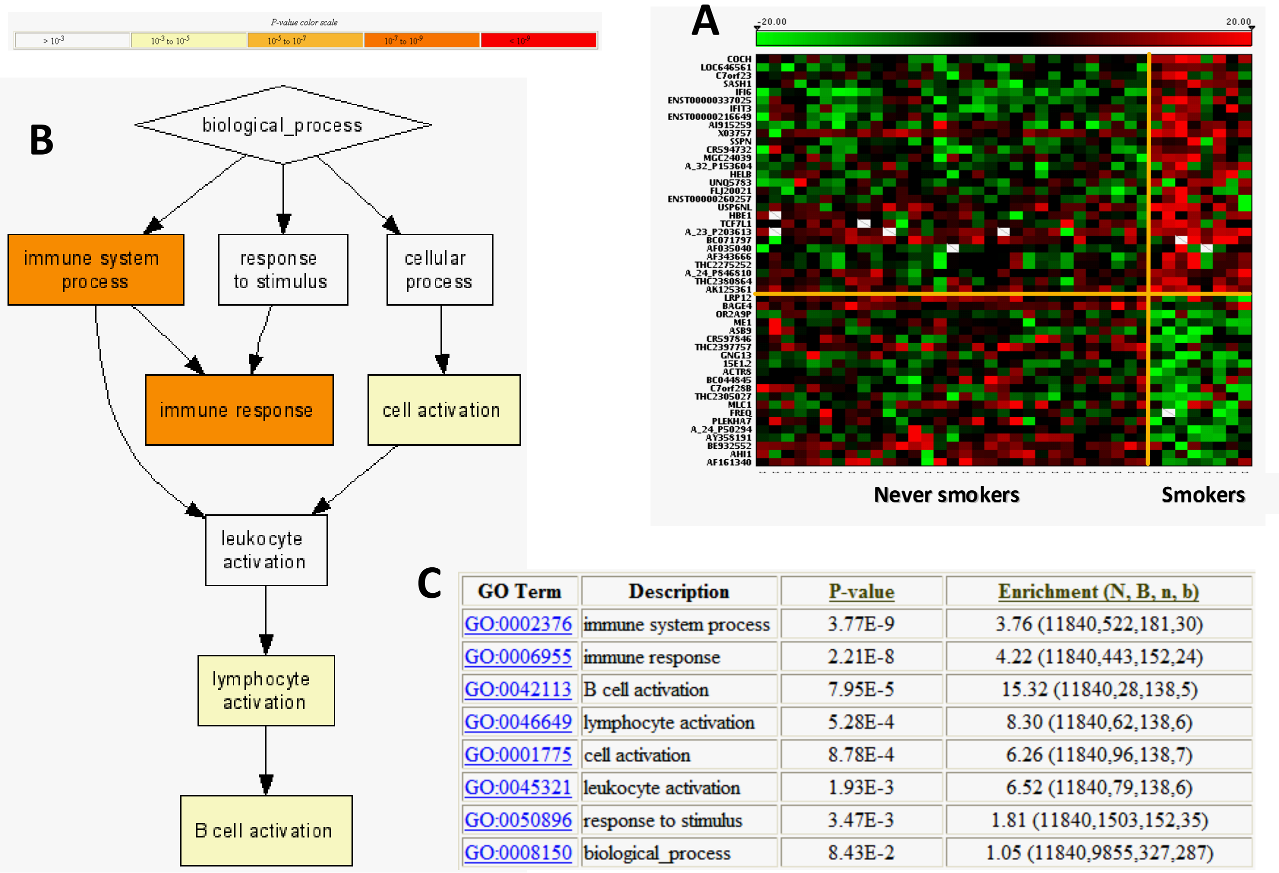
Table 6: Clinical characteristics of current smokers compared to non smokers

Variables	Never Smokers (N=31)	Current Smokers (N=8)	p
Gender [M/F (%)] **	12/19 (38.7/61.3)	2/6 (25.0/75.0)	NS
Age [yy]	35 ± 7	41 ± 10	NS
Smoking habit [cigarettes/week]	-	65 (63)	-
Smoking history [years]	-	14.5 (11.0)	-
BMI [kg/m2]	23.1 ± 2.5	23.9 ± 4.9	NS
Waist [cm]	84 ± 9	85 ± 11	NS
Systolic BP [mmHg]	112 ± 13	110 ± 18	NS
Diastolic BP [mmHg]	71 ± 17	73 ± 15	NS
Fasting plasma glucose [mg/dL]	87 ± 5	87 ± 8	NS
2-hour plasma glucose [mg/dL]	93 ± 24	86 ± 14	NS
Total-Cholesterol [mg/dL]	195 ± 37	209 ± 29	NS
HDL-Cholesterol [mg/dL]	65 ± 14	64 ± 13	NS
Triglycerides [mg/dL] *	59 (39)	63 (25)	NS
LDL-Cholesterol [mg/dL]	115 ± 35	133 ± 26	NS
WBC [Count*10^3/mm3]	5.81 ± 1.25	6.04 ± 1.92	NS
hsCRP [mg/L] *	0.64 (0.71)	0.17 (1.34)	NS

Footnote to table: Continuous variables are expressed as mean ± SD or median (IQR), if not normally distributed (*); nominal variables (**) are represented as number (percent). Significance of between-group difference has been evaluated using t-test or Chi-square (**) as appropriate. Not normally distributed variables (*) were log-transformed before performing t-test. CAD = Coronary Artery Disease, CVD = Cardio-Vascular Diseases, BMI = Body Mass Index, BP = Blood Pressure, HDL = High-Density Lipoprotein, LDL = Low-Density Lipoprotein, WBC = White Blood Cells, hsCRP = high sensitivity C-Reactive Protein.

Despite this similar clinical portrait, smokers showed a different gene expression profile, mainly characterized by a significant over-expression of genes related to generic GO terms of inflammation, like “immune system process” ($p = 3.8 * 10^{-9}$), “immune response“ ($p = 2.2 * 10^{-8}$) and “B cell activation” ($p = 7.9 * 10^{-5}$), as shown in this figure.

Figure 30: Gene expression profile in current cigarette smokers compared to subjects who have never smoked.



GO Term	Description	P-value	Enrichment (N, B, n, b)
GO:0002376	immune system process	3.77E-9	3.76 (11840,522,181,30)
GO:0006955	immune response	2.21E-8	4.22 (11840,443,152,24)
GO:0042113	B cell activation	7.95E-5	15.32 (11840,28,138,5)
GO:0046649	lymphocyte activation	5.28E-4	8.30 (11840,62,138,6)
GO:0001775	cell activation	8.78E-4	6.26 (11840,96,138,7)
GO:0045321	leukocyte activation	1.93E-3	6.52 (11840,79,138,6)
GO:0050896	response to stimulus	3.47E-3	1.81 (11840,1503,152,35)
GO:0008150	biological_process	8.43E-2	1.05 (11840,9855,327,287)

Footnote to figure: Panel A: Heatmap of the 50 most differentially expressed genes between current and never smokers. Each column corresponds to a patient and each row to a gene. Color represents log2 ratio of sample compared to reference (see Methods). For illustration purposes, only genes with less than 25% of missing values have been represented. However, the full list of genes ranked according to their TNoM and t-test p-value (see methods) is available as supplementary material. Panel B: Graphical representation of the portion of the GO hierarchy that includes the terms significantly over-represented in the list of genes over-expressed in current smokers. Color is a function of the p-value of GO term enrichment by mHG analysis and is coded following the scale depicted on top of the panel. Panel C: GO terms significantly enriched in the list of genes over-expressed in current smokers. Enrichment is defined as (b/n) / (B/N), where N: total number of genes, B: total number of genes associated with a specific GO term, n: number of genes in the 'target set', and b: number of genes in the 'target set' associated with a specific GO term. The reported p-values are not corrected for multiple testing.

Amongst the genes differentially expressed in smokers, we observed a significant up-regulation of several genes related to the response to interferon, like interferon alpha-inducible protein 6 (IFI6 [NM_022873], TNoM p=0.0027) and interferon-induced protein with tetratricopeptide repeats 3 (IFIT3 [NM_001549], TNoM p= 0.0027). In addition, smokers were also characterized by a significant over-expression of beta-chorionic gonadotropin (□ -hCG [NM_000737], TNoM p=0.0014), a marker of germ cell neoplasms[95]; sarcospan (SSPN [NM_005086], TNoM p= 0.0027), a gene inducing cell apoptosys in response to hypoxia[96]; and white blood cell-specific transcription factors such as the T-cell specific transcription factor 7-like 1 (TCF7L1 [NM_031283], TNoM p= 0.0030). The presence of an overall up-regulation of genes involved in the innate immunity in current smokers is also confirmed by a significant enrichment of several expression profiles previously observed in response to interferon and published in the Molecular Signatures Database [97] [data not shown].

To identify molecular pathways selectively enriched in current smokers, we searched the KEGG database and we observed a significant up-regulation ($p<0.001$) of several pathways associated with immune response and cancer (next table), as well as the down-regulation of the PPAR-g pathway (p= 0.002).

Table 7: Molecular pathways of the KEGG database enriched in genes over-expressed in current smokers compared to non-smokers

Term Name	Description	P-value	Enrichment (N, B, n, b)	Genes
04662	B cell receptor signaling pathway	9.22E-5	9.12 (15598,64,187,7)	CD22 - cd22 molecule FCGR2B - fc fragment of igg, low affinity iib, receptor (cd32) RASGRP3 - ras guanyl releasing protein 3 (calcium and dag-regulated) BLNK - b-cell linker PIK3CD - phosphoinositide-3-kinase, catalytic, delta polypeptide CD79A - cd79a molecule, immunoglobulin-associated alpha CD79B - cd79b molecule, immunoglobulin-associated beta
04080	Neuroactive ligand-receptor interaction	1.01E-4	2.62 (15598,160,855,23)	ADRA2C - adrenergic, alpha-2c-, receptor CNR1 - cannabinoid receptor 1 (brain) AGTRL1 - angiotensin ii receptor-like 1 P2RX5 - purinergic receptor p2x, ligand-gated ion channel, 5 HRH3 - histamine receptor h3 TSHB - thyroid stimulating hormone, beta GABBR1 - gamma-aminobutyric acid (gaba) b receptor, 1 NPY5R - neuropeptide y receptor y5 PRSS1 - protease, serine, 1 (trypsin 1) GH1 - growth hormone 1 GABRA1 - gamma-aminobutyric acid (gaba) a receptor, alpha 1 P2RY6 - pyrimidinergic receptor p2y, g-protein

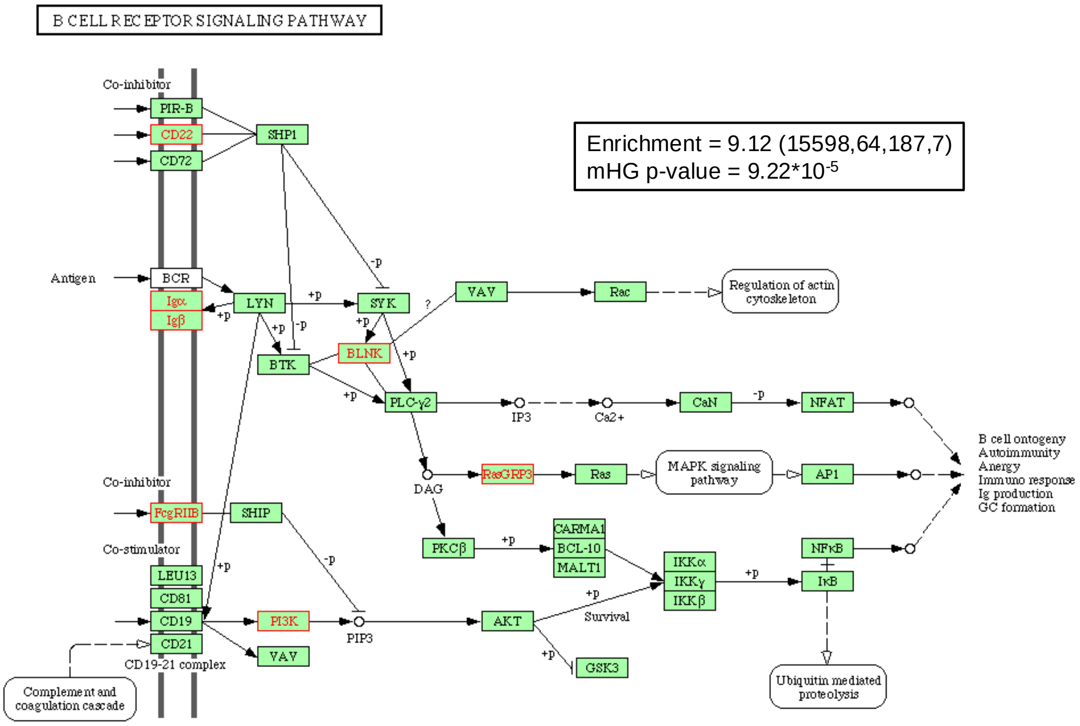
				coupled, 6 TAAR8 - trace amine associated receptor 8 GRM4 - glutamate receptor, metabotropic 4 P2RY10 - purinergic receptor p2y, g-protein coupled, 10 CYSLTR1 - cysteinyl leukotriene receptor 1 PTGER1 - prostaglandin e receptor 1 (subtype ep1), 42kda NMUR2 - neuromedin u receptor 2 OPRK1 - opioid receptor, kappa 1 NPBWR2 - neuropeptides b/w receptor 2 VIPR1 - vasoactive intestinal peptide receptor 1 CRHR2 - corticotropin releasing hormone receptor 2 CHRM1 - cholinergic receptor, muscarinic 1
00602	Glycosphingolipid biosynthesis - neo-lactoseries	1.42E-3	11.44 (15598,19,287,4)	tFUT4 - fucosyltransferase 4 (alpha (1,3) fucosyltransferase, myeloid-specific) FUT5 - fucosyltransferase 5 (alpha (1,3) fucosyltransferase) FUT6 - fucosyltransferase 6 (alpha (1,3) fucosyltransferase) GCNT2 - glucosaminyl (n-acetyl) transferase 2, i-branching enzyme (i blood group)
05213	Endometrial cancer	3.4E-3	7.27 (15598,49,219,5)	TCF7L1 - transcription factor 7-like 1 (t-cell specific, hmg-box) PIK3CD - phosphoinositide-3-kinase, catalytic, delta polypeptide ARAF - v-raf murine sarcoma 3611 viral oncogene homolog TCF7 - transcription factor 7 (t-cell specific, hmg-box) GRB2 - growth factor receptor-bound protein 2
05221	Acute myeloid leukemia	4.12E-3	4.68 (15598,55,424,7)	PIM2 - pim-2 oncogene TCF7L1 - transcription factor 7-like 1 (t-cell specific, hmg-box) PIK3CD - phosphoinositide-3-kinase, catalytic, delta polypeptide CCND1 - cyclin d1 ARAF - v-raf murine sarcoma 3611 viral oncogene homolog TCF7 - transcription factor 7 (t-cell specific, hmg-box) GRB2 - growth factor receptor-bound protein 2
04620	Toll-like receptor signaling pathway	4.55E-3	5.70 (15598,85,193,6)	MAP3K7IP1 - mitogen-activated protein kinase kinase kinase 7 interacting protein 1 PIK3CD - phosphoinositide-3-kinase, catalytic, delta polypeptide FADD - fas (tnfrsf6)-associated via death domain CXCL10 - chemokine (c-x-c motif) ligand 10 IRF7 - interferon regulatory factor 7

