Emilia Maria Cristina MAZZA

Università degli Studi di Milano
Dottorato in Medicina Molecolare e Traslazionale

Curriculum di Genomica, Proteomica e Tecnologie Correlate

Ciclo XXVI

Anno Accademico 2012/2013

Dottorando: Emilia Maria Cristina MAZZA

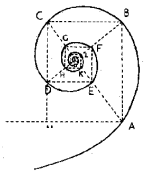
Identification of gene regulatory modules in a human model of physiological inflammation: a bioinformatics approach

CICLO XXVI

Tutore: Prof.ssa Cristina BATTAGLIA
Correlatore: Prof. Silvio BICCIATO

Direttore del Dottorato: Ch.mo Prof. Mario Clerici

UNIVERSITÀ DEGLI STUDI DI MILANO



SCUOLA DI DOTTORATO IN MEDICINA MOLECOLARE

CICLO XXVI

Anno Accademico 2012/2013

TESI DI DOTTORATO DI RICERCA

Settore BIO/10

Identification of gene regulatory modules in a human model of physiological inflammation: a bioinformatics approach

Dottorando : Emilia Maria Cristina Mazza

Matricola R09200

TUTORE: Prof.ssa Cristina BATTAGLIA

CO-TUTORE: Prof. Silvio BICCIATO

DIRETTORE DEL DOTTORATO: Ch.mo Mario CLERICI

SOMMARIO

Negli organismi sani l'infiammazione è il primo meccanismo di difesa contro i microorganismi patogeni. I monociti e i macrofagi hanno un ruolo chiave in questo processo e una alterazione della loro attività è alla base di molte condizioni patologiche. Per questi motivi, capire i meccanismi molecolari dell'attivazione monocito-macrofagica è il primo passo per lo studio delle malattie infiammatorie ed eventualmente, per lo sviluppo di strategie terapeutiche. Nonostante ciò non sono ancora del tutto chiari i meccanismi di attivazione funzionale dei monociti e la loro caratterizzazione fenotipica durante la polarizzazione M1/M2. In questa tesi viene descritta la costruzione in vitro di un modello umano della reazione infiammatoria e lo sviluppo di un metodo bioinformatico mirato alla ricostruzione delle reti di regolazione genica alla base di questo processo. In particolare, monociti umani primari, isolati dal sangue di soggetti sani, sono stati messi in coltura ed esposti ad una serie di stimoli microambientali (chemiochine, citochine, temperatura, molecole di derivazione batterica, ecc.) propri dell'infiammazione fisiologica e i loro profili di espressione genica sono stati monitorati per 48 h tramite analisi con la tecnologia dei microarray. L'analisi computazionale inizia con l'identificazione dei geni differenzialmente espressi durante l'arco temporale e che, attraverso analisi di arricchimento, sembrano essere coinvolti nel processo infiammatorio. Questi geni possono essere considerati i regolatori del processo infiammatorio e per questo sono stati utilizzati come geni chiave per l'identificazione dei moduli di regolazione genica. Grazie a questo approccio bioinformatico è stato possibile identificare i geni caratterizzanti le differenti fasi della risposta infiammatoria ed è stato possibile ricostruire i loro moduli di connessione. Infine, è stata fatta una comparazione tra i risultati ottenuti dall'analisi computazionale del modello e dati pubblici ottenuti con la tecnica di immunoprecipitazione della cromatina associata al sequenziamento (ChIP-seq).

ABSTRACT

In healthy organisms, inflammation is the first defense mechanism and monocytes and macrophages are among the key players of this process. Since a alteration of the activity of these cell populations is at the base of several pathological conditions, elucidating the molecular mechanisms of monocyte/macrophage activation represents a major step to study inflammatory disorders and, eventually, develop new therapeutic strategies. However, these mechanisms and their interplay during monocyte/macrophage activation still remain poorly characterized. Here, we report the setup of a physiological inflammation model, based on human primary cells, and of a bioinformatics approach that allow studying the development of the inflammatory reaction during its entire course and elucidating networks of molecular interactions which are at the basis of this process. Specifically, human blood monocytes isolated from blood of normal healthy donors have been cultured and exposed to a combination of factors reproducing physiological inflammatory conditions and their gene expression profiles monitored during a time course of 48 hours. The computational process starts with the identification of those genes whose expression changes during the time course and that, through enrichment analysis, appear to be involved in inflammatory processes. These genes can be considered as controllers of the process and thus are further used as regulators to identify regulatory modules. Using these computational methods we have been able to obtain genes that characterize the various steps of the inflammatory process and to reconstruct their connection modules. Finally, to validate our results we performed a comparison between data from the physiological inflammation model and data obtained from ChIP-seq that combines chromatin immunoprecipitation with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins

CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

1.1 Omics data and their analysis

The word that changed the science's lexis and that gave the origins to a new era, was coined in a McDonald's Raw Bar by Dr. Thomas H. Roderick, a geneticist at the Jackson Laboratory, Bar Harbor, ME, in 1986. He was looking with some colleagues for the name of a new journal, and he was looking for a word that would encompass sequencing, mapping, and new technologies, a word that could describe the genome as a functioning whole beyond just single genes or sequences spread around a chromosome (Kuska B., 1998). He came up with the word "genomics". From that moment on, begins the "omics" era and a large amount of words with suffix –*omics* spread out, every time to describe a big field in life sciences that focuses on large-scale data/information to understand life summarized in "omics" such as proteomics, genomics, metabolomics, and transcriptomics. In the case of transcriptomics the word encompass the study of the set of all RNA molecules, their structures and functions. Unlike the genome, the transcriptome can change with external conditions and stimuli, because it includes all mRNA transcripts and reflects the genes that are actively expressed in a given cell at a given moment under a particular condition. The study of transcriptomics, also referred to as expression profiling, is the study of the expression level of mRNAs through

high-throughput techniques like microarrays. The basic principle of microarray technology is complementary hybridization of nucleotides, as explained by the Watson–Crick double helical model of DNA. The mRNA from a given cell line or tissue is used to generate a labeled sample, sometimes termed the *target*, which is hybridized in parallel to a large number of DNA sequences, immobilized on a solid surface in an ordered topology. Although academic groups and commercial suppliers have developed many different microarray systems, in the most commonly used technology the arrayed material, generally termed the *probe* (being the equivalent to the probe used in a northern blot analysis), is an oligonucleotide sequence. In oligonucleotide arrays, short 20–25mers are synthesized in situ, either by photolithography onto silicon wafers (high-density-oligonucleotide arrays from Affymetrix) or by ink-jet technology (developed by Rosetta Inpharmatics and licensed to Agilent Technologies). Microarrays allow the simultaneous measurement of tens of thousands of messenger RNA (mRNA) transcripts, this is why they are so powerful and why the use of high-throughput techniques has become routine in genome-wide studies. Public databases of microarray gene expression data have been quickly growing as the use of high-throughput techniques has become very common. Major repositories of microarray data, as Gene Expression Omnibus (Edgar et al.,2002) and ArrayExpress (Brazma et al.,2003), are exceptionally rich mines of genomic information about the immune cells and

exploiting their content represents an unprecedented opportunity to improve the interpretation and validation of expression studies.