				CD40 - cd40 molecule, tnf receptor superfamily member 5
05215	Prostate cancer	6.2E-3	6.67 (15598,86,136,5)	TCF7L1 - transcription factor 7-like 1 (t-cell specific, hmg-box) PIK3CD - phosphoinositide-3-kinase, catalytic, delta polypeptide TCF7 - transcription factor 7 (t-cell specific, hmg-box) GRB2 - growth factor receptor-bound protein 2 INS - insulin

Footnote: P-value' is the enrichment p-value computed according to the mHG or HG model. This p-value is not corrected for multiple testing. Enrichemnt (N, B, n, b) is defnied as follows: "N" is the total number of genes, "B" is the total number of genes associated with a specific term ('target' set and 'background set'), "n" is the number of genes in the 'target set', and "b" is the number of genes in the 'target set' associated with a specific term

Among the pathways over-expressed in smokers, the most enriched ($p = 9.22*10^{-5}$) resulted to be the "B-cell receptor signaling pathway", a key step in innate response immunity and direct recognition of the antigen[98], as well as in B cells selection and survival[99]. The next figure shows the "B-cell receptor signaling pathway" as represented in the KEGG encyclopedia. Genes shown in red appear to be upregulated in presence of smoking.

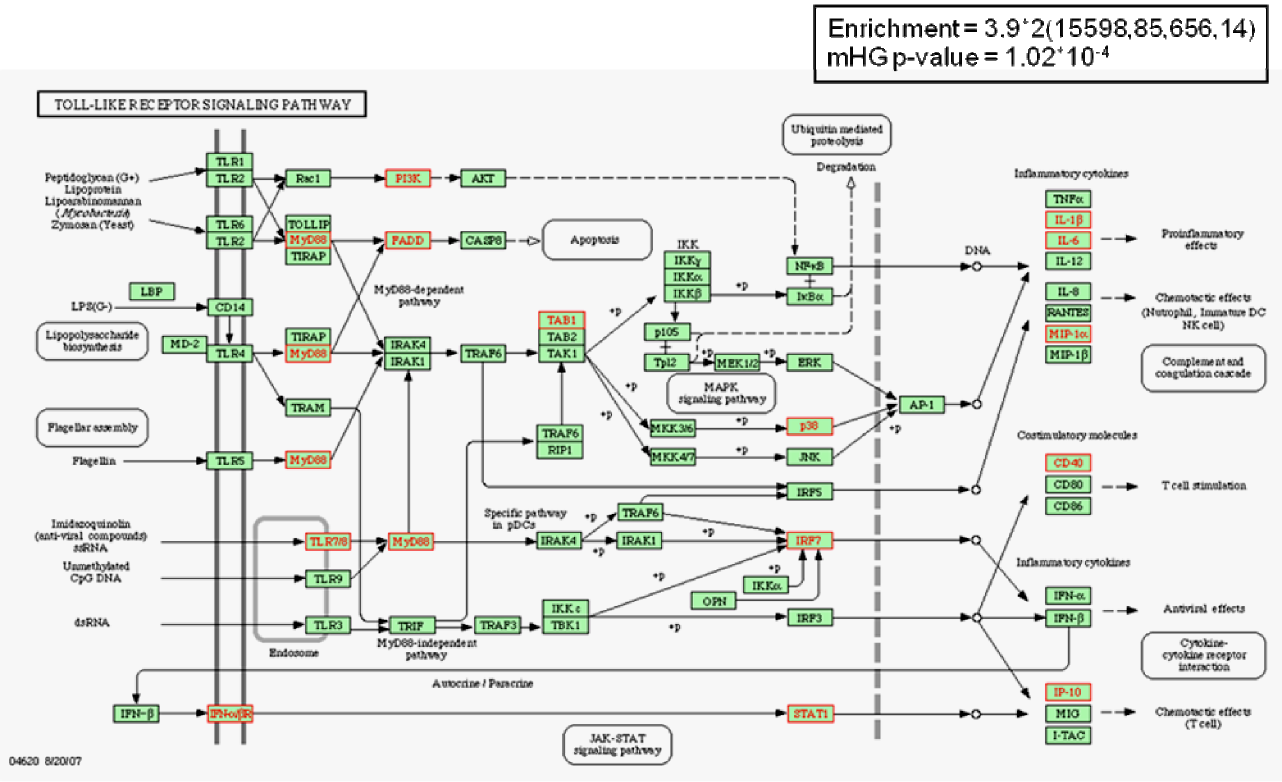
Figure 31: Molecular pathway most over-expressed in current compared to never smokers by enrichment analysis on KEGG database: B-cell receptor signaling pathway.



Footnote to figure: Graphical representation of the "B-cell receptor pathway" as provided in the KEGG database. Genes highlighted in red are over-expressed in smokers compared to non-smokers. Enrichment is defined by mHGM analysis as (b/n) / (B/N), where N: total number of genes, B: total number of genes included in a specific KEGG pathway, n: number of genes in the 'target set', and b: number of genes in the 'target set' included in the specific KEGG pathway.

In addition to this pathway, also the "Toll-like receptor signaling pathway" resulted over-expressed in smokers, further confirming an over-activation of mechanisms related to innate immunity (shown in the following graph).

Figure 32: Molecular pathway most over-expressed in current compared to never smokers by enrichment analysis on KEGG database: Toll-like receptor signaling pathway

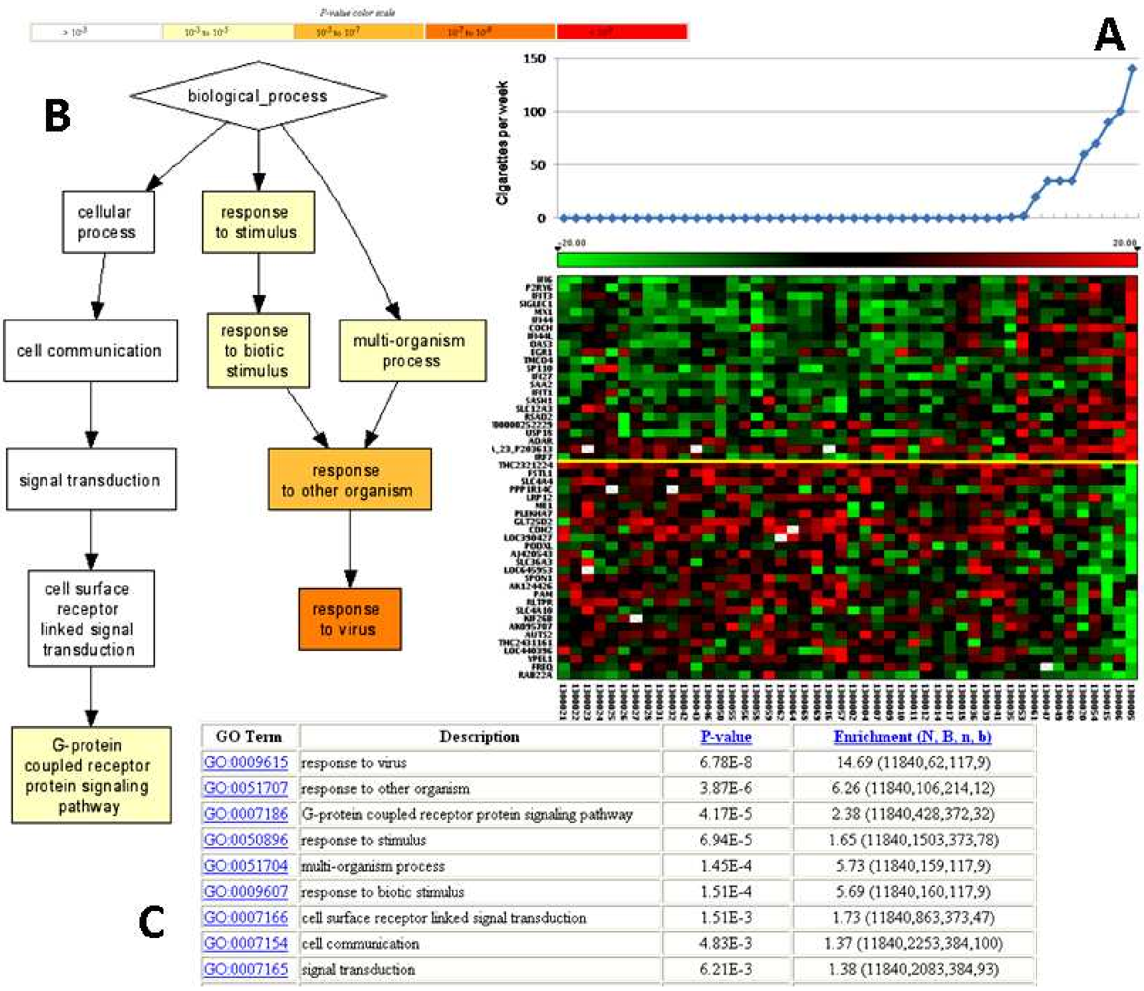


Enrichment = 3.9*2(15598,85,656,14)
mHG p-value = $1.02*10^{-4}$

Footnote to figure: Graphical representation of the "Toll-like receptor signaling pathway" as provided in the KEGG database. Genes highlighted in red are over-expressed in smokers compared to non-smokers. Enrichment is defined by mHGM analysis as (b/n) / (B/N), where N: total number of genes, B: total number of genes included in a specific KEGG pathway, n: number of genes in the 'target set', and b: number of genes in the 'target set' included in the specific KEGG pathway.

The pro-inflammatory signature found in smoker's gene expression profile was also proportional to the level of smoking exposure, as assessed by Pearson's correlation analysis of gene expression with weekly cigarette consumption. Next figure depicts the groups of genes that were significantly and positively correlated with cigarette consumption, assessed as number of cigarettes per week. Panel A shows the heatmap of the 50 most correlated genes (lower part of the panel) sorted on the x-axis in relation to number of smoked cigarettes (upper part of the panel). The enrichment of functional groups is very close to that observed in analyzing binary differential expression for smokers vs non-smokers (Panels B and C).

Figure 33: Gene expression profile correlated with degree of exposure to cigarette smoking



GO Term	Description	P-value	Enrichment (N, B, n, b)
GO:0009615	response to virus	6.78E-8	14.69 (11840,62,117,9)
GO:0051707	response to other organism	3.87E-6	6.26 (11840,106,214,12)
GO:0007186	G-protein coupled receptor protein signaling pathway	4.17E-5	2.38 (11840,428,372,32)
GO:0050896	response to stimulus	6.94E-5	1.65 (11840,1503,373,78)
GO:0051704	multi-organism process	1.45E-4	5.73 (11840,159,117,9)
GO:0009607	response to biotic stimulus	1.51E-4	5.69 (11840,160,117,9)
GO:0007166	cell surface receptor linked signal transduction	1.51E-3	1.73 (11840,863,373,47)
GO:0007154	cell communication	4.83E-3	1.37 (11840,2253,384,100)
GO:0007165	signal transduction	6.21E-3	1.38 (11840,2083,384,93)

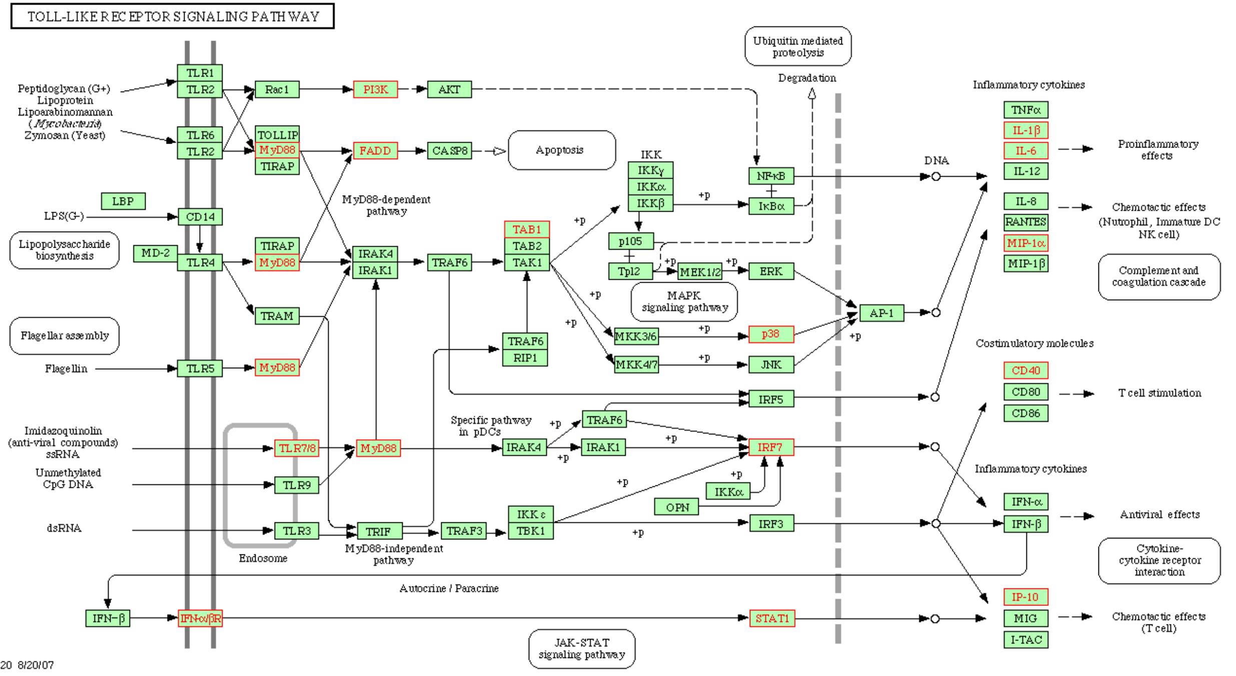
Footnote to figure. Panel A: Heatmap of the 50 genes most correlated with number of cigarettes smoked per-week. Each column corresponds to a patient and each row to a gene. Color represents log2 ratio of sample compared to reference (see Methods). For illustration purposes, only genes with less than 4 missing values have been represented. The full list of genes ranked according to correlation's p-value is available as supplementary material. Panel B: Graphical representation of the portion of the GO hierarchy that includes the terms significantly over-represented in the list of genes directly correlated with degree of current smoking. Color is a function of the p-value of GO term enrichment by mHG analysis and is coded following the scale depicted on top of the panel. Panel C: GO terms significantly enriched in the list of genes directly correlated with degree of current smoking. Enrichment is defined as (b/n) / (B/N), where N: total number of genes, B:

total number of genes associated with a specific GO term, n: number of genes in the 'target set', and b: number of genes in the 'target set' associated with a specific GO term. The reported p-values are not corrected for multiple testing spanning 5081 GO terms.