Meta-analysis of large microarray expression datasets allows researchers to confirm biological hypotheses, formulated from results of a study, in a relatively inexpensive way, i.e. using data independently obtained in another laboratory, without the need of novel experiments. Meta-analysis also offers the opportunity of re-analyzing formerly available data, in combination with new samples and new computational methods, thus increasing the reliability and robustness of results. However, performing a meta-analysis of independent microarray studies requires to carefully handle the heterogeneity of array designs, which complicates cross-platform integration. Moreover, although the power of microarrays and their capability in studying gene expression profile in many physiological or pathological conditions has been largely demonstrated, it's very difficult to extrapolate the entire amount of data they contain and to interpret it to the best.

The interpretation of the immunological data is complicated by the complexity of the immune system itself: several different cell types activated by several different stimuli and under several different conditions cooperate to make sure that pathogens are recognized and neutralized, and that infected

cells are killed. New molecules and new molecular mechanisms involved in pathogen recognition and immune response are still being discovered. For this reason, when evaluating an immune response, one should consider taking an approach for its characterization that elevates the study of the single components of the system (e.g., genes, proteins) to higher hierarchies, as entire genomic regions, groups of co-expressed genes, functional modules, and networks of interactions. Since the networks of signals and relations among genes and regulators (e.g., transcription factors (TFs)) control the development of physio-pathological states, understanding how elementary biological objects act together and interact is fundamental for the advancement of biological knowledge.

In this context, the goal is to analyze the correlation between genes, between their products, and the mechanisms of interaction that determine the physiological state of a cell or of a tissue and to recapitulate regulatory interactions of biological systems into mathematical models. However, the inference of gene regulatory networks is a challenging task because of incomplete knowledge of the involved molecules, the combinatorial nature of the problem and the fact that, often, available data are limited and inaccurate. Moreover, the presence of several feedback loops among these regulatory processes makes their organization and functioning very complex and this level of complexity cannot be addressed

using standard computational methods. Currently there are some methods to infer gene interactions using microarray expression profiles, as for example: i) Ingenuity Pathway Analysis (IPA; www.ingenuity.com) that allows to identify the most relevant signaling and metabolic pathways, molecular networks, and biological functions from a list of genes ii) Graphite web (Sales et al., 2013; www.graphiteweb.bio.unipd.it) that is a public web server for the analysis and visualization of biological pathways using high-throughput gene expression data, iii) Context Likelihood Relatedness (CLR; Faith et al., 2007) that infers regulatory interactions between transcription factors and their targets using a compendium of gene expression profiles and iv) The Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE; Margolin et al., 2006) that is based on a mutual information approach and it allows to reconstruct regulatory networks in mammalian cells. However, none of these allows identifying networks and modules of gene interaction using an integrated approach, which takes advantage of information from differentially expressed genes and functional enrichments. Therefore, standard methodologies for the analysis of gene expression profiles, which aim at identifying relevant genes from the statistical analysis of microarray signals, seem to be severely limited in unveiling the mechanisms governing the transcriptional cascade.

1.2 Monocytes, macrophages and dendritic cells

Monocytes are produced by the bone marrow from hematopoietic stem cell precursors called monoblasts. Monoblasts are characterized by the expression of surface marker CD34. Monocytes circulate in the bloodstream for about one to three days and then move into tissues where they can differentiate into macrophages in order to replenish the pool of tissue macrophages following homeostatic loss, or to become inflammatory macrophages upon tissue damage. Monocytes have been considered as the systemic reservoir of myeloid precursors for the renewal of tissue macrophages and antigen-presenting Dendritic Cells (DC) (*Figure 1.1*). Macrophages are specialized phagocytic cells that attack and destroy foreign substances, cellular debris, infectious microbes and cancer cells; they also stimulate lymphocytes and other immune cells to respond to pathogens. Each type of macrophage, depending on where it reside, has a specific name: macrophages in connective tissue are often called histiocytes, macrophages in the skin and in the liver are respectively known as Langerhans and Kupffers cells, osteoclasts are macrophages in the bone and macrophages in the brain are microglial cells (Castagna et al., 2012).

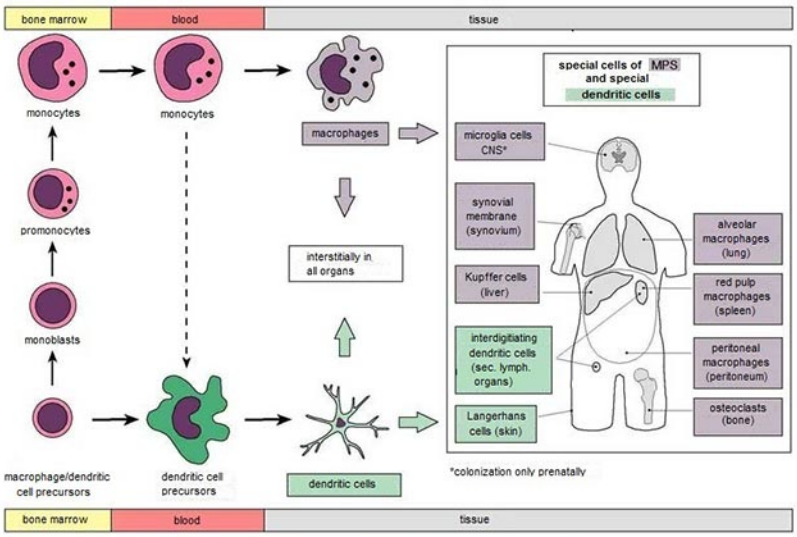


Figure 1.1: Genealogy and nomenclature of monocyte, macrophages and dendritic cells.

The differentiation of monocytes into macrophages is regulated by several cytokines, like interleukins and interferons. Depending on the stimuli that drive the differentiation, macrophage can be divided in distinct subpopulations, and they are being referred to generally as polarized macrophages. In particular two main macrophage phenotypes have been proposed: the inflammatory (M1) and the repair or anti-inflammatory phenotype (M2). The classically activated macrophages (M1) develop in response to inflammatory factors like the Th1 cytokine IFN-γ, LPS and TNF-α, and mediate resistance against intracellular parasites and tumors (Sica et al., 2012). Alternative M2 macrophages

are activated by Th2 cytokines, or FcγR binding in the presence of TLR agonist, or glucocorticoids and anti-inflammatory molecules (M2a, M2b, M2c respectively), and they take part in parasite clearance, dampening of inflammation, tissue remodelling, and tumor promotion (Matzinger et al.,2007). Several in vitro and in vivo studies suggest that polarised M1 and M2 macrophages can switch from a phenotype to the other. A controversial issue is whether M1 and M2 macrophages consist of phenotypically distinct subpopulations that can serve different functions, or the same cells can shift from one to another functional phenotype based on microenviromental signals. While in several pathological conditions the latter seems to be the case (obesity-induced insulin resistance, type-2 diabetes, atherosclerotic lesions, cancer, endotoxin tolerance) (*Figure 1.2*), if M1 and M2 macrophages can undergo dynamic transitions between different functional states during a “physiological” inflammatory response is still unknown.

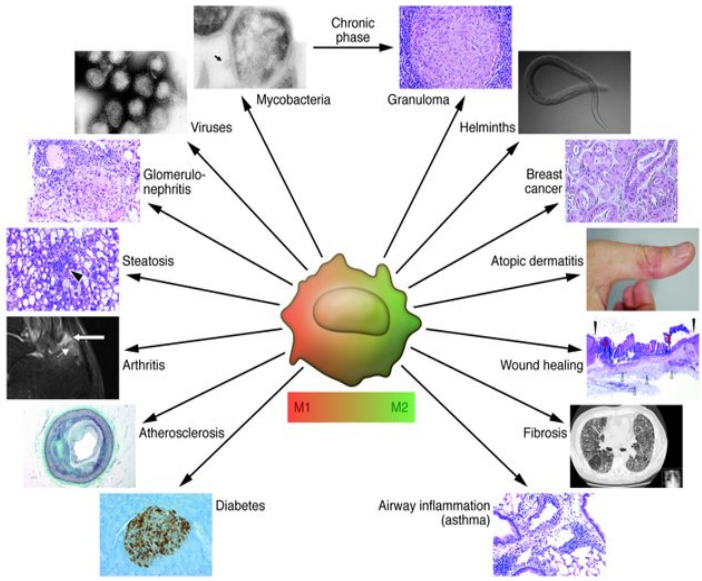


Figure 1.2: Schematic representation of macrophage plasticity and polarization in pathology (Sica et al., 2012)

Monocytes can differentiate also into dendritic cells and can be found as immature DCs in the blood. Dendritic cells are a critical link in the immune system. Their role is to monitor the body seeking out and capturing foreign invaders, called antigens (whether these are bacteria, viruses, or dangerous toxins), afterwards dendritic cells convert them into smaller pieces and display the antigenic fragments on their cell surfaces. Then DCs move to lymph nodes or to the spleen where they activate other cells of the immune system against the invaders, in particular B cells and killer T cells that, respectively, make antibodies to neutralize the invaders and

activate specific attacks to destroy them (Clark, 1997). There are two major types of DCs in human blood and these are myeloid DC (mDC) CD11c+ BDCA1+ and plasmacytoid DC (pDC) CD11c− BDCA2+ (CD303+). The two subsets differ in the expression of highly conserved microbial pattern recognition receptors, known as Toll-like receptors (TLR), but both are able to induce the stimulation of naive T cells. The dendritic cells also participate in the mechanism known as tolerance that restrains the T cells activity. If dendritic cells are too tolerant, this can create a permissive environment for chronic infectious agents or for tumors. If DCs are not enough tolerant, the immune system can lead to autoimmune diseases (rheumatoid arthritis and multiple sclerosis) (Ganguly et al., 2013). Chronic inflammatory diseases and autoimmune disorders are characterized by persistent inflammation and immune activation. This chronic activation is probably triggered by exogenous stimuli (infection or mechanical stress) which facilitate the wrong recognition of self-antigens. In physiological conditions the inflammatory reaction is generally stimulated at the tissue level as a response to an event of danger (for example, a bacterial infection). After the contact with the microorganism, tissue produces factors such as chemokines that attract monocytes to the site of inflammation. In the tissue, activated monocytes begin the inflammation reaction that, usually, is followed by the destruction and the phagocytosis of the microorganism. This kind of inflammatory reaction is considered as a mechanism of innate immunity

which is quicker than the mechanism of adaptive immunity. When the injurious stimulus is cleared, tissue produces anti-inflammatory cytokines. This kind of cytokines can induce the switch of macrophages polarization from a pro-inflammatory phenotype to an anti-inflammatory phenotype. In fact, in this stage macrophages begin to produce growth factors, tissue factors and anti-inflammatory cytokines, including TGF-β, responsible for the reconstruction and tissue remodeling. There is no information at present on the features of the entire course of the inflammatory reaction and on the possibility that the same cell population could be first polarized towards an effector inflammatory program and subsequently re-polarized to the deactivation program.

1.3 Aim of the study

The research activity presented in this thesis wants to contribute to filling the gaps in the bioinformatics analysis of microarray data of immune cells.

Specifically, we focused on

i) the development and the application of computational strategies for the meta-analysis of gene expression data of the immune system cells, obtained from public repositories.

ii) the development of a bioinformatics approach that allow studying the inflammatory reaction during its entire course

iii) the development of a bioinformatics pipeline to reconstruct gene regulatory networks and decipher transcriptional modules in this process.

MATERIALS AND METHODS

This chapter contains a description of gene expression and ChIP-seq data used in this thesis. Each dataset is fully reviewed and detailed. Following paragraphs describe: i) how different datasets were combined to construct and analyze a proprietary meta-database of gene expression profiles of monocytes macrophages and dendritic cells; ii) the tools used to set up a pipeline to analyze an *in vitro* model of physiological inflammation. Finally, the methods used for ChIP-seq analysis are briefly presented.

2.1 Model of physiological inflammation

We developed an *in vitro* model of physiological inflammation, based on human primary cells, that could allow us to study the development of the inflammatory reaction during its entire course, thus opening the possibility of accurately characterizing the development and regulation of human macrophage functions. The *in vitro* model of physiological inflammation consists in a 48-hour culture of primary human monocytes isolated from 9 healthy donors. During the culture, monocytes are exposed sequentially to a series of stimuli (*Figure 2.1)* that mimic conditions in a simplified microenvironment that develops in inflamed tissue (Italiani et al., 2011). In particular, cells were cultured at the following experimental conditions:

- time 0: beginning of the culture at 37° C and addition of CCL2;
- time +2 hours: removal of CCL2, addition of LPS and increasing temperature to 39° C;
- time +3 hours: addition of TNF;
- time +7 hours: addition of IFNγ;
- time +14 hours: removal of inflammatory stimuli, and addition of IL10. Lowering the temperature to 37° C;
- time +24 hours: IL10 removal, addition of TGFβ, and maintenance of temperature at 37° C.

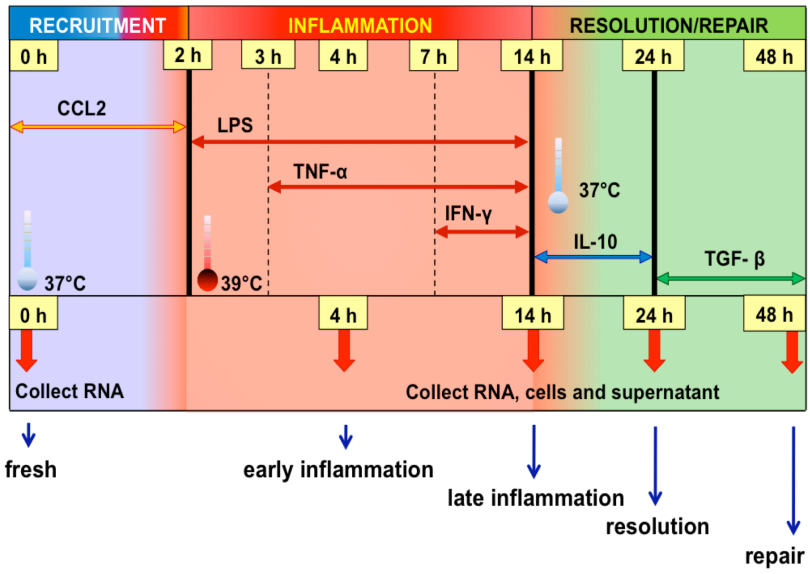


Figure 2.1: Graphic representation of the in vitro model of inflammation based on human primary monocytes. Freshly isolated monocytes are first exposed to the chemokine CCL2 for 2 h at 37°C. At 2 h, monocytes are exposed to LPS and the temperature is raised to 39°C. Temperature is then kept at 39°C until 14 h while TNF-α and IFN-γ are added at 3 and 7 hours, respectively. At 14 h all the inflammatory stimuli are washed off, the temperature brought down to 37°C and fresh medium containing IL-10 added. Finally, monocytes are exposed to TGF-β at 24 hours.