In addition, the "Toll-like receptor signaling pathway" resulted the most enriched KEGG pathway amongst the genes positively correlated also with the degree of smoking (next figure).

Figure 34: Molecular pathway most correlated with degree of exposure to cigarette smoking by enrichment analysis on KEGG database.

Enrichment = 3.9*2(15598,85,656,14)
mHG p-value = $1.02*10^{-4}$



Footnote to figure. Graphical representation of the "Toll-like receptor pathway" as provided in the KEGG database. Genes highlighted in red are correlated with degree of exposure to cigarette smoking. Enrichment is defined by mHGM analysis as (b/n) / (B/N), where N: total number of genes, B: total number of genes included in a specific KEGG pathway, n: number of genes in the 'target set', and b: number of genes in the 'target set' included in the specific KEGG pathway.

7.5.3 High LDL-cholesterol is associated with a pro-inflammatory gene expression profile characterized by an over-expression of genes related to cell-mediated immune response

The next table illustrates the clinical characteristics of the study population stratified by levels of LDL-cholesterol: subjects of the third LDL tertile were characterized, in average, by moderately elevated LDL-cholesterol levels (157±18 mg/dL) and did not differ in any of the other clinical parameters from the individuals with lower LDL-cholesterol concentrations (1st and 2nd tertiles). Only hsCRP resulted slightly, but not significantly, elevated in the 3rd LDL-cholesterol tertile.

Table 8: Clinical characteristics of the study population stratified by tertiles of LDL Cholesterol.

Variables	LDL Cholesterol tertiles			Trend
	1st (40-104 mg/dL)	2nd (106-138 mg/dL)	3rd (140-197 mg/dL)	p
Gender [M/F (%)] **	5/10 (33.3/66.7)	6/10 (37.5/62.5)	9/7 (56.2/43.8)	NS
Age [yy]	34 ± 8	36 ± 7	37 ± 9	NS
Family history of CAD [count (%)] **	2 (14)	1 (7)	2 (15)	NS
Family history of CVD [count (%)] **	3 (21)	3 (23)	3 (21)	NS
Smoking habit [current/former/never (%)] **	1/2/11 (7.1/14.3/78.6)	4/3/8 (26.7/20.0/53.3)	3/3/8 (21.4/21.4/57.1)	NS
BMI [kg/m2]	24 ± 2	22 ± 3	25 ± 3	NS
Waist [cm]	85 ± 7	81 ± 10	90 ± 8	NS
Systolic BP [mmHg]	112 ± 16	110 ± 15	116 ± 13	NS
Diastolic BP [mmHg]	66 ± 21	72 ± 13	76 ± 12	NS
Fasting plasma glucose [mg/dL]	85 ± 5	85 ± 6	91 ± 5	NS
2-hour plasma glucose [mg/dL]	87 ± 27	91 ± 17	100 ± 21	NS

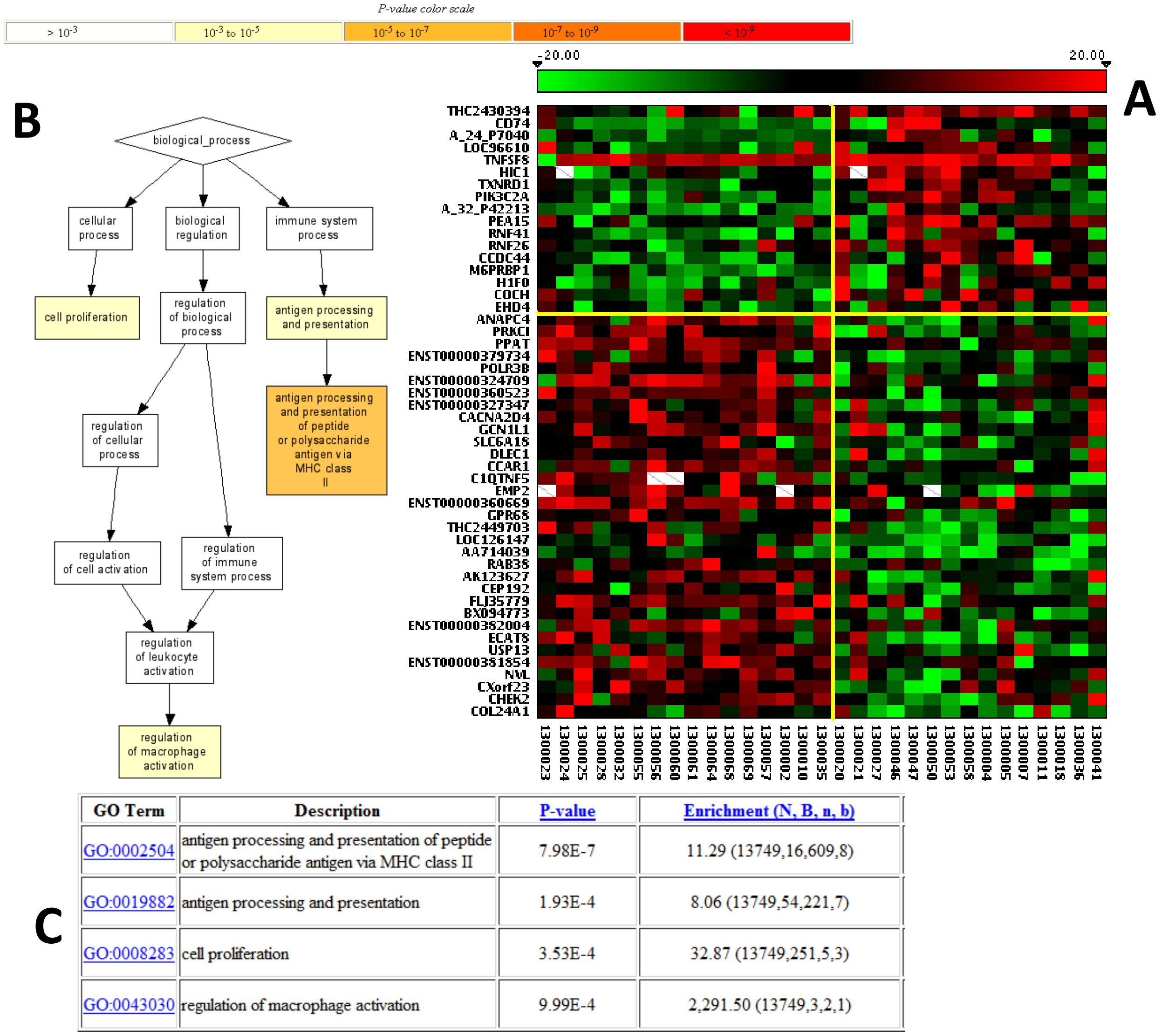
Variables	**LDL Cholesterol tertiles**			**Trend**
	1st (40-104 mg/dL)	**2nd** (106-138 mg/dL)	**3rd** (140-197 mg/dL)	**p**
Total-Cholesterol [mg/dL]	168 ± 25	203 ± 22	240 ± 24	NS
HDL-Cholesterol [mg/dL]	71 ± 14	65 ± 15	59 ± 14	NS
Triglycerides [mg/dL] *	53 (32)	66 (40)	70 (132)	NS
WBC count [# * 10^3/mm3]	5.52 ± 0.82	6.24 ± 1.75	6.00 ± 1.11	NS
hsCRP [mg/L] *	0.35 (0.72)	0.36 (0.85)	0.91 (2.00)	NS

Footnote to table: Continuous variables are expressed as mean ± SD or median (IQR), if not normally distributed (*); nominal variables (**) are represented as number (percent). Significance of trend among groups has been evaluated using analysis of variance (ANOVA) or Chi-square test (**) as appropriate. Not normally distributed variables (*) were log-transformed before performing ANOVA test. CAD = Coronary Artery Disease, CVD = Cardio-Vascular Diseases, BMI = Body Mass Index, BP = Blood Pressure, HDL = High-Density Lipoprotein, LDL = Low-Density Lipoprotein, WBC = White Blood Cells, hsCRP = high sensitivity C-Reactive Protein.

However, we detected a significant association between the expression level of many genes and the individual's LDL-cholesterol concentration. To better characterize the genes that were differentially expressed in high compared to low LDL-cholesterol concentrations, we compared LDL≤ 104 to LDL≥ 140 mg/dL, the extreme tertiles of the cohort distribution.

As shown in the next figure, the gene expression signature associated with high LDL-cholesterol was characterized by an over-representation of genes involved in the inflammatory response, with a preferential enrichment of GO terms associated with the cell-mediated immunity such as "antigen processing and presentation via MHC class II" ($p = 8.0*10^{-7}$).

Figure 35: Gene expression profile in subjects with high (> 140 mg/dL) compared to low (< 104 mg/dL) plasma LDL-cholesterol concentrations.

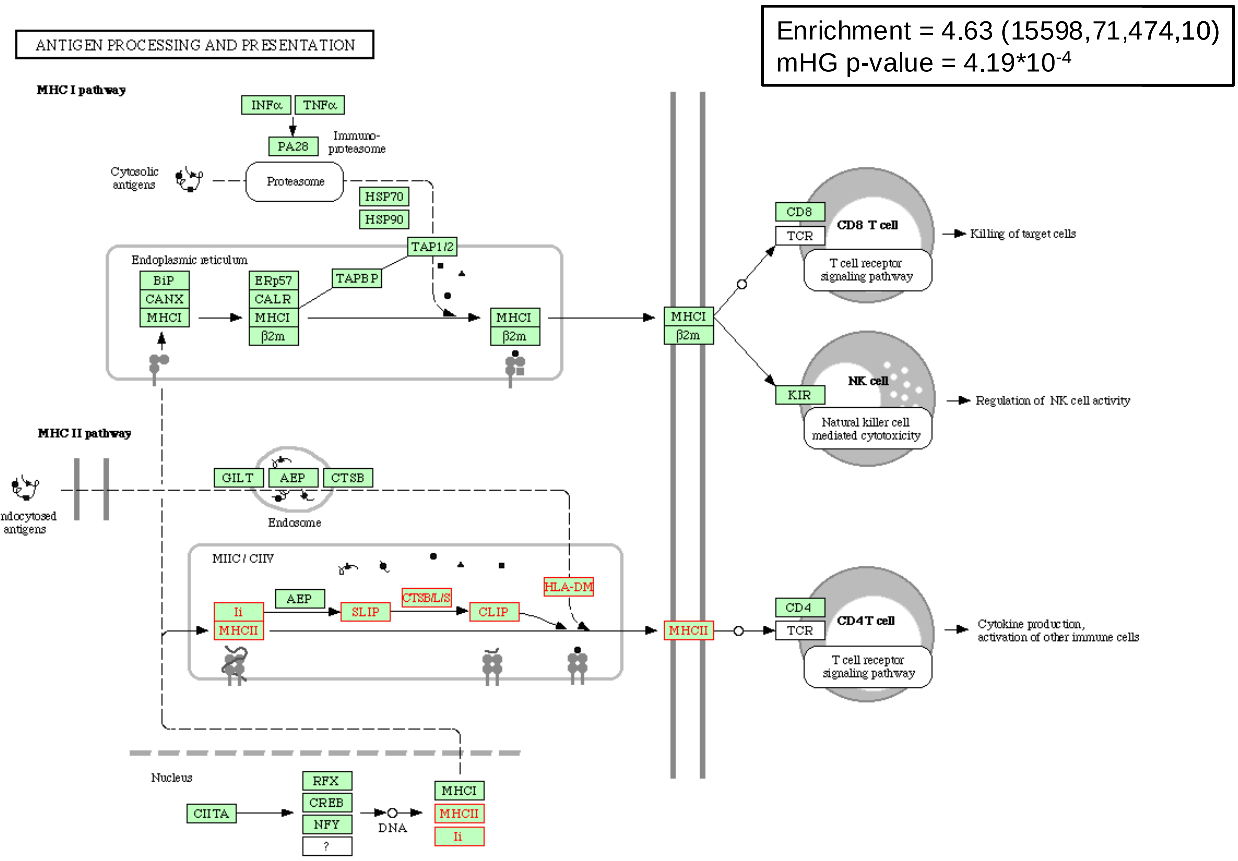


GO Term	Description	P-value	Enrichment (N, B, n, b)
GO:0002504	antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	7.98E-7	11.29 (13749,16,609,8)
GO:0019882	antigen processing and presentation	1.93E-4	8.06 (13749,54,221,7)
GO:0008283	cell proliferation	3.53E-4	32.87 (13749,251,5,3)
GO:0043030	regulation of macrophage activation	9.99E-4	2,291.50 (13749,3,2,1)

Footnote to figure. Heatmap of the 50 most differentially expressed genes between high and low LDL-cholesterol concentrations determined by semi-supervised class discovery. Each column corresponds to a patient and each row to a gene. Color represents log2 ratio of sample compared to reference. For illustration purposes, only genes with less than 4 missing values have been represented. Panel B: Graphical representation of the portion of the GO hierarchy that includes the terms significantly over-represented in the list of genes over-expressed in high LDL. Color is a function of the p-value of GO term enrichment by mHG analysis and is coded following the scale depicted on top of the panel. Panel C: GO terms significantly enriched in the list of genes over-expressed in high LDL. Enrichment is defined as (b/n) / (B/N), where N: total number of genes, B: total number of genes associated with a specific GO term, n: number of genes in the 'target set', and b: number of genes in the 'target set' associated with a specific GO term. The reported p-values are not corrected for multiple testing.

In addition, the most enriched cellular pathway in presence of high LDL-cholesterol resulted to be the antigen processing and presentation through the MHC class II pathway ($p = 1.8*10^{-4}$) whose key genes were all significantly up-regulated (next figure).