Total RNA was extracted from monocytes of 12 individual donors (3 for the "early" series: 0, 2.0-3.5 h; and 9 for the "late" series: 0, 4-48 h), using Qiagen miRNeasy kit (Qiagen), quantified spectrophotometrically (ND-1000, NanoDrop Technologies, Wilmington, DE), and checked for integrity by microcapillary electrophoresis (Agilent 2100 Bioanalyzer; Agilent Technologies, Palo Alto, CA). Samples were prepared starting from 0.1-1 µg total RNA, using the GeneChip®3' IVT Express kit or the GeneChip® One Cycle cDNA Synthesis kit (Affymetrix, Santa Clara, CA), with identical results. Biotinylated cRNAs (15 µg) were fragmented and hybridized for 16 h at 45°C onto GeneChip® HG-U133 Plus 2.0 Arrays (Affymetrix). After washing and staining, arrays were scanned with the GeneChip® Scanner 3000 7G (Affymetrix) and fluorescent images were acquired and analyzed using GCOS software (Affymetrix) to generate a total of 60 raw intensity files (CEL files). The Affymetrix microarray HG-U133 Plus 2.0. allows simultaneous monitoring of 47,401 genes and human transcripts for a total of 54,675 probe set.

2.2 Genomic data

The large amount of genomic data produced using microarray technology and DNA sequencing induced the creation of public repositories where storing and making publicly available to the scientific community this huge amount of data. Genomic data obtained during experiments designed to study a particular biological pathway, contains indeed a wealth of information not necessarily used in the original study and therefore available to other researchers for validating and confirming biological hypotheses. The larger repository of gene expression data is Gene Expression Omnibus (GEO; www.ncbi.nlm.nih.gov/geo/) at the National Center for Biotechnology Information (NCBI, Bethesda, MD, USA). Actually, the database contains 1.031.592 samples divided into 42.965 experiments and obtained using 12.258 different platforms. We have exploited this very useful resource to validate the results obtained through the analysis of the model. Gene expression data stored in this database are organized into three categories:

- GSE that indicates the code of the series;
- GSM that indicates the code of the sample;
- GPL that indicates the code of the platform.

2.2.1 Meta-database of gene expression profiles of monocytes macrophages and dendritic cells.

Datasets of human primary monocytes, macrophages, and dendritic cells (DC) were retrieved from Gene Expression Omnibus. Twenty-four series comprising 474 samples were downloaded from GEO of which 303 samples were used and organised in a proprietary database using the software A-MADMAN (Bisognin et al 2009; *table 2.1*).

GEO series	Platform	Total samples in series	Samples used in our study	Reference
GSE4984	HG-U133 Plus2.0	12	6	Fulcher et al., 2006
GSE5099	HG-U133A	30	14	Martinez et al., 2006
GSE5547	HG-U133 Plus2.0	24	6	Humphrey et al., 2007
GSE6965	HG-U133 Plus2.0	4	4	Mezger et al., 2008
GSE7509	HG-U133 Plus2.0	26	26	Dhodapkar et al., 2007
GSE7568	HG-U133 Plus2.0	25	25	Gratchev et al., 2008
GSE7807	HG-U133 Plus2.0	8	4	Woszczek et al., 2008
GSE8286	HG-U133A	9	9	Liu et al., 2008
GSE8515	HG-U133A	15	15	Jura et al., 2008
GSE8608	HG-U133 Plus2.0	6	1	Hofer et al., 2008
GSE8658	HG-U133 Plus2.0	63	30	Szatmari et al., 2007
GSE9080	HG-U133Av2	6	3	---
GSE9874	HG-U133A	60	11	Hägg et al., 2008
GSE9946	HG-U133A	12	12	Popov et al., 2008
GSE9988	HG-U133 Plus2.0	62	58	Dower et al., 2008
GSE10856	HG-U133 Plus2.0	4	4	Chang et al., 2008
GSE11393	HG-U133Av2	9	3	Llaverias et al., 2008
GSE11430	HG-U133 Plus2.0	10	10	Maouche et al., 2008
GSE11864	HG-U133 Plus2.0	10	10	Hu et al., 2008
GSE12108	HG-U133 Plus2.0	14	13	Butchar et al., 2008
GSE12773	HG-U133 Plus2.0	10	5	Rate et al., 2009
GSE12837	HG-U133A	24	3	Coppe et al., 2009
GSE13762	HG-U133 Plus2.0	15	15	Széles et al., 2009
GSE14419	HG-U133Av2	16	16	---

Table 2.1: Complete list of the datasets used in this study and their sources. Genome-wide expression levels and meta-information of 303 samples were organized in a proprietary meta-database using A-MADMAN.

A-MADMAN is an open source web application which allows the automatic import of metadata from GEO records into a local relational database, the subsequent manual annotation and selection of samples through user-defined tags, and the selection of samples to be analyzed using a complex logical query on tags (*Figure 2.2*).

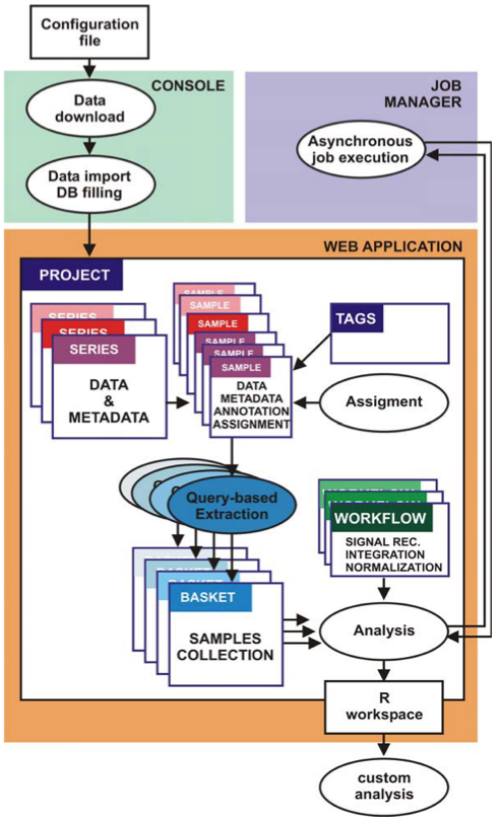


Figure 2.2: A-MADMAN architecture. A-MADMAN includes console, job server and web-application. The console allows data retrieval, import and database filling. The web application is the user friendly and collaborative core of the system allowing data inspection,

annotation and analysis. A Project is a collection of samples, series, tags, baskets and analyses owned by a user or by a group of users. Series and samples data and metadata come from GEO. The user can create an annotation system based on tags and assign samples to individuals. Queries on Boolean combinations of annotation tags are used to select and extract groups of samples, giving rise to baskets.

All samples have been manually re-annotated and tagged based on the meta-information provided by GEO and by the original publications. In particular, we labeled 62 samples as untreated monocytes and 46 and 20 samples as M1 and M2 activated monocytes/macrophages, respectively. Gene expression profiles have been generated from raw .CEL files using an *ad-hoc* procedure called *Virtual-Chip (*Bisognin et al., 2010; Fallarino et al., 2010) and described in paragraph 2.3.2. The expression matrix has been analyzed to validate the results obtained from the analysis of our inflammation model.

2.2.2 GSE16723

This series contains chromatin immunoprecipitation sequencing (ChIP-seq) data from 10 samples of mouse primary bone marrow-derived macrophages unstimulated or treated with LPS (Barish et al., 2010). The experiment was performed using different antibodies but we focused on sample GSM611116 in which ChIP was carried out with an antibody specific for p65 (a subunit that forms the NF-κB complex). Specifically, bone marrow was purified and differentiated in

DMEM containing 20% fetal bovine serum, 30% L929 conditioned media, and antibiotics for 5 days, then re-plated in macrophage serum free media (Invitrogen) overnight. The sample was treated with LPS for 3 h. Cells were fixed with 2 mM disuccinimidyl glutarate for 30 minutes, then with 1% formaldehyde for 10 minutes, then glycine-quenched and harvested. Following fixation, nuclei were isolated, lysed in buffer containing 1% SDS, 10 mM EDTA, 50 mM Tris-HCl pH 8.0, and protease inhibitors, and sheared with a Diagenode Bioruptor to chromatin fragment sizes of 200 – 1000 base pairs. Chromatin was immunoprecipitated with antibody to p65. Short DNA reads were aligned against the mouse mm9 reference genome using the Illumina Pipeline Suite v1.4. Peak detection was performed with the HOMER software suite (http://biowhat.ucsd.edu/homer/). IgG antibody was used as a negative control.

2.2.3 GSE32324

This series contains chromatin immunoprecipitation sequencing (ChIP-seq) data from THP1 cells (Iglesias MJ et al., 2012), a human monocytic cell line. We used samples GSM869213 and GSM869215 for our analysis. THP-1 cells were maintained in culture in RPMI 1640 medium containing 10% fetal bovine serum, 1mM sodium pyruvate, 100 units/ml penicillin and 100 µg/ml streptomycin at 37°C with 5% CO2. For differentiation of THP-1 cells into macrophages a protocol

using conditioned media was used. To confirm differentiation of THP-1 monocytes into macrophages CD11b expression was measured using FACS and quantitative polymerase chain reaction (qPCR). THP-1 macrophages were stimulated with 1ug/ml LPS (LPS-stimulated) from Escherichia coli for 2 hours to induce an acute inflammatory response. To confirm an induction of the inflammatory response in the THP-1 macrophages, TNF mRNA levels were measured in control and LPS-stimulated samples. Immunoprecipitation was performed with specific antibodies raised against Sp1 transcription factor. DNA-protein complexes were eluted, treated with RNase for 4–6 hours at 45°C and Proteinase K overnight at 65°C. DNA was extracted by phenol/chloroform/isoamyl alcohol extraction, purified and resuspended in water. Sequencing was performed on the Illumina Genome Analyzer I (SNP&SEQ technology platform, Uppsala University, Sweden). Sequence reads were 35 bases or longer but truncated to 35 bases to ensure base quality over the entire read. Reads were aligned to the human reference genome (GRCh37/hg19) with Burrows-Wheeler Alignment tool (BWA; Li H and Durbinv R, 2009)). Peak detection was performed with MACS (Zhang et al., 2008).

2.3 Signal quantification

In Affymetrix microarrays, the expression signal of each gene is quantified summarizing the intensities of all oligonucleotides,

i.e. the probes, of a probe set matching a target gene or transcript. The signal can be generated using a series of statistical or model-based algorithms (e.g., MAS5.0, MBEI, RMA, GCRMA) that transform the intensity level into a number representing the gene expression level. In this thesis expression levels were quantified using directly RMA for data produced using the same microarray platform whereas a novel procedure was developed for integration of data produced by different types of microarray.

2.3.1 Robust Multichip Average

Robust Multichip Average (RMA) (Irizarry et al., 2003) consists of three steps: background adjustment, quantile normalization, and summarization. The RMA method begins by computing background-corrected perfect match (PM) intensities for each perfect match cell on every array. These PM intensities are computed in such a way that all background-corrected values must be positive and log-2 transformed. Then, PM intensities are normalized using the quantile normalization method developed by Bolstad et al. (Bolstad et al., 2003). Following quantile normalization, an additive linear model is fit to the normalized data to obtain an expression measure for each probe on each array. Finally, signals are summarized using the median polish algorithm. The output is a matrix of intensities where each column corresponds to a chip and each row is a probe set.

2.3.2 Virtual-Chip procedure

When data have been produced using the same microarray platform, the normalization-quantification can be directly performed using RMA. If instead different platforms were used to obtain the data, as in the case of our meta-database, RMA cannot be directly applied but it's necessary to apply a data combination strategy. Data combination integrates multiple datasets directly at the level of raw data and generates a unique matrix of gene expression signals. The direct merging of raw data from different studies is applicable only when expression profiles have been obtained using the same array technology (e.g. Affymetrix, Agilent, Illumina, etc.) and requires an *ad-hoc* normalization step. We called this procedure *Virtual-Chip*. In *Virtual-Chip*, raw expression data (i.e., CEL files) obtained from at least two different platforms are integrated using an approach inspired by the generation of custom Chip Definition Files, CDFs (Dai et al., 2005; Ferrari et al., 2007). In custom CDFs, probes matching the same transcript, but belonging to different probes sets, are aggregated into putative custom-probe sets, each one including only those probes with a unique and exclusive correspondence with a single transcript. Similarly, probes matching the same transcript but located at different coordinates on different type of arrays may be merged in custom-probe sets and positioned in a *virtual-grid* whose geometry can be arbitrarily defined (*Figure 2.3*).

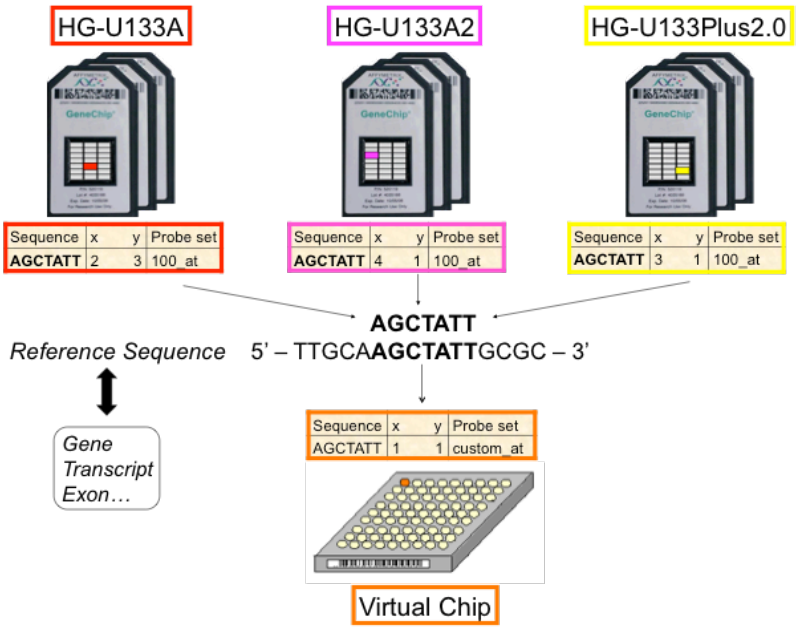


Figure 2.3: Virtual-Chip procedure.