Figure 36: Molecular pathway most over-expressed in high (> 140 mg/dL) compared to low (< 104 mg/dL) LDL-cholesterol by enrichment analysis on KEGG database.



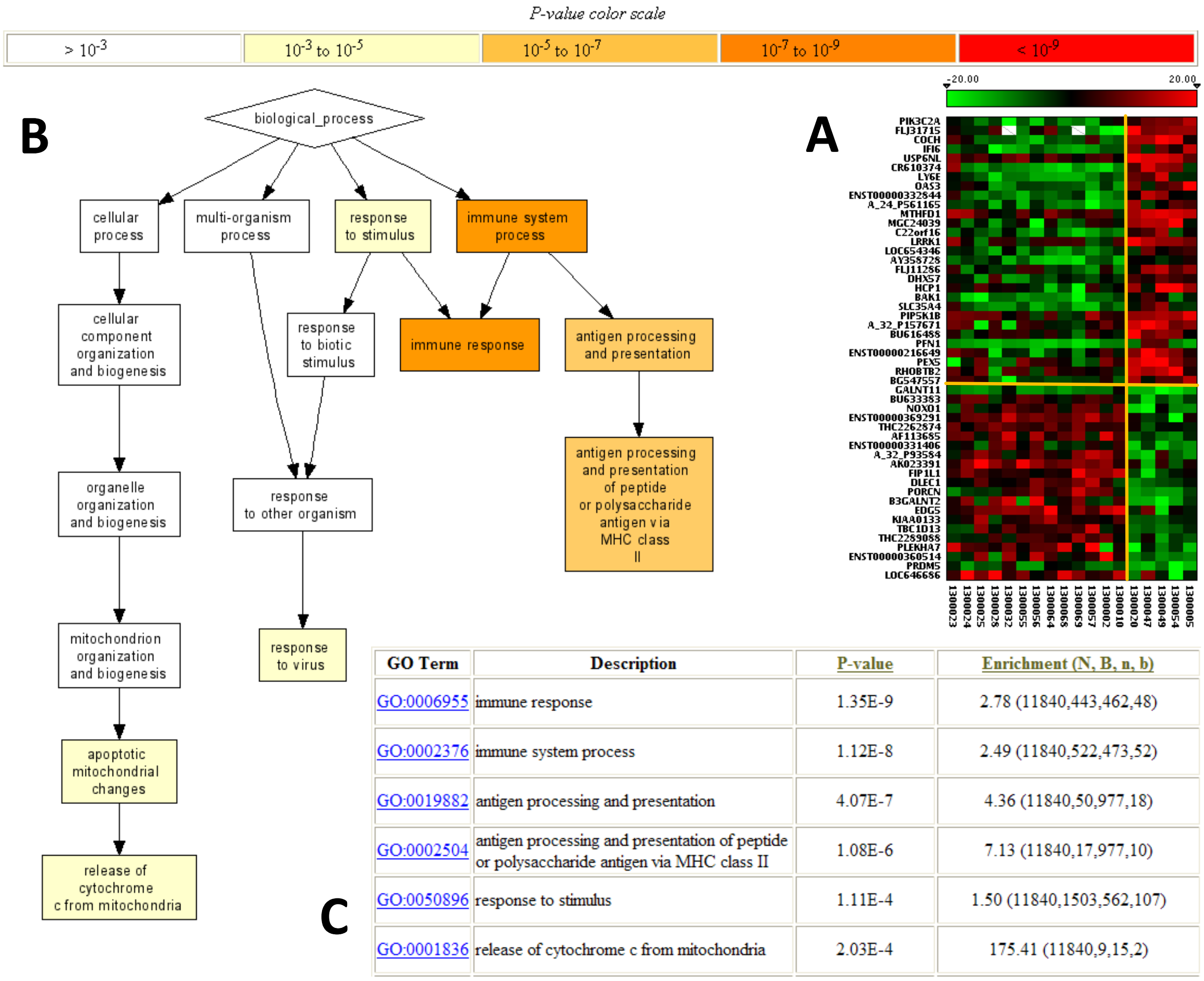
Enrichment = 4.63 (15598,71,474,10)
mHG p-value = $4.19*10^{-4}$

Footnote to figure. Graphical representation of the "Antigen processing and presentation pathway" as provided in the KEGG database. Genes highlighted in red are over-expressed in subjects with high LDL-cholesterol concentrations. Enrichment is defined by mHGM analysis as (b/n) / (B/N), where N: total number of genes, B: total number of genes included in a specific KEGG pathway, n: number of genes in the 'target set', and b: number of genes in the 'target set' included in the specific KEGG pathway.

7.5.4 Cigarette smoking and high LDL-cholesterol activate both innate as well as cell-mediated immune response

In the previous sections, we described the response of PBMCs to cigarette smoking and to high LDL-cholesterol, in terms of changes in mRNA expression levels. Exposure to smoking leads to an activation of the innate immune response system while high plasma LDL-cholesterol concentrations are associated with an activation of the cell-mediated immune response pathways. It is therefore likely that different risk factors act through distinct pathophysiological mechanisms and the pathogenesis of atherosclerosis might be driven by each factor in a distinct manner. To test the effect of the presence of both factors in the same individual we compared the PBMC expression profiles of current smokers with high LDL-cholesterol concentration to that of never smokers with low plasma LDL-cholesterol. As depicted in the following figure, we observed a strong enrichment of gene sets related to both types of immune response, including general innate inflammatory response (Response to stimulus with $p \sim 10^{-4}$) as well as cell-mediated immunity. The latter is exemplified by the GO term including genes involved in antigen presentation through the HMC class II pathway, which contains 17 members in the entire human genome, 10 of which appeared significantly differentially expressed in our comparison. The latter enrichment yields a p-value of $1.08*10^{-6}$.

Figure 37: Gene expression profile in current cigarette smokers with high plasma LDL-cholesterol concentrations compared to non-smokers with low LDL-cholesterol.



GO Term	Description	P-value	Enrichment (N, B, n, b)
GO:0006955	immune response	1.35E-9	2.78 (11840,443,462,48)
GO:0002376	immune system process	1.12E-8	2.49 (11840,522,473,52)
GO:0019882	antigen processing and presentation	4.07E-7	4.36 (11840,50,977,18)
GO:0002504	antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	1.08E-6	7.13 (11840,17,977,10)
GO:0050896	response to stimulus	1.11E-4	1.50 (11840,1503,562,107)
GO:0001836	release of cytochrome c from mitochondria	2.03E-4	175.41 (11840,9,15,2)

Footnote to figure. Panel A: Heatmap of the 50 most differentially expressed genes between current smokers with high LDL-cholesterol and never smokers with low LDL-cholesterol. Each column corresponds to a patient and each row to a gene. Color represents log2 ratio of sample compared to reference (see methods) and legend scale is depicted above the heatmap. For illustration purposes, only genes with less than 25% of missing values have been represented. Panel B: Graphical representation of the portion of GO tree including the terms significantly over-represented in the list of genes over-expressed in current smokers with high LDL-cholesterol concentrations. Color is a function of the p-value of GO term enrichment by mHG analysis and is coded following the scale depicted on top of the panel. Panel C: GO terms significantly enriched in the list of genes over-expressed in current smokers with high LDL-cholesterol concentrations. Enrichment is defined as (b/n) / (B/N), where N: total number of genes, B: total number of genes associated with a specific GO term, n: number of genes in the 'target set', and b: number of genes in the 'target set' associated with a specific GO term.

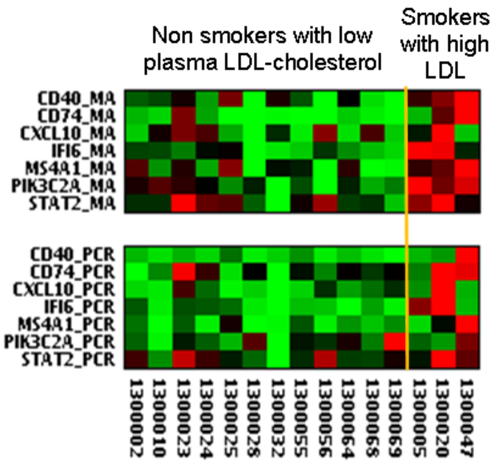
7.5.5 RT-PCR validation

Out of the list of genes over-expressed in smokers with high plasma LDL-cholesterol concentrations and ranked according to TNoM and t-test p-value, we identified 7 highly meaningful candidate genes representing both innate and cell-mediated immunity, including chemokine (CXCL10), interferon (IFI6 and STAT2) and TNF (CD40) related genes, HLA class II histocompatibility (CD74) and immune cell activation/ proliferation (CD20).

As shown in the following figure, the expression profile of the validation genes resulted significantly higher in smokers with high LDL compared to non smoker with low LDL concentrations. The expression pattern in each subject appeared to be similar when assayed by microarray and traditional RT-PCR evaluation as shown by head-to-head comparison of heatmaps.

Figure 38: Expression profile of the seven validation genes assessed by microarray and RT-PCR essay.

Systematic name	Gene Name	Description	T-test p-value
NM_001565	CXCL10	Small inducible cytokine B10 precursor	0.029
NM_001250	CD40	Tumor necrosis factor receptor superfamily member 5 precursor	0.022
NM_004355	CD74	HLA class II histocompatibility antigen gamma chain	0.001
NM_022873	IFI6	Interferon-induced protein 6-16 precursor	9.21E-05
NM_152866	MS4A1	Lymphocyte antigen CD20	0.007
NM_002645	PIK3C2A	Homo sapiens phosphoinositide-3-kinase, class 2, alpha polypeptide	8.15E-05
NM_005419	STAT2	Signal transducer and activator of transcription 2	0.014



Foonote to figure. Table lists annotation data of the seven validation genes and p-value for Student's T-test comparison between smokers with high plasma LDL-cholesterol concentrations and non-smokers with low LDL-cholesterol. Figure shows the heatmap of expression profile of the seven validation genes assessed by microarray (top-panel) and RT-PCR essay (bottom panel). Color represents log2 ratio of sample compared to reference for microarray data and arbitrary units for RT-PCR.

7.6 Gene expression profile in relation to carotid intima-media thickness

Although ours as well as other group's data have been showing the presence of a significant correlation between CV risk factors and certain gene expression profiles, commonly characterized by a pro-inflammatory signature, no data so far have emerged investigating whether the gene expression profile is correlated with the extent and severity of atherosclerotic vascular damage, also in a pre-clinical stage. To pursue this aim, we investigate the relationship between gene expression profile and carotid IMT values in our population of apparently ehalty young adults. The following table shows the characteristics of our study population when stratified by quartiles of IMT value. As it is appears, also the subjects in top IMT quartile showed relatively normal IMT values, in most of the cases below 1 mm. Despite this substantially normal carotid IMT profile, subjects in the fourth quartile showed a significantly higher CV risk profile characterized by a prevalent proportion of males over females, a greater age and adiposity (both BMI and waist circumference), and higher plasma glucose and LDL-cholesterol concentrations.

Table 9: Clinical characteristics of the study overall population stratified by IMT values

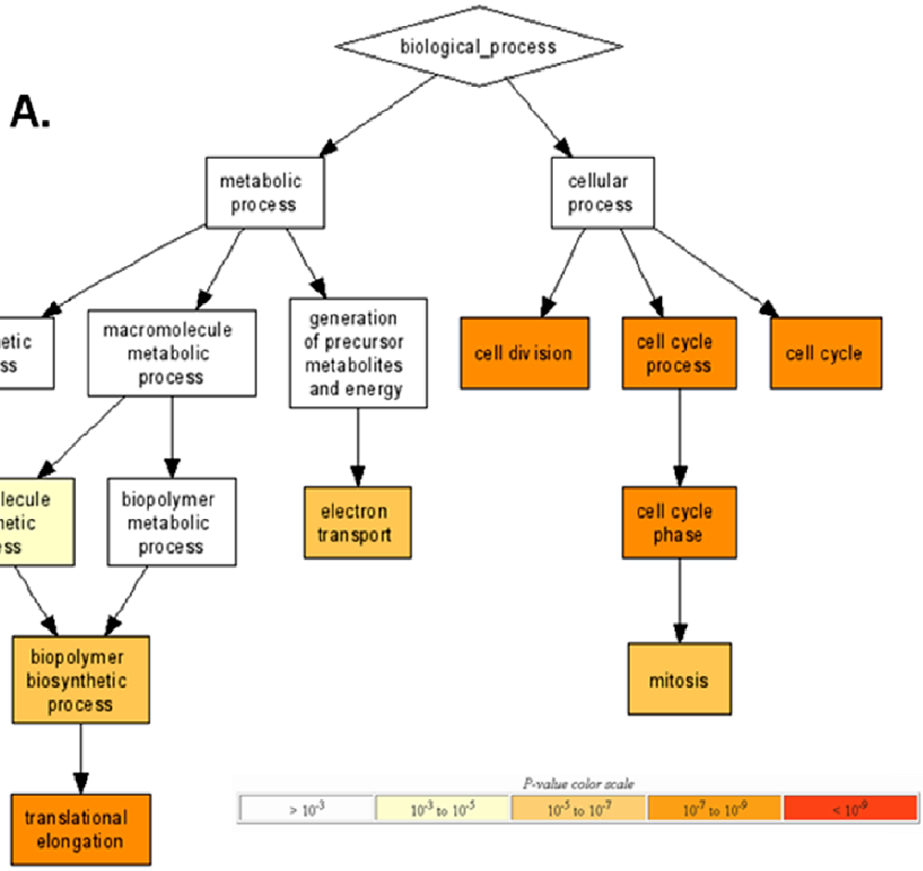
Variable	IMT quartile				Linear contrast ANOVA	
	1st 0.60-0.74 mm	**2nd** 0.75-0.80 mm	**3rd** 0.81-0.88 mm	**4th** 0.90-1.06 mm	**F**	**P**
Gender [M(%)]	11 (35.5)	18 (48.6)	30 (76.9)	21 (58.3)	-	,004
Age [yy]	33 ± 6	36 ± 8	37 ± 7	42 ± 9	7,892	,000
BMI [kg/m2]	23.6 ± 2.8	23.2 ± 3.3	25.3 ± 3.6	25.9 ± 5.6	3,846	,011
Waist [cm]	86 ± 8	85 ± 9	90 ± 11	93 ± 13	4,264	,006
SBP [mmHg]	114 ± 11	115 ± 14	116 ± 16	122 ± 17	1,988	NS
DBP [mmHg]	72 ± 16	75 ± 9	77 ± 12	77 ± 10	1,067	NS
FPG [mg/dL]	85 ± 5	86 ± 6	88 ± 6	91 ± 14	4,139	,008