As for any other microarray geometry, this *virtual grid* may be used as a reference to create a *virtual CDF* file containing the probes of the *Virtual-Chip* and their coordinates on the *virtual platform*. The probes included in the *virtual CDF* are those shared among the platforms of interest, with the additional condition of generating custom probe set of at least 4 probes. The *virtual CDF* can be derived from any custom CDF, e.g., those developed by Dai and publicly accessible at the Molecular and Behavioral Neuroscience Institute Microarray Lab website. Finally, the *virtual CDF* can be used as the geometry file in RMA as far as the original CEL files are

properly re-mapped to match the topology described in the *virtual CDF*. Re-mapped CEL files, called *virtual CEL file*, are homogeneous in terms of platform and gene expression data can be generated with a single step of background correction, normalization and summarization directly from the fluorescence signals of all microarrays composing the meta-dataset. CEL file re-mapping requires re-defining:

- the content of the [HEADER] field, i.e., all physical coordinates (total number of cells containing the probes, indicated by *Cols, Rows*, *TotalX*,and *TotalY*, and localization of the 4 border cells) and the name of the platform;

- all data contained in the [INTENSITY] field, i.e., physical localization (*X* e *Y*) and fluorescence intensity (*MEAN*) of any probe.

2.4 Pipeline to reconstruct module networks during physiological inflammation

The transcriptional network of macrophage activation can exhibit many distinct steady-states which are associated with tissue- and infection-specific macrophage functions. The transcriptional response is dynamic and is characterized by clusters of gene activated during time and controlled by different combinations of transcription factors. Time course

genomic data, coupled with bioinformatics tools for their analysis, represents a promising starting point to elucidate the transcriptional network of macrophage activation, to identify key regulators and their functions during the different stages of innate immune response. In this paragraph are described the different tools used for the setup of an *ad-hoc* bioinformatics pipeline aimed at reconstructing gene regulatory modules during the macrophages activation and polarization. The computational process starts with the identification of those genes whose expression changes during the time course and that, through enrichment analysis, appear to be involved in inflammatory processes. These genes can be considered as controllers of the process and thus are further used as regulators to identify regulatory modules (*Figure 2.4*).

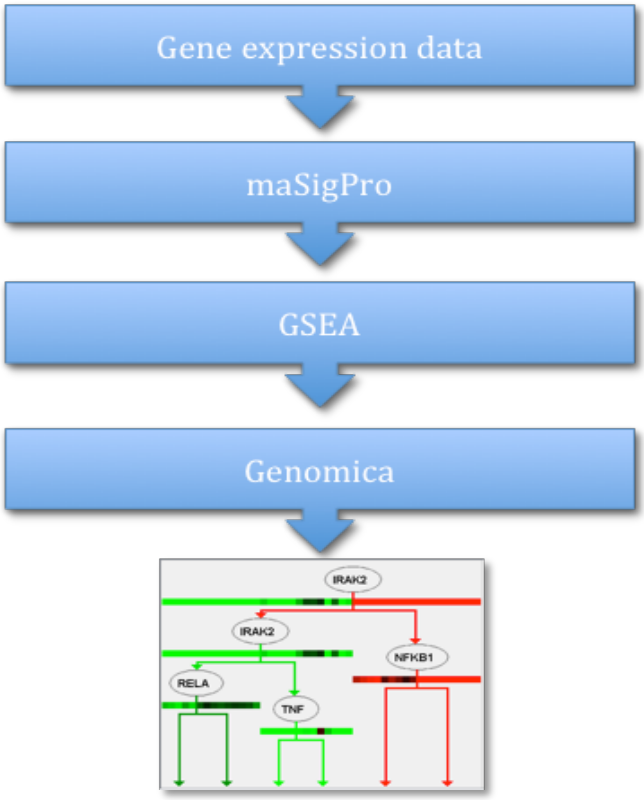


Figure 2.4: bioinformatics pipeline for module network reconstruction.

2.4.1 Microarray Significant Profiles (maSigPro)

Time-course microarray experiments are useful approaches for exploring biological processes because this type of experiments allows scientists to study gene expression changes along time and to evaluate trend differences between various experimental groups. The large amount of data, the presence of several experimental conditions and the dynamic

nature of experiments make the analysis very complex. So recently, many tools have been developed for time-course microarray analysis. In our study genes showing different expression profiles along our time-course experiment have been identified using the microarray Significant Profiles method coded in the maSigPro R package (Conesa et al., 2006). MaSigPro follows a two steps regression strategy to find genes with significant temporal expression changes and significant differences between experimental groups. The method defines a general regression model for the data where the experimental groups are identified by dummy variables. In the first step the procedure adjusts this global model by the least-squared technique to identify differentially expressed genes and selects significant genes applying false discovery rate control procedures. In the second step a variable selection strategy is applied to identify statistically significant profile differences between experimental groups and to find statistically significant different profiles. The coefficients obtained in this second regression model will be useful to cluster together significant genes with similar expression patterns and to visualize the results. The maSigPro package contains different types of regression methods and this permits to choose an adequate regression model for the data. The input is a matrix with gene expression data; the obtained output is a matrix with as many rows as significant genes and as many columns as parameters in the complete regression model.

2.4.2 Gene Set Enrichment Analysis (GSEA)

To understand the biological role of the genes that are significantly modulated during the inflammatory response, each cluster was subjected to an over-representation analysis using the gene set enrichment analysis (GSEA) tool. All genes of the microarray are ranked in a list *L* based on the correlation between their expression and a reference gene profile (i.e. linearly or exponentially increasing in time). Given an a priori defined set of genes *S* (e.g., genes encoding products in a metabolic pathway), the goal of GSEA is to determine whether the members of *S* are randomly distributed throughout *L* or primarily found at the top or bottom rank positions. GSEA is characterized by three key elements:

1. calculation of an Enrichment Score (ES) that reflects the degree to which a set *S* is over-represented at top or bottom of the entire ranked list *L*. The score is calculated by walking down the list *L*, increasing a running-sum statistic when a gene in *S* is encountered and decreasing it when genes not in *S* are encountered. The magnitude of the increment can be equal at every step (classic statistic) or can depend on the correlation of the gene with the reference profile (weighted statistic). The enrichment score is the maximum deviation from zero encountered in the random walk; in the first case it corresponds to a standard Kolmogorov–Smirnov statistic (Hollander and

Wolfe, 1999); in the second case to a weighted Kolmogorov–Smirnov-like statistic;

2. estimation of significance level of ES (nominal p-value) using an empirical gene-set-based permutation test procedure. Random gene sets, size matched to the actual gene set, are created and their enrichment scores calculated. These enrichment scores are used to create a null distribution from which the significance of the actual enrichment score (for the actual gene set) is calculated;

3. adjustment for multiple hypothesis testing is performed by first normalizing the ES for each gene set to account for the size of the set, thus yielding a normalized enrichment score (NES). Then, the FDR (the estimated probability that a set with a given NES represents a false positive finding) corresponding to each NES is computed to control the proportion of false positives.

Gene set enrichment analysis is implemented with Molecular Signature Database (MSigDb). MSigDb is a publicly accessible collection of curated gene sets that is maintained by the GSEA team (www.broadinstitute.org/gsea/msigdb/index.jsp; Subramanian et al., 2005). The MSigDB gene sets are divided into five major collections:

C1: **positional gene sets** for each human chromosome and each cytogenetic band;

C2: curated gene sets from online pathway databases, publications in PubMed, any knowledge of domain experts;

C3: motif gene sets based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat and dog genomes;

C4: computational gene sets defined by expression neighborhoods centered on 380 cancer-associated genes;

C5: GO gene sets consist of genes annotated by the same GO terms.

In this thesis, we used a subset of the C2 collection (version 2.5), i.e. those gene sets derived from BIOCARTA, KEGG, and REACTOME pathways database.

2.4.3 Bayesian networks

In the last step of our pipeline we used an approach based on Bayesian network theory to reconstruct gene regulatory network and their associated modules. Bayesian networks (BNs), belong to the family of probabilistic graphical models (GMs). Each node in the graph represents a random variable, while the edges between the nodes represent probabilistic dependencies among the corresponding random variables. BNs correspond to a GM structure known as a directed acyclic graph (DAG). They enable an effective representation and computation of the joint probability distribution (JPD) over a set

of random variables (Ruggeri et al., 2009). The structure of a DAG is constituted by two entities: nodes or vertices and directed edges. The nodes represent random variables and are usually drawn as circles labeled by the variable names. The edges represent direct dependence among the variables and are usually drawn by arrows between nodes. In particular, an edge from node X_i to node X_j represents a statistical dependence between the corresponding variables. Thus, the arrow indicates that a value taken by variable X_j depends on the value taken by variable X_i, (X_i "influences" X_j). Node X_i is then referred to as a parent of X_j and, similarly X_j is referred to as the child of X_i. More formally, a Bayesian network B is an annotated acyclic graph that represents a JPD over a set of random variables V. The network is defined by a pair B = ⟨G, Θ⟩ where G is the DAG whose nodes X_1, X_2, …,X_n represent random variables, and whose edges represent the direct dependencies between these variables. The graph G encodes independence assumptions, by which each variable X_i is independent of its non descendants given its parents in G. The second component Θ denotes the set of parameters of the network. This set contains the parameter $\theta_{\chi_i|\pi_i} = P_B(\chi_i|\pi_i)$ for each realization χ_i of X_i conditioned on π_i, the set of parents of X_i in *G*. Accordingly, *B* defines a unique JPD over V, namely:

$$P_B(\mathrm{X}_1, \mathrm{X}_2, ..., \mathrm{X}_n) = \prod_{i=1}^{n} P_B(\mathrm{X}_i | \pi_i) = \prod_{i=1}^{n} \theta_{\chi_i | \pi_i}$$

2.4.4 Genomica: the algorithm

Bayesian networks are a powerful approach to learn regulatory networks but, since in a gene expression dataset the number of variables is normally bigger than the number of samples, this method tends to overfit the data. To overcome this problem Segal and his team implemented in Genomica software the module network method, which is a special type of Bayesian network algorithm (Segal et al., 2003). In this method, each module represents a set of variables that share: i) a single variable or a set of variables as their parents and ii) local distributions. Compared to standard Bayesian network algorithms, this design significantly reduces the number of parameters to be learned and consequently leads to more accurate inferences. The process of learning module networks consists of two steps that are called M-step and E-step: the first requires clustering genes into modules and the second inferring the regulation program of each module. Segal et al. designed an expectation-maximization-based learning algorithm that alternates between these two steps. In fact the procedure is iterative, which means that involves the recurrence of these steps. In each iteration, the method searches for a regulation program for each module and then reallocates each gene to the module that best describes its

behaviour. At every iteration, the algorithm gives a score to the steps, called Bayesian score, and the steps are repeated until convergence is reached. The iterative process searches for the model with the highest Bayesian score using an Expectation Maximization algorithm (Dempster et al., 1977).

2.4.5 Genomica: the parameters

The procedure involves loading a matrix containing gene expression levels, the choice of analysis parameters and, in case, insertion of a list of known and/or putative regulators. The algorithm needs to set several parameters:

- initial clustering method (default is Agglomerative, correlation centered);
- maximum number of modules (based on the number of matrix genes);
- maximum number of iterations;
- module merge method;
- scoring method (only Bayesian type is available);
- lookahead depth;
- maximum tree depth (based on how big you want regulation tree);
- minimum experiments for contest;
- candidate regulator genes (it can be load a list of genes or can be choose genes from the all in the matrix);

- experiment split constraints (experiment sample can be divided into different groups based on experiment characteristics);
- known regulator to split on first (it's possible to choose the main regulator of the regulation tree).

Given these inputs, the algorithm searches simultaneously for a partition of genes into modules and for a regulation program for each module that explains the expression behavior of genes in the module. The regulation program of a module specifies the set of regulatory genes that control the module and the mRNA expression profile of the genes in the module as a function of the expression of the module's regulators. The procedure gives as output a list of modules and associated regulation programs. This procedure identifies groups of coregulated genes, their regulators, the behavior of the module as a function of the regulators expression and the conditions under which regulation takes place.

2.5 Pscan: finding over-represented transcription factor binding site motifs in sequences

Once established, through the analysis of the pipeline, that genes in the modules are regulated by the same gene network and have a common biological function, we want to investigate if they are also regulated by the same transcription factors. To perform this analysis we used the online method Pscan

(http://159.149.160.51/pscan/). Pscan is a software tool that scans promoter sequences from co-regulated or co-expressed genes, looking for over- or under-represented motifs describing the binding specificity of known TFs (Zambelli et al., 2009), it provides a quick hints on which factors could be responsible for the patterns of expression observed, or vice versa seem to be avoided. As first step, users have to input a set of gene identifiers and to select the organism of provenience. Then, users have to specify the promoter region to investigate with respect to the transcription start sites (TSSs) of the genes and choose whether the analysis has to be performed with the TFBSs matrices available in the JASPAR or TRANSFAC databases. Given these inputs, for each profile, the average matching score obtained from the input sequence set can be compared to the mean and the standard deviation of the score on the whole genome promoter set. The over- (or under-) representation for each profile is finally assessed with a *z*-test, that associates with each profile the probability of obtaining the same score on a random sequence set (*Figure 2.5*).

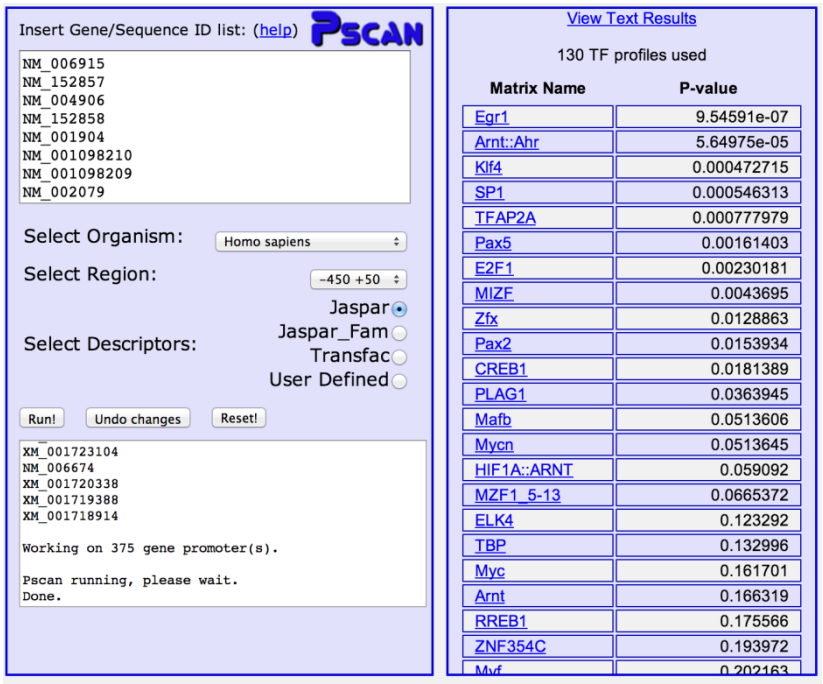


Figure 2.5: Pscan outup interface. Motif profiles are ranked according to their z-test P-value.