Variable	IMT quartile				Linear contrast ANOVA	
	1st 0.60-0.74 mm	**2nd** 0.75-0.80 mm	**3rd** 0.81-0.88 mm	**4th** 0.90-1.06 mm	**F**	**P**
2h PG [mg/dL]	99 ± 26	90 ± 20	88 ± 20	102 ± 39	2,242	,086
Tot Chol [mg/dL]	199 ± 35	202 ± 34	196 ± 30	216 ± 31	2,857	,039
HDL Chol [mg/dL]	62 ± 17	64 ± 15	56 ± 13	59 ± 14	1,784	NS
TG [mg/dL]	58 (75)	58 (38)	72 (34)	58 (34)	,916	NS
LDL Chol [mg/dL]	119 ± 32	125 ± 31	125 ± 29	142 ± 27	3,721	,013
hsCRP [mg/dL]	0.75 (0.81)	0.50 (0.55)	0.63 (0.93)	1.04 (1.89)	1,297	NS
Smokers [N(%)]	6 (18.8)	6 (18.8)	12 (36.4)	9 (27.3)	-	NS

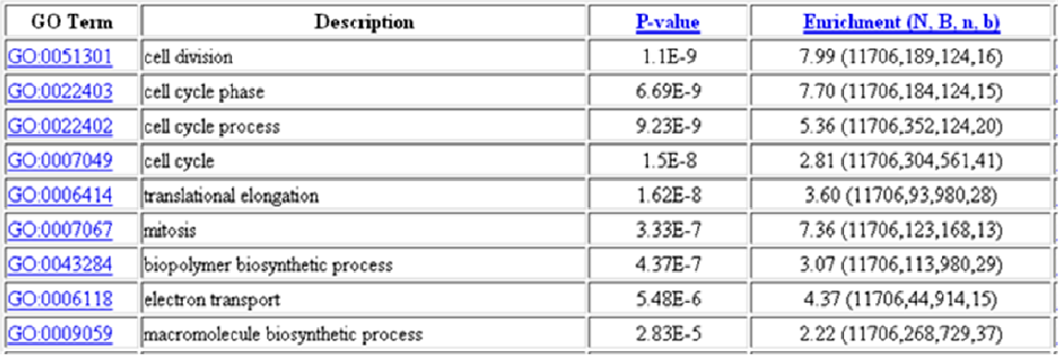
Footnote to the table: Continuous variables are expressed as mean ± SD or median (IQR), if not normally distributed (*); nominal variables (**) are presented as number (percent). BMI = Body Mass Index, BP = Blood Pressure, HDL = High-Density Lipoprotein, LDL = Low-Density Lipoprotein, WBC = White Blood Cells, hsCRP = high sensitivity C-Reactive Protein.

In addition to these findings, IMT values were significantly correlated with a PBMC transcriptomic profile characterized by increased expression of groups of genes involved in cell cycle and mitosis (GO terms: “Cell division” p=1.1E-9; “Cell cycle phase” p=6.7E-9; “Cell cycle process” p=1.5E-8;”Mitosis” p=3.3E-7; etc.) and mithocondrial oxidative phosphorylation (“electron transport” p=5.5E-6; “mitochondrial electron transport” p=5.4E-5),as shown in the following figure.

Figure 39: Groups of genes significantly correlated with IMT values



B.



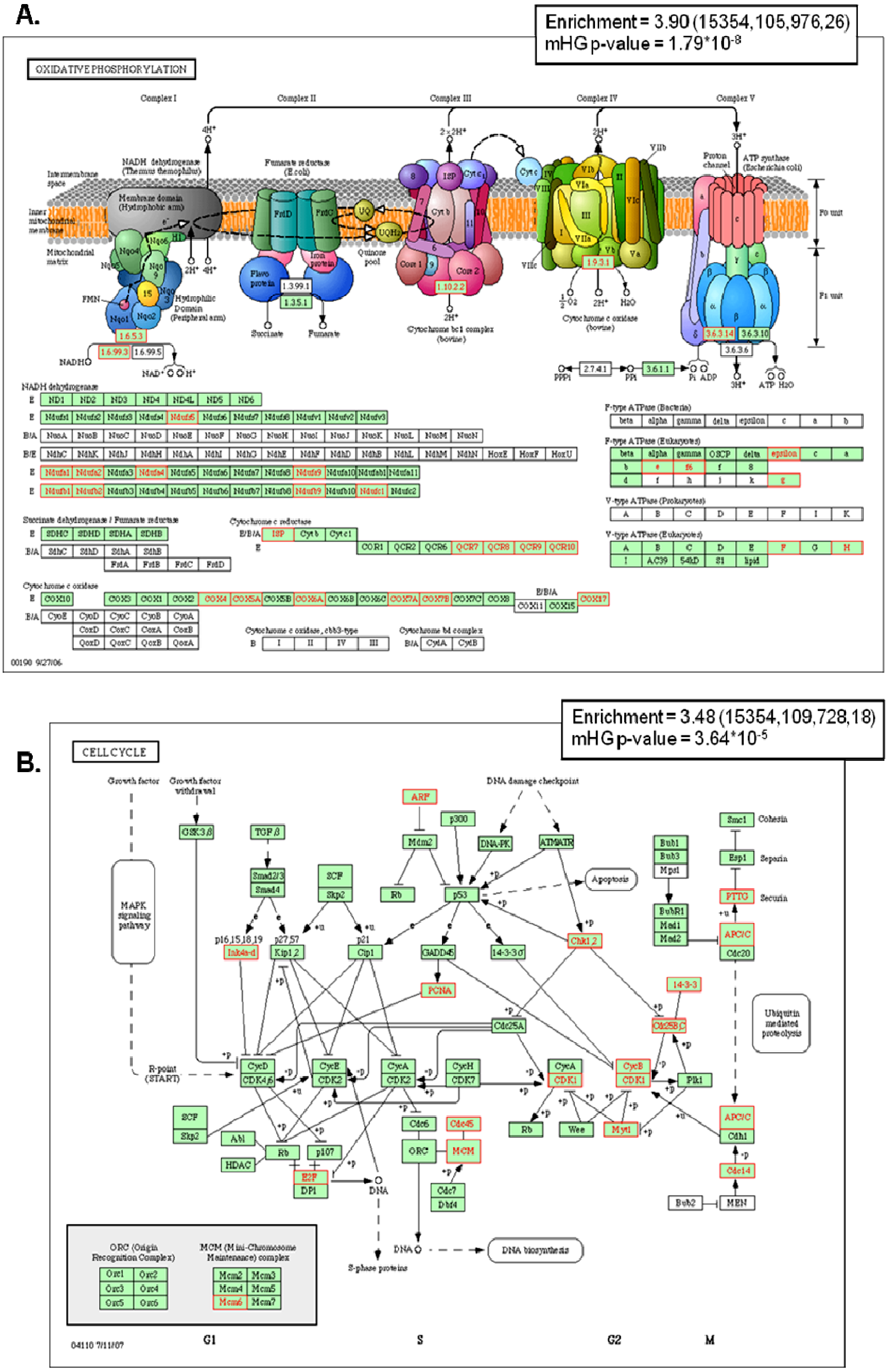
GO Term	Description	P-value	Enrichment (N, B, n, b)
GO:0051301	cell division	1.1E-9	7.99 (11706,189,124,16)
GO:0022403	cell cycle phase	6.69E-9	7.70 (11706,184,124,15)
GO:0022402	cell cycle process	9.23E-9	5.36 (11706,352,124,20)
GO:0007049	cell cycle	1.5E-8	2.81 (11706,304,561,41)
GO:0006414	translational elongation	1.62E-8	3.60 (11706,93,980,28)
GO:0007067	mitosis	3.33E-7	7.36 (11706,123,168,13)
GO:0043284	biopolymer biosynthetic process	4.37E-7	3.07 (11706,113,980,29)
GO:0006118	electron transport	5.48E-6	4.37 (11706,44,914,15)
GO:0009059	macromolecule biosynthetic process	2.83E-5	2.22 (11706,268,729,37)

Footnote to figure. Panel A shows a graphical representation of the portion of the GO hierarchy that includes the terms significantly over-represented in the list of genes over-expressed in correlation with carotid IMT values. Panel B enlists in a tabular form the GO terms significantly enriched in the list of genes over-expressed in presence of higher TNF-alpha concentrations.

Consistently with these findings, the “oxidative phosphorylation” pathway resulted as the most significantly up-regulated KEGG pathway in presence of high IMT (p=1.8E-8), followed by the cell cycle pathway (p=3.6E-5).

The genes significantly up-regulated in relation with IMT values of the two pathways are depicted in red in the following figure.

Figure 40: Molecular pathways significantly over-expressed in presence of higher IMT values



Footnote to figure: Graphical representation of the two molecular pathways most correalted with IMT values as provided in the KEGG database. Genes highlighted in red are over-expressed in smokers compared to non-smokers. Enrichment is defined by mHGM analysis as (b/n) / (B/N), where N: total number of genes, B: total number of genes included in a specific KEGG pathway, n: number of genes in the 'target set', and b: number of genes in the 'target set' included in the specific KEGG pathway.

Although this profile resulted consistent in both types of analysis (GO and KEGG enrichment analysis) and significantly associated with IMT values, it was not possible from a bivariate approach to exclude the possible presence of confounders, as IMT resulted associated with several clinical characteristics and CVD risk factors.

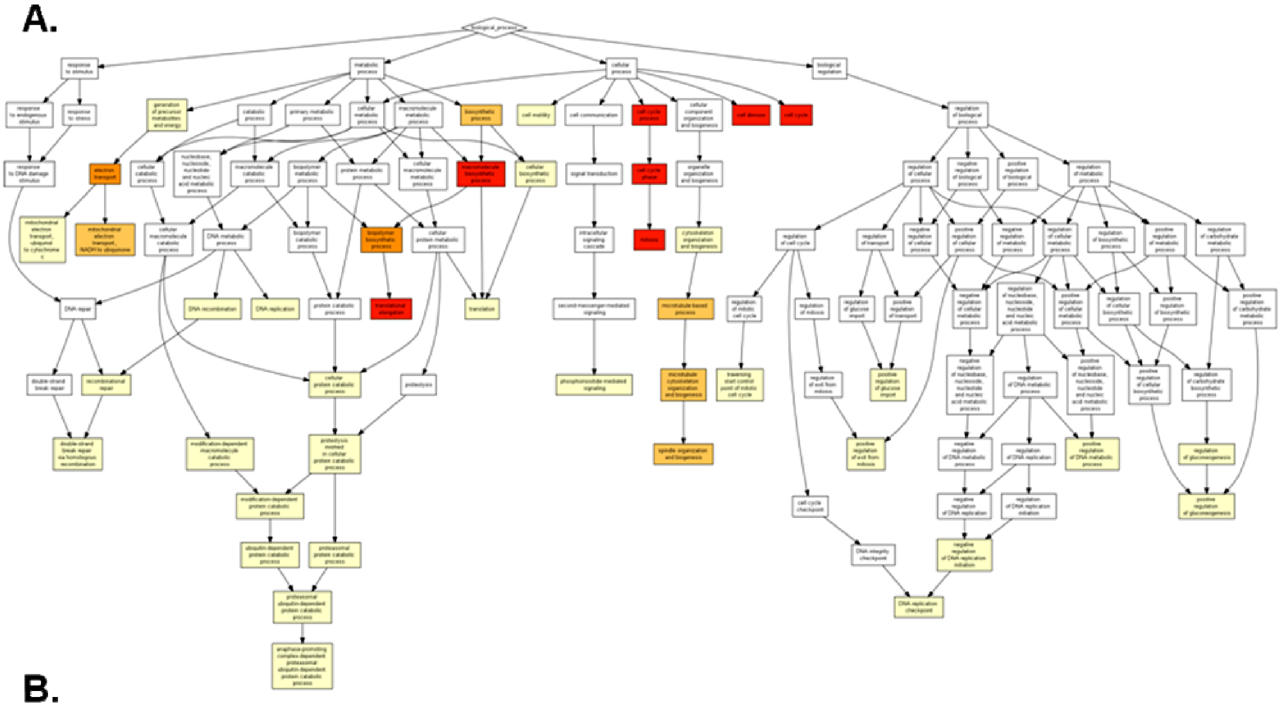
As shown by table 4, IMT resulted increate in subjects of male gender, and higher age, BMI, waist circumference, plasma glucose and LDL-cholesterol concentrations. To adjust IMT values for the presence of confounders, we performed a multivariate linear regression analysis, including gender, age, BMI, plasma glucose and LDL-cholesterol as independent variables. Out of these variables, only age resulted independently correlated with IMT values (next table).

Table 10: IMT predictors by linear regression analysis

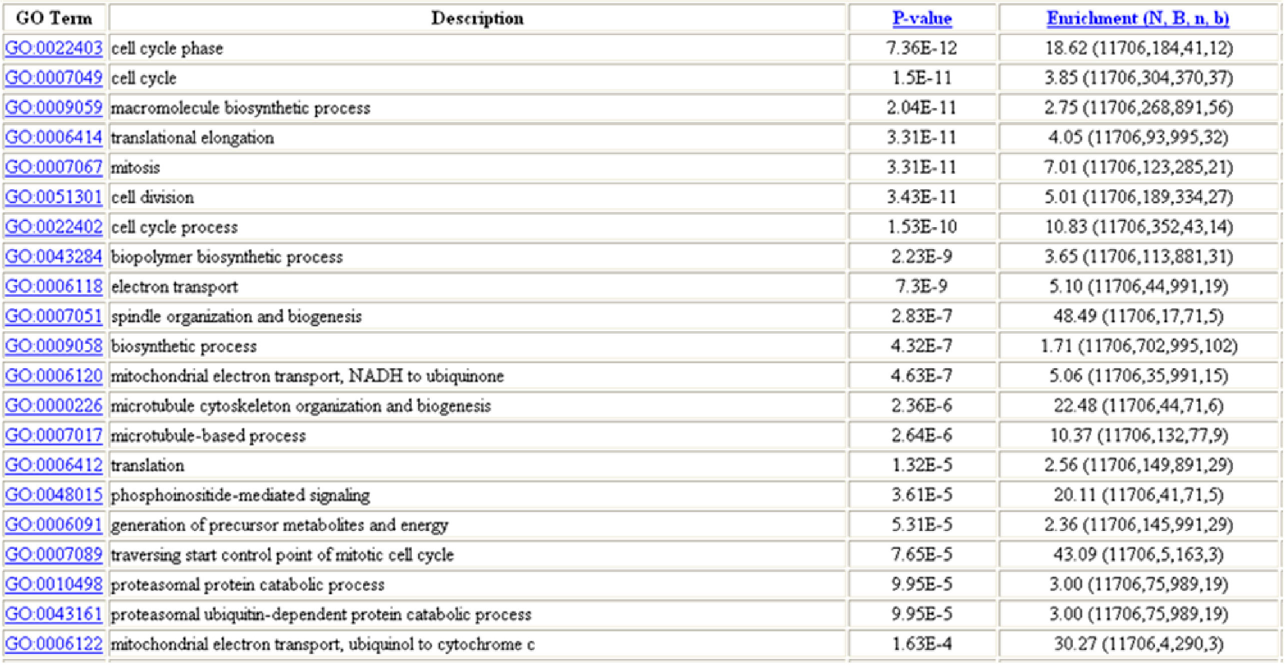
Variabile	Beta	p	R^2
Sesso	0.112	N.S.	0,187
Età	0.440	<0.001	
BMI	-0.021	N.S.	
Glicemia a digiuno	0.070	N.S.	
Glicemia post-carico	-0.070	N.S.	
LDL colesterolo	0.021	N.S.	

We therefore adjusted IMT values by age using linear regression analysis. After age-adjustment, IMT values resulted significantly correlated with the same gene groups and pathways with a higher level of significance. An increase in transcription of oxidative phosphorylation genes resulted the bio-signal most strongly correlated with age-adjusted IMT values (p=6.4E-12), as shown in the following figure. The value of the significance of the association resulted even increased after age-adjustment.

Figure 41: Groups of genes significantly correlated with IMT values after age-adjustment


A.
biological_process
response to stimulus
metabolic process
cellular process
biological regulation
response to endogenous stimulus
response to stress
generation of precursor metabolites and energy
catabolic process
primary metabolic process
cellular metabolic process
macromolecule metabolic process
biosynthetic process
cell motility
cell communication
cell cycle process
cellular component organization and biogenesis
cell division
cell cycle
regulation of biological process
response to DNA damage stimulus
electron transport
cellular catabolic process
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
macromolecule catabolic process
biopolymer metabolic process
protein metabolic process
cellular macromolecule metabolic process
macromolecule biosynthetic process
cellular biosynthetic process
signal transduction
cell cycle phase
organelle organization and biogenesis
regulation of cellular process
negative regulation of biological process
positive regulation of biological process
regulation of metabolic process
mitochondrial electron transport, ubiquinol to cytochrome c
mitochondrial electron transport, NADH to ubiquinone
cellular macromolecule catabolic process
DNA metabolic process
biopolymer catabolic process
biopolymer biosynthetic process
cellular protein metabolic process
intracellular signaling cascade
mitosis
cytoskeleton organization and biogenesis
regulation of cell cycle
regulation of transport
negative regulation of cellular process
positive regulation of cellular process
negative regulation of metabolic process
regulation of cellular metabolic process
regulation of biosynthetic process
positive regulation of metabolic process
regulation of carbohydrate metabolic process
DNA repair
DNA recombination
DNA replication
protein catabolic process
translational elongation
translation
second-messenger-mediated signaling
microtubule-based process
regulation of mitotic cell cycle
regulation of mitosis
regulation of glucose import
positive regulation of transport
negative regulation of cellular metabolic process
regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
positive regulation of cellular metabolic process
regulation of cellular biosynthetic process
positive regulation of biosynthetic process
positive regulation of carbohydrate metabolic process
double-strand break repair
recombinational repair
cellular protein catabolic process
proteolysis
phosphoinositide-mediated signaling
microtubule cytoskeleton organization and biogenesis
traversing start control point of mitotic cell cycle
regulation of exit from mitosis
positive regulation of glucose import
negative regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
regulation of DNA metabolic process
positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
positive regulation of cellular biosynthetic process
regulation of carbohydrate biosynthetic process
double-strand break repair via homologous recombination
modification-dependent macromolecule catabolic process
proteolysis involved in cellular protein catabolic process
spindle organization and biogenesis
positive regulation of exit from mitosis
negative regulation of DNA metabolic process
regulation of DNA replication
positive regulation of DNA metabolic process
regulation of gluconeogenesis
modification-dependent protein catabolic process
cell cycle checkpoint
negative regulation of DNA replication
regulation of DNA replication initiation
positive regulation of gluconeogenesis
ubiquitin-dependent protein catabolic process
proteasomal protein catabolic process
DNA integrity checkpoint
negative regulation of DNA replication initiation
proteasomal ubiquitin-dependent protein catabolic process
DNA replication checkpoint
anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process

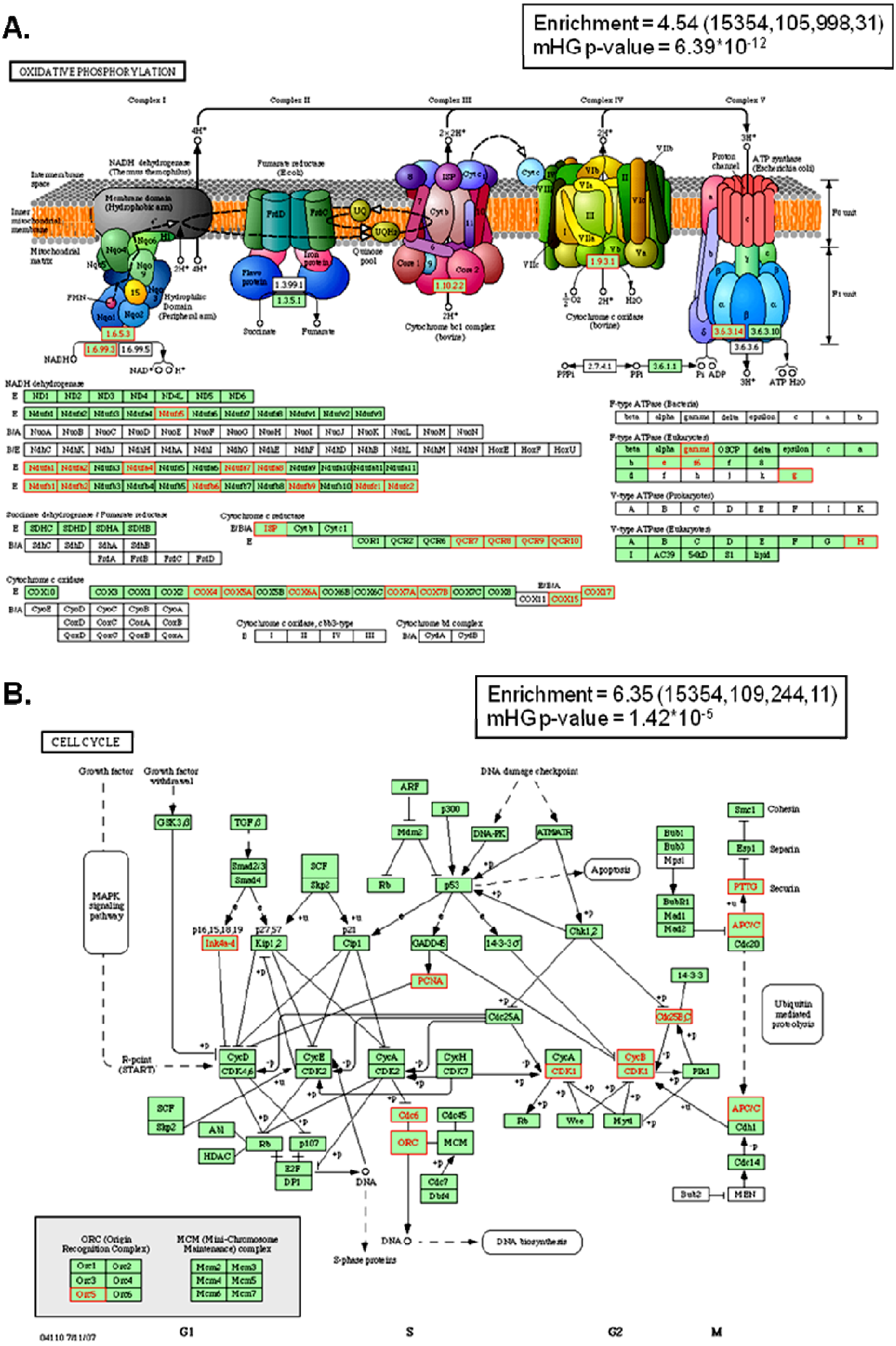
B.



GO Term	Description	P-value	Enrichment (N, B, n, b)
GO:0022403	cell cycle phase	7.36E-12	18.62 (11706,184,41,12)
GO:0007049	cell cycle	1.5E-11	3.85 (11706,304,370,37)
GO:0009059	macromolecule biosynthetic process	2.04E-11	2.75 (11706,268,891,56)
GO:0006414	translational elongation	3.31E-11	4.05 (11706,93,995,32)
GO:0007067	mitosis	3.31E-11	7.01 (11706,123,285,21)
GO:0051301	cell division	3.43E-11	5.01 (11706,189,334,27)
GO:0022402	cell cycle process	1.53E-10	10.83 (11706,352,43,14)
GO:0043284	biopolymer biosynthetic process	2.23E-9	3.65 (11706,113,881,31)
GO:0006118	electron transport	7.3E-9	5.10 (11706,44,991,19)
GO:0007051	spindle organization and biogenesis	2.83E-7	48.49 (11706,17,71,5)
GO:0009058	biosynthetic process	4.32E-7	1.71 (11706,702,995,102)
GO:0006120	mitochondrial electron transport, NADH to ubiquinone	4.63E-7	5.06 (11706,35,991,15)
GO:0000226	microtubule cytoskeleton organization and biogenesis	2.36E-6	22.48 (11706,44,71,6)
GO:0007017	microtubule-based process	2.64E-6	10.37 (11706,132,77,9)
GO:0006412	translation	1.32E-5	2.56 (11706,149,891,29)
GO:0048015	phosphoinositide-mediated signaling	3.61E-5	20.11 (11706,41,71,5)
GO:0006091	generation of precursor metabolites and energy	5.31E-5	2.36 (11706,145,991,29)
GO:0007089	traversing start control point of mitotic cell cycle	7.65E-5	43.09 (11706,5,163,3)
GO:0010498	proteasomal protein catabolic process	9.95E-5	3.00 (11706,75,989,19)
GO:0043161	proteasomal ubiquitin-dependent protein catabolic process	9.95E-5	3.00 (11706,75,989,19)
GO:0006122	mitochondrial electron transport, ubiquinol to cytochrome c	1.63E-4	30.27 (11706,4,290,3)

Footnote to figure. Panel A shows a graphical representation of the portion of the GO hierarchy that includes the terms significantly over-represented in the list of genes over-expressed in correlation with carotid IMT values. Color is a function of the p-value of GO term enrichment by mHG analysis and is coded following the scale depicted in the figure. Panel B enlists in a tabular form the GO terms significantly enriched in the list of genes over-expressed in presence of higher TNF-alpha concentrations. Enrichment is defined as (b/n) / (B/N), where N: total number of genes, B: total number of genes associated with a specific GO term, n: number of genes in the 'target set', and b: number of genes in the 'target set' associated with a specific GO term. The reported p-values are not corrected for multiple testing.

Figure 42: Molecular pathways significantly over-expressed in presence of higher IMT values after age-adjustment


A.
OXIDATIVE PHOSPHORYLATION
Complex I
Complex II
Complex III
Complex IV
Complex V
NADH dehydrogenase (Thermus thermophilus)
Membrane domain (Hydrophobic arm)
Hydrophilic Domain (Peripheral arm)
Intermembrane space
Inner mitochondrial membrane
Mitochondrial matrix
Fumarate reductase (E.coli)
Quinone pool
Cytochrome bc1 complex (bovine)
Cytochrome c oxidase (bovine)
Proton channel
ATP synthase (Escherichia coli)
F0 unit
F1 unit
NADH dehydrogenase
Succinate dehydrogenase / Fumarate reductase
Cytochrome c reductase
Cytochrome c oxidase
Cytochrome c oxidase, cbb3-type
Cytochrome bd complex
F-type ATPase (Bacteria)
F-type ATPase (Eukaryotes)
V-type ATPase (Prokaryotes)
V-type ATPase (Eukaryotes)
B.
CELL CYCLE
Growth factor
Growth factor withdrawal
DNA damage checkpoint
MAPK signaling pathway
Apoptosis
Cohesin
Separin
Securin
Ubiquitin mediated proteolysis
R-point (START)
DNA
DNA biosynthesis
S-phase proteins
ORC (Origin Recognition Complex)
MCM (Mini-Chromosome Maintenance) complex
04110 7/11/07
G1
S
G2
M

A.

Enrichment = 4.54 (15354,105,998,31)
mHG p-value = 6.39*10-12

B.

Enrichment = 6.35 (15354,109,244,11)
mHG p-value = 1.42*10-5

Footnote to figure: Graphical representation of the two molecular pathways most correalted with IMT values as provided in the KEGG database. Genes highlighted in red are over-expressed in smokers compared to non-smokers. Enrichment is defined by mHGM analysis as (b/n) / (B/N), where N: total number of genes, B: total number of genes included in a specific KEGG pathway, n: number of genes in the 'target set', and b: number of genes in the 'target set' included in the specific KEGG pathway.

8 Discussion

White blood cells, especially lympho-monocytes, are high involved in the genesis and progression of atherosclerosis. Despite this well known paradigma, only a very limited experience has been made trying to identify disese biomarkers directly related to the phenotype of these cells. In the current set of studies we provide exploratory data (both from an initial pilot dataset and a pivotal, larger population) that consistenlty show the presence of pro-inflammatory signatures in gene expression profiles of circulating PBMCs in presence of exposure to CVD risk factors in volunteers of young-adult age, without evident clinical diseases and characterized by a low CVD risk profile.

These data provide for the first time a coherent framework that can be used as exploratory finding to build further gene expression signature profiling study for the identification of potential biomarkers of CVD risk.

In our population, gene expression profile of PBMCs varies significantly according to several parameters, including gender and age.

To verify whether PBMCs expression profile could provide intelligible information on inflammatory processes, we verified how the transcriptome of these cells could be influenced by the WBC environment itself. We therefore investigated the correlation between gene expression profile and WBC population composition. As already mentioned in the results, the amount of RNA hybridized on the array is constant and therefore not influenced by the presence of a greater or smaller PBMC count. What influences the profile is the cytokine environment to which the PBMCs are exposed, which is on the contrary high dependent upon the relative proportion of circulating PBMCs over the other WBC components, especially neutrophiles. In presence of a relative imbalance toward neutrophiles, which are mainly deputed to the defense against bacterial infections, PBMC expression profile was indeed characterized by an over-expression of genes associated to the immune response to bacteria. Conversely, the relative abundance of PBMCs was associated with an over-expression of genes involved in MCH calls are response, which is the typical cellular response to viral exposure. In both cases, the down-regulation of the opposite signature pattern was also present, further supporting the capacity of this approach to grasp subtle biological differences.

In presence of well known CVD risk factors, such as LDL-cholesterol and smoking, PBMCs consistently show a pro-inflammatory profile, observed both at the level of groups of transcripts differentially regulated and over-represented molecular pathways.

Even more interestingly, different CVD risk factors appear to be associated to different gene expression profiles, although all in the direction of bursting the inflammatory response. Indeed, we observed that exposure to cigarette smoking produces a significant over-expression of genes involved in the innate immunity and the first stages of the inflammatory process. This type of immunity is more implicated than cellular-mediated immunity in responding to chemical stimuli. On the contrary, the presence of significantly higher (although close to the range of clinical normality) concentrations of LDL-cholesterol appears to be associated to a different pro-inflammatory profile, closer to the cellular-mediated immunitary response. Although it could be argued that there are random and assay-related factors of variability that can differentiate the two profiles without the existence of a true biological difference, the statistical signal behind these two profiles appears to be sufficiently strong and consistent independently of the second level analysis used. In addition to the analyses here reported, we also compared the two expression profiles with the expression profiles collected in the Molecular Signature database (MSig). A search for similar profiles in the database, identified several different molecular signatures characteristics of the exposure of in-vitro cells to chemical agents or generic stimuli such as toxins in the first case, and the response to bacteria or cytokines of the cellular mediated response in the latter.

Finally, the co-presence of exposure to both CVD risk factors was characterized by an additive effect on gene expression where characteristics of both profiles were identifiable.

It is of interest to highlight that this striking molecular signature was observed in subjects who –for the same sample size- did not show any clinically meaningful difference in CVD risk factors. Besides, also hs-CRP resulted perfectly overlapping in values between the two groups, although it is widely considered a bio-marker of the low-grade inflammation induced by CVD risk factors.

Also in this case, gene expression profile appeared consistent with this finding when the transcriptome of PBMCs was correlated with circulating hs-CRP concentrations. Hs-CRP in fact resulted not associated with a specific gene expression profile, representative of CVD –related inflammation. On the other hand, a pro-inflamamtory profile was instead observed in correlation with circulating levels of other cytokines, like TNF-alpha and IL6 [not shown].

Although some preliminary data of correlation between gene expression in circulating cells and CVD risk factors are already available in the literature, no study tried to verify whether the presence of vascular damage (even at a early stage) was associated with a specific expression profile. From this point of view, it should be noted that several CVD biomarkers such as hs-CRP are only poorly associated with vascular markers of the atherosclerotic burden, like carotid IMT and vascular Flow-Mediated Dilation (FMD).

In a bivariate analysis we observed that a higher IMT (although within the boundaries of normality) was associated with a specific gene expression profile. The same profile was observed (even more strongly) after adjustment for relevant clinical confounders by a multivariate model. In addition, in a secondary analysis, the random sampling of less than 50% of the total sample showed, in correlation analysis with IMT, a similar result with similar level of statistical significance [data not shown], thus showing a relevant consistency of the biological signal

The profile associated with high IMT appears to be characterized by a dramatic over-expression of genes involved in the cell cycle and in the mitochondrial respiratory chain. Taken as such these data appear difficult to be put into a clear clinical or patho-physiological perspective. Both of these signatures can be considered typical of activated cells that are moving toward self-replication and activation. Even sticking to this conservative explanation, it is clear that the fact that inflammatory cells of an healthy individual are consistently and highly activated in concordance with the presence of a sign of vascular organ damage it is of high interest. In addition, increased expression of respiratory chain proteins are also a typical feature of cells with pro-oxidant unbalance of the redox system. The increase in generation of oxidative stress is indeed a cornerstone of CVD inflammation.

Unfortunately the descriptive nature of cross-sectional studies and array profiling do not allow to further speculate over this intense statistical signal. Further experiments are therefore needed to establish the direction of these changes and their impact on cellular phenotype.

However, it appears consistent that the expression profile of circulating PBMCs of subjects exposed – in a relatively limited extent- to CVD risk factors are collectively characterized by pro-inflammatory features. This inflammatory signature appears, in our preliminary experience, to be different in presence of different CVD risk factors. Although this dataset is not sufficient to confirm whether this features are universally identifiable and whether different risk factors can trigger different cellular responses, it provides a pilotal data framework that support the interest for further investigations in this area of research.

The present data have been collected using a high-throughput technique for gene expression profiling called microarray. Microarrays become possible after the extensive search of the genome project that provided sufficient meta-data for almost all the human transcripts. Microarrays are complex not in terms of laboratory execution, but foremost in terms of interpretation. Several of the common statistical rules do not apply in presence of huge amount of data of skewed and inconsistent distribution. A typical analysis of microarray data is devoted to decrease data dimensionality, usually comparing gene-by-gene the expression in two different groups of samples. The p-value of the comparison (or any other figure of merit) is then used to isolate a group of differentially expressed genes using more or less arbitrary thresholds. The list of selected genes is then usually explored gene-by-gene searching for biologically interesting genes. This approach is prone to selection bias, since we are naturally more interested in interesting genes than in unknown or apparently unrelated trascript extracting selectively nicer (but less informative) results.

The analysis techniques applied in this set of study is one of the strongest points of the project, as they preserve all the unique characteristics of microarray (for instance the fact of using always all the measured genes to extract results) and provide a more objective way of classifying the biological relevance of results.

In brief, all bivariate analyses have been performed using correlation analysis or non-parametric comparisons to ensure a reliable proportion of significance among different genes. The list of all assayed genes has then been resorted based on p-value as figure of merit of the correlation/ comparison. No direct analysis of the bivariate list of genes has been provided in order to avoid identification bias. The overall gene set has been therefore tested for presence of enrichment of all the known groups of genes classified in the GO DAVID database. The list of genes was searched to identify the relative position in the ranking of each gene of each GO group. Then the distribution of the relative ranking of all genes in each GO term has been tested against what expected by random allocation. In presence of an high prevalence of genes at higher or lower end of the ranking spectrum in one GO group, more than what significantly plausible by chance, the group has been considered significantly over-enriched and a p-value has been assigned. The same type of analysis was conducted on the molecular pathways stored in the KEGG database. The two databases are completely independent and representative of different type of gene associations. Therefore the consistency of results between the two is a valuable indicator of the presence of a true biological signature.

On the other hand, the study is prone –by design- to several weaknesses: first, it is cross-sectional and therefore no causality can be assessed between the clinical phenotypes and the gene signatures identified; then, the relatively small sample size does not allow to identify more subtle differences between signatures or signatures of other clinical parameters; third, a second, independent population is needed to confirm all the findings.

These limitations are implicit in the exploratory nature of the project. Said that, it should on the other hand noted that several signs of internal consistency are present, as reviewed in this discussion section.

Therefore, this study can be considered the first structured dataset suggesting that the exposure to CVD risk factors is associated to a significant change in gene expression profile of circulating PBMCs, which appear to switch over a pro-inflammatory phenotype. This phenomenon is perfectly in line with the inflammatory hypothesis of atherogenesis and provides further insights into the very early steps of the bursting of the inflammatory response that accompanies CVD origin and progression. The further definition of these pro-inflammatory signatures is of high relevance, as they can constitute significant biomarkers of biological response to CVD risk exposure, helping to identify which subjects are responding to a pro-atherogenic environment in the least adaptative way: with disease progression. Pivotal confirmation of these findings, exploration of the effect of exposure to treatment and cohort follow-up of these signatures appear to be the necessary steps to be taken in order to procede in the direction of candidate discovery for pro-atherogenic PBMC inflammation.

9 References

1 Castelli WP. Cholesterol and lipids in the risk of coronary artery disease--the Framingham Heart Study. The Canadian journal of cardiology 1988; 4 Suppl A:5A-10A.

2 Yano K, McGee D, Reed DM. The impact of elevated blood pressure upon 10-year mortality among Japanese men in Hawaii: the Honolulu Heart Program. Journal of chronic diseases 1983; 36(8):569-79.

3 S Debey, U Schoenbeck, M Hellmich, BS Gathof, R Pillai, T Zander, JL Schultze. Comparison of different isolation techniques prior gene expression profiling of blood derived cells: impact on physiological responses, on overall expression and the role of different cell types. The pharmacogenomics journal 2004; 4(3):193-207.

4 Castelli WP. Lipids, risk factors and ischaemic heart disease. Atherosclerosis. 1996;124(suppl):S1-S9.

5 National Cholesterol Education Program (NCEPT), National Heart, Lung, and Blood Institute. Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). National Institutes of Health (NIH) Publication No. 02-5215; September 2002

6 Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure (JNC 7). NIH Publication No. 04-5230.

7 American Diabetes Association (ADA) statements. Hypertension Management in Adults With Diabetes Diabetes Care 2004 27: 65-67.

8 American Diabetes Association (ADA) statements. Dyslipidemia Management in Adults With Diabetes. Diabetes Care 2004 27: 68-71

9 Law MR, Wald NJ, Morris JK. The performance of blood pressure and other cardiovascular risk factors as screening tests for ischaemic heart disease and stroke. J Med Screen 2004;11:3–7.

10 Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther 2001; 69(3):89-95.

11 H Chon, MC Verhaar, HA Koomans, JA Joles, B Braam. Role of circulating karyocytes in the initiation and progression of atherosclerosis. Hypertension 2006; 47(5):803-10.

12 L Rainen, U Oelmueller, S Jurgensen et al. Stabilization of mRNA expression in whole blood samples. Clinical chemistry 2002; 48(11):1883-90.

13 T Reichert, M DeBruyere, V Deneys et al. Lymphocyte subset reference ranges in adult Caucasians. Clinical immunology and immunopathology 1991; 60(2):190-208.

14 S Debey, T Zander, B Brors, A Popov, R Eils, JL Schultze. A highly standardized, robust, and cost-effective method for genome-wide transcriptome analysis of peripheral blood applicable to large-scale clinical trials. Genomics 2006; 87(5):653-64.

15 P Sethu, LL Moldawer, MN Mindrinos et al. Microfluidic isolation of leukocytes from whole blood for phenotype and gene expression analysis. Analytical chemistry 2006; 78(15):5453-61.

16 PM Ridker, MJ Stampfer, N Rifai. Novel risk factors for systemic atherosclerosis: a comparison of C-reactive protein, fibrinogen, homocysteine, lipoprotein(a), and standard cholesterol screening as predictors of peripheral arterial disease. Jama 2001; 285(19):2481-5.

17 SE Nissen, EM Tuzcu, P Schoenhagen et al. Statin therapy, LDL cholesterol, C-reactive protein, and coronary artery disease. N Engl J Med 2005; 352(1):29-38.

18 TA Pearson, GA Mensah, RW Alexander et al. Markers of inflammation and cardiovascular disease: application to clinical and public health practice: A statement for healthcare professionals from the Centers for Disease Control and Prevention and the American Heart Association. Circulation 2003; 107(3):499-511.

19 JP Cobb, MN Mindrinos, C Miller-Graziano et al. Application of genome-wide expression analysis to human health and disease. Proc Natl Acad Sci U S A 2005; 102(13):4801-6.

20 Herder C, Baumert J, Thorand B, Martin S, Lowel H, Kolb H, Koenig W. Chemokines and incident coronary heart disease: results from the MONICA/KORA Augsburg case-cohort study, 1984–2002. Arterioscler Thromb Vasc Biol 2006;26:2147–52.

21 Rothenbacher D, Muller-Scholze S, Herder C, Koenig W, Kolb H. Differential expression of chemokines, risk of stable coronary heart disease, and correlation with established cardiovascular risk markers. Arterioscler Thromb Vasc Biol 2006;26:194–9.

22 JP Radich, M Mao, S Stepaniants et al. Individual-specific variation of gene expression in peripheral blood leukocytes. Genomics 2004; 83(6):980-8.

23 AR Whitney, M Diehn, SJ Popper, AA Alizadeh, JC Boldrick, DA Relman, PO Brown. Individuality and variation in gene expression patterns in human blood. Proceedings of the National Academy of Sciences of the United States of America 2003; 100(4):1896-901.

24 JJ Eady, GM Wortley, YM Wormstone, JC Hughes, SB Astley, RJ Foxall, JF Doleman, RM Elliott. Variation in gene expression profiles of peripheral blood mononuclear cells from healthy volunteers. Physiological genomics 2005; 22(3):402-11.

25 IC Macaulay, P Carr, A Gusnanto, WH Ouwehand, D Fitzgerald, NA Watkins. Platelet genomics and proteomics in human health and disease. J Clin Invest 2005; 115(12):3370-7.

26 RJ Feezor, HV Baker, M Mindrinos *et al.* Whole blood and leukocyte RNA isolation for gene expression analyses. Physiol Genomics 2004; 19(3):247-54.

27 S Debey, U Schoenbeck, M Hellmich, BS Gathof, R Pillai, T Zander, JL Schultze. Comparison of different isolation techniques prior gene expression profiling of blood derived cells: impact on physiological responses, on overall expression and the role of different cell types. The pharmacogenomics journal 2004; 4(3):193-207.

28 EC Baechler, FM Batliwalla, G Karypis *et al.* Expression levels for many genes in human peripheral blood cells are highly sensitive to ex vivo incubation. Genes and immunity 2004; 5(5):347-53.

29 MC Muller, K Merx, A Weisser, S Kreil, T Lahaye, R Hehlmann, A Hochhaus. Improvement of molecular monitoring of residual disease in leukemias by bedside RNA stabilization. Leukemia 2002; 16(12):2395-9.

30 L Rainen, U Oelmueller, S Jurgensen *et al.* Stabilization of mRNA expression in whole blood samples. Clinical chemistry 2002; 48(11):1883-90.

31 P Stordeur, L Zhou, B Byl, F Brohet, W Burny, D de Groote, T van der Poll, M Goldman. Immune monitoring in whole blood using real-time PCR. J Immunol Methods 2003; 276(1-2):69-77.

32 JP Cobb, MN Mindrinos, C Miller-Graziano et al. Application of genome-wide expression analysis to human health and disease. Proc Natl Acad Sci U S A 2005; 102(13):4801-6.

33 S Debey, T Zander, B Brors, A Popov, R Eils, JL Schultze. A highly standardized, robust, and cost-effective method for genome-wide transcriptome analysis of peripheral blood applicable to large-scale clinical trials. Genomics 2006; 87(5):653-64.

34 P Sethu, LL Moldawer, MN Mindrinos *et al.* Microfluidic isolation of leukocytes from whole blood for phenotype and gene expression analysis. Analytical chemistry 2006; 78(15):5453-61.

35 T Reichert, M DeBruyere, V Deneys *et al.* Lymphocyte subset reference ranges in adult Caucasians. Clinical immunology and immunopathology 1991; 60(2):190-208.

36 C Palmer, M Diehn, AA Alizadeh, PO Brown. Cell-type specific gene expression profiles of leukocytes in human peripheral blood. BMC genomics [electronic resource] 2006; 7:115.

37 T Chtanova, RA Kemp, AP Sutherland, F Ronchese, CR Mackay. Gene microarrays reveal extensive differential gene expression in both CD4(+) and CD8(+) type 1 and type 2 T cells. J Immunol 2001; 167(6):3057-63.

38 Y Tang, H Xu, X Du *et al.* Gene expression in blood changes rapidly in neutrophils and monocytes after ischemic stroke in humans: a microarray study. J Cereb Blood Flow Metab 2006; 26(8):1089-102.

39 RL Khan, GE Gonye, G Gao, JS Schwaber. A universal reference sample derived from clone vector for improved detection of differential gene expression. BMC genomics [electronic resource] 2006; 7:109.

40 AD Hershey, Y Tang, SW Powers, MA Kabbouche, DL Gilbert, TA Glauser, FR Sharp. Genomic abnormalities in patients with migraine and chronic migraine: preliminary blood gene expression suggests platelet abnormalities. Headache 2004; 44(10):994-1004.

41 Y Tang, DL Gilbert, TA Glauser, AD Hershey, FR Sharp. Blood gene expression profiling of neurologic diseases: a pilot microarray study. Arch Neurol 2005; 62(2):210-5.

42 AH Iglesias, S Camelo, D Hwang, R Villanueva, G Stephanopoulos, F Dangond. Microarray detection of E2F pathway activation and other targets in multiple sclerosis peripheral blood mononuclear cells. J Neuroimmunol 2004; 150(1-2):163-77.

43 Y Tang, MB Schapiro, DN Franz et al. Blood expression profiles for tuberous sclerosis complex 2, neurofibromatosis type 1, and Down's syndrome. Ann Neurol 2004; 56(6):808-14.

44 F Borovecki, L Lovrecic, J Zhou et al. Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease. Proc Natl Acad Sci U S A 2005; 102(31):11023-8.

45 J Kalman, K Kitajka, M Pakaski, A Zvara, A Juhasz, G Vincze, Z Janka, LG Puskas. Gene expression profile analysis of lymphocytes from Alzheimer's patients. Psychiatric genetics 2005; 15(1):1-6.

46 RH Segman, N Shefi, T Goltser-Dubner, N Friedman, N Kaminski, AY Shalev. Peripheral blood mononuclear cell gene expression profiles identify emergent post-traumatic stress disorder among trauma survivors. Molecular psychiatry 2005; 10(5):500-13, 425.

47 K Morita, T Saito, M Ohta, T Ohmori, K Kawai, S Teshima-Kondo, K Rokutan. Expression analysis of psychological stress-associated genes in peripheral blood leukocytes. Neuroscience letters 2005; 381(1-2):57-62.

48 NC Twine, JA Stover, B Marshall et al. Disease-associated expression profiles in peripheral blood mononuclear cells from patients with advanced renal cell carcinoma. Cancer Res 2003; 63(18):6069-75.

49 ME Burczynski, NC Twine, G Dukart et al. Transcriptional profiles in peripheral blood mononuclear cells prognostic of clinical outcomes in patients with advanced renal cell carcinoma. Clin Cancer Res 2005; 11(3):1181-9.

50 MS Forrest, Q Lan, AE Hubbard et al. Discovery of novel biomarkers by microarray analysis of peripheral blood mononuclear cell gene expression in benzene-exposed workers. Environmental health perspectives 2005; 113(6):801-7.

51 H Hakonarson, US Bjornsdottir, E Halapi et al. Profiling of genes expressed in peripheral blood mononuclear cells predicts glucocorticoid sensitivity in asthma patients. Proc Natl Acad Sci U S A 2005; 102(41):14789-94.

52 TM Bull, CD Coldren, M Moore, SM Sotto-Santiago, DV Pham, SP Nana-Sinkam, NF Voelkel, MW Geraci. Gene microarray analysis of peripheral blood cells in pulmonary arterial hypertension. American journal of respiratory and critical care medicine 2004; 170(8):911-9.

53 SY Yu, YW Hu, XY Liu, W Xiong, ZT Zhou, ZH Yuan. Gene expression profiles in peripheral blood mononuclear cells of SARS patients. World J Gastroenterol 2005; 11(32):5037-43.

54 F Diaz-Mitoma, I Alvarez-Maya, A Dabrowski et al. Transcriptional analysis of human peripheral blood mononuclear cells after influenza immunization. J Clin Virol 2004; 31(2):100-12.

55 CF Ockenhouse, WB Bernstein, Z Wang, MT Vahey. Functional genomic relationships in HIV-1 disease revealed by gene-expression profiling of primary human peripheral blood mononuclear cells. The Journal of infectious diseases 2005; 191(12):2064-74.

56 MB Hansen, L Skov, T Menne, J Olsen. Gene transcripts as potential diagnostic markers for allergic contact dermatitis. Contact dermatitis 2005; 53(2):100-6.

57 Z Liu, RW Yelverton, B Kraft, SB Tanner, NJ Olsen, TM Aune. Highly conserved gene expression profiles in humans with allergic rhinitis altered by immunotherapy. Clin Exp Allergy 2005; 35(12):1581-90.

58 N Olsen, T Sokka, CL Seehorn, B Kraft, K Maas, J Moore, TM Aune. A gene expression signature for recent onset rheumatoid arthritis in peripheral blood mononuclear cells. Ann Rheum Dis 2004; 63(11):1387-92.

59 LF Bovin, K Rieneck, C Workman et al. Blood cell gene expression profiling in rheumatoid arthritis. Discriminative genes and effect of rheumatoid factor. Immunol Lett 2004; 93(2-3):217-26.

60 FM Batliwalla, EC Baechler, X Xiao et al. Peripheral blood gene expression profiling in rheumatoid arthritis. Genes and immunity 2005; 6(5):388-97.

61 FK Tan, X Zhou, MD Mayes, P Gourh, X Guo, C Marcum, L Jin, FC Arnett, Jr. Signatures of differentially regulated interferon gene expression and vasculotrophism in the peripheral blood cells of systemic sclerosis patients. Rheumatology (Oxford, England) 2006; 45(6):694-702.

62 L Bennett, AK Palucka, E Arce, V Cantrell, J Borvak, J Banchereau, V Pascual. Interferon and granulopoiesis signatures in systemic lupus erythematosus blood. J Exp Med 2003; 197(6):711-23.

63 AK Palucka, JP Blanck, L Bennett, V Pascual, J Banchereau. Cross-regulation of TNF and IFN-alpha in autoimmune diseases. Proc Natl Acad Sci U S A 2005; 102(9):3372-7.

64 V Pascual, F Allantaz, E Arce, M Punaro, J Banchereau. Role of interleukin-1 (IL-1) in the pathogenesis of systemic onset juvenile idiopathic arthritis and clinical response to IL-1 blockade. J Exp Med 2005; 201(9):1479-86.

65 EE Mannick, JC Bonomolo, R Horswell et al. Gene expression in mononuclear cells from patients with inflammatory bowel disease. Clinical immunology (Orlando, Fla 2004; 112(3):247-57.

66 N Kaushik, D Fear, SC Richards et al. Gene expression in peripheral blood mononuclear cells from patients with chronic fatigue syndrome. Journal of clinical pathology 2005; 58(8):826-32.

67 RM Rutherford, J Kehren, F Staedtler, SD Chibout, JJ Egan, M Tamm, JJ Gilmartin, MH Brutsche. Functional genomics in sarcoidosis--reduced or increased apoptosis? Swiss Med Wkly 2001; 131(31-32):459-70.

68 RA Ahokas, KJ Warrington, IC Gerling et al. Aldosteronism and peripheral blood mononuclear cell activation: a neuroendocrine-immune interface. Circ Res 2003; 93(10):e124-35.

69 ML Jison, PJ Munson, JJ Barb et al. Blood mononuclear cell gene expression profiles characterize the oxidant, hemolytic, and inflammatory stress of sickle cell disease. Blood 2004; 104(1):270-80.

70 A Gladkevich, HF Kauffman, J Korf. Lymphocytes as a neural probe: potential for studying psychiatric disorders. Progress in neuro-psychopharmacology & biological psychiatry 2004; 28(3):559-76.

71 JE Blalock. Shared ligands and receptors as a molecular mechanism for communication between the immune and neuroendocrine systems. Ann N Y Acad Sci 1994; 741:292-8.

72 Y Tang, H Xu, X Du et al. Gene expression in blood changes rapidly in neutrophils and monocytes after ischemic stroke in humans: a microarray study. J Cereb Blood Flow Metab 2006; 26(8):1089-102.

73 Y Tang, AC Nee, A Lu, R Ran, FR Sharp. Blood genomic expression profile for neuronal injury. J Cereb Blood Flow Metab 2003; 23(3):310-9.

74 Y Tang, A Lu, BJ Aronow, FR Sharp. Blood genomic responses differ after stroke, seizures, hypoglycemia, and hypoxia: blood genomic fingerprints of disease. Ann Neurol 2001; 50(6):699-707.

75 DF Moore, H Li, N Jeffries *et al.* Using peripheral blood mononuclear cells to determine a gene expression profile of acute ischemic stroke: a pilot investigation. Circulation 2005; 111(2):212-21.

76 JW Lampe, SB Stepaniants, M Mao, JP Radich, H Dai, PS Linsley, SH Friend, JD Potter. Signatures of environmental exposures using peripheral leukocyte gene expression: tobacco smoke. Cancer Epidemiol Biomarkers Prev 2004; 13(3):445-53.

77 NA Cacalano, D Sanden, JA Johnston. Tyrosine-phosphorylated SOCS-3 inhibits STAT activation but binds to p120 RasGAP and activates Ras. Nat Cell Biol 2001; 3(5):460-5.

78 H Chon, CA Gaillard, BB van der Meijden *et al.* Broadly altered gene expression in blood leukocytes in essential hypertension is absent during treatment. Hypertension 2004; 43(5):947-51.

79 V Wibaut-Berlaimont, AM Randi, V Mandryko, MW Lunnon, DO Haskard, RP Naoumova. Atorvastatin affects leukocyte gene expression in dyslipidemia patients: in vivo regulation of hemostasis, inflammation and apoptosis. J Thromb Haemost 2005; 3(4):677-85.

80 J Davignon, R Laaksonen. Low-density lipoprotein-independent effects of statins. Curr Opin Lipidol 1999; 10(6):543-59.

81 PA Kiener, PM Davis, JL Murray, S Youssef, BM Rankin, M Kowala. Stimulation of inflammatory responses in vitro and in vivo by lipophilic HMG-CoA reductase inhibitors. Int Immunopharmacol 2001; 1(1):105-18.

82 R Tibshirani, T Hastie, B Narasimhan, G Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci U S A 2002; 99(10):6567-72.

83 SB Glueck, VJ Dzau. Physiological genomics: implications in hypertension research. Hypertension 2002; 39(2 Pt 2):310-5.

84 M Pravenec, C Wallace, TJ Aitman, TW Kurtz. Gene expression profiling in hypertension research: a critical perspective. Hypertension 2003; 41(1):3-8.

85) Zavaroni I, Bonora E, Pagliara M, Dall'Aglio E, Luchetti L, Buonanno G, et al. Risk factors for coronary artery disease in healthy persons with hyperinsulinemia and normal glucose tolerance. N Engl J Med 1989;320:702-6.

86) Zavaroni I, Bonini L, Gasparini P, Barilli AL, Zuccarelli A, Dall'Aglio E, et al. Hyperinsulinemia in a normal population as a predictor of non-insulin-dependent diabetes mellitus, hypertension, and coronary heart disease: the Barilla factory revisited. Metabolism 1999;48:989-94.

87) Ardigo D, Stüehlinger M, Franzini L, Valtueña S, Piatti PM, Pachinger O, Reaven GM, Zavaroni I. ADMA is independently related to flow-mediated vasodilation in subjects at low cardiovascular risk. Eur J Clin Invest. 2007;37(4):263-9.

88 Valtuena S,Numeroso F, ArdigoD, et al. Relationship among leptin, insulin, body composition and liver steatosis in non-diabetic moderate drinkers with normal transaminase levels. Eur J Endocrinol 2005;153(2):283–90.

89 Ardigo D, Numeroso F, Valtuena S, et al. Hyperinsulinemia predicts hepatic fat content in healthy individuals with normal transaminase concentrations. Metabolism 2005;54(12):1566–70.

90 O'Leary DH, Polak JF, Kronmal RA, et al. Carotid-artery intima and media thickness as a risk factor for myocardial infarction and stroke in older adults. Cardiovascular Health Study Collaborative Research Group. *N Engl J Med.* 1999;340:14-22

91 Redberg RF, Vogel RA, Criqui MH, et al. Task Force #3 - What is the spectrum of current and emerging techniques for the noninvasive measurement of atherosclerosis? *J Am Coll Cardiol.* 2003;41:1886-98.

92 http://www1.qiagen.com/HB/RNeasyMiniKit_EN

93http://www.chem.agilent.com/Library/usermanuals/Public/G4140-90050_Two-Color_GE_5.7.pdf

94 Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation 1998;97:1837-47.

95 Demirtas E, Krishnamurthy S, Tulandi T. Elevated serum beta-human chorionic gonadotropin in nonpregnant conditions. Obstet Gynecol Surv. 2007;62(10):675-9;

96 Zhou D, Wang J, Zapala MA, Xue J, Schork NJ, Haddad GG. Gene expression in mouse brain following chronic hypoxia: role of sarcospan in glial cell death. Physiol Genomics 2008;32(3):370-9.

97 http://www.broad.mit.edu/gsea/msigdb/

98 Gray D, Gray M, Barr T. Innate responses of B cells. Eur J Immunol 2007;37(12):3304-10.

99 Niiro H, Clark EA. Regulation of B-cell fate by antigen-receptor signals. Nat Rev Immunol 2002;2(12):945-56.