2.6 ChIP-seq analysis

ChIP-sequencing, also known as ChIP-seq, is a method used to analyze interactions between DNA and transcription factors or other chromatin-associated proteins. ChIP-seq is a two-step method that combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing (seq) to identify the binding sites of DNA-associated proteins. Briefly, proteins bound to the DNA are fixed with a cross-linked agent, then DNA is fragmented and complexes (protein-DNA) are harvested with target antibodies. Finally, cross-links are

broken and only DNA fragments from binding sites remain. The obtained fragments are sequenced (*Figure 2.6*). Determining how proteins interact with DNA to regulate gene expression is essential for fully understanding many biological processes and disease states (Park, 2010).

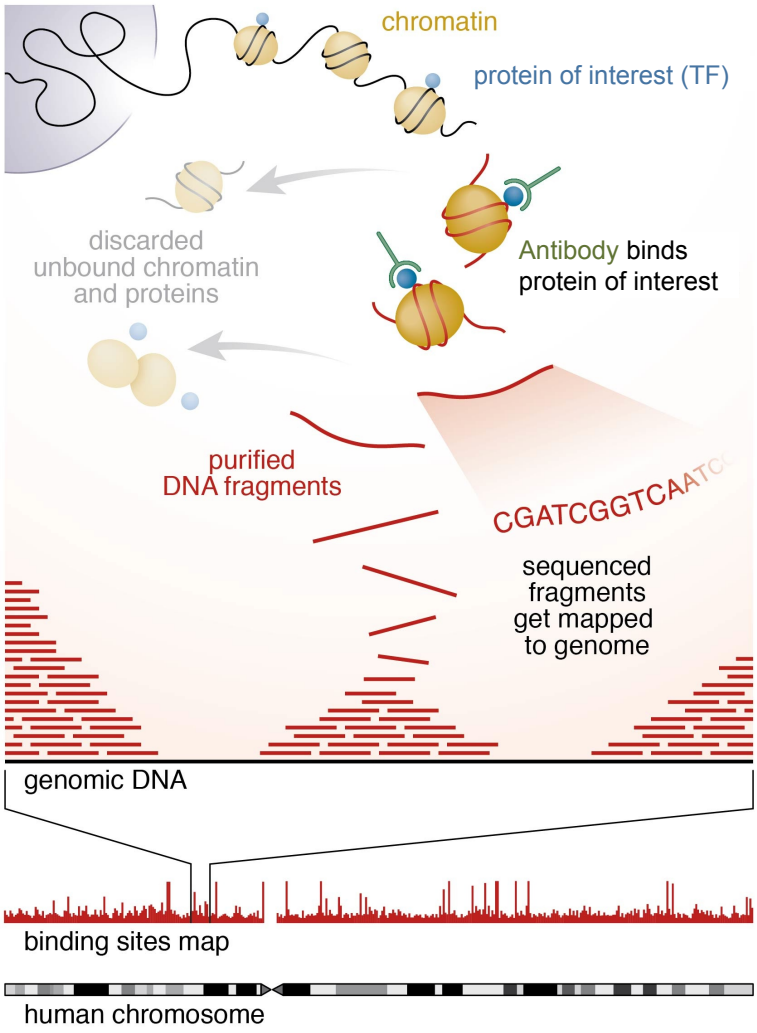


Figure 2.6: ChIP-sequencing workflow.

For our study, datasets of ChIP-seq were downloaded from GEO and were analyzed with methods reported below. We

performed standard ChIP-seq analysis: peak call, peak annotation and motif discovery.

2.6.1 Model-based Analysis of ChIP-seq (MACS)

The first step in a ChIP-seq analysis is the peak calling. We used MACS (Zhang et al., 2008), that is a ChIP-seq peak-finding algorithm. MACS performs removal of redundant reads, performs read-shifting to account for the offset in forward or reverse strand reads and uses control samples and local statistics to minimize bias and calculates an empirical FDR (*Figure 2.7*). MACS can be applied to ChIP-seq experiments without controls, and to those with controls with improved performance. For experiments with a control, MACS linearly scales the total control tag count to be the same as the total ChIP tag count. MACS allows each genomic position to contain no more than one tag and removes all the redundancies, moreover tag distribution along the genome could be modeled by a Poisson distribution. MACS shifts every tag by d/2, it slides 2d windows across the genome to find candidate peaks with a significant tag enrichment (Poisson distribution). The location with the highest fragment pileup, hereafter referred to as the summit, is predicted as the precise binding location. If, instead, a ChIP-seq experiment was performed with controls, MACS empirically estimates the false discovery rate (FDR) for each detected peak using the same procedure employed in the previous ChIP-chip peak finders

MAT (Johnson et al., 2006) and MA2C (Song et al., 2007). At each p-value, MACS uses the same parameters to find ChIP peaks over control and control peaks over ChIP (i.e. a sample swap). The empirical FDR is defined as Number of control peaks / Number of ChIP peaks.

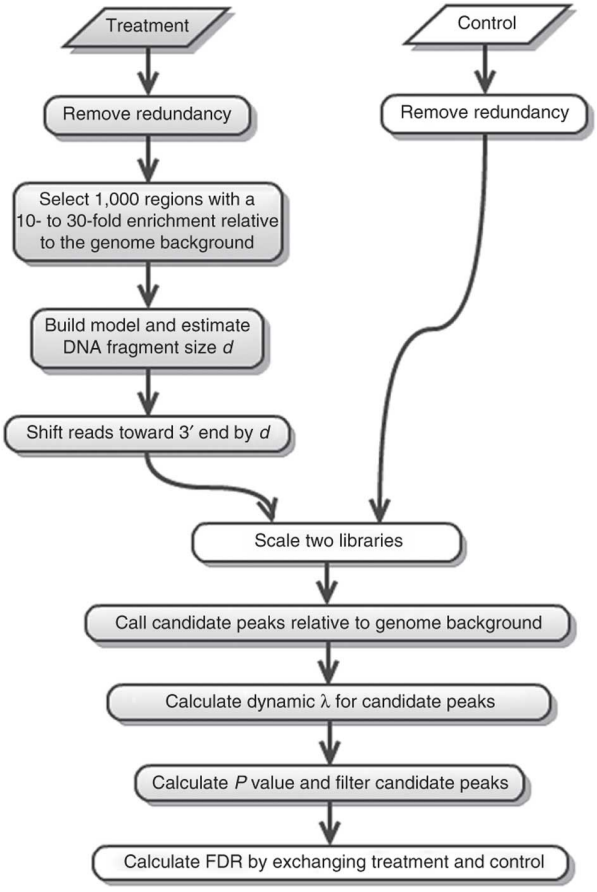


Figure 2.7: MACS workflow (Feng et al., 2012). If the control sample is missing, then the steps shown in white boxes will be skipped (remove redundancy of the control sample, scale two libraries and calculate FDR by exchanging treatment and control).

2.6.2 Genomic Regions Enrichment of Annotations Tool (GREAT)

GREAT is a web tool to perform peak annotation and to analyze the functional significance of cis-regulatory regions identified by localized measurements of DNA binding events across an entire genome (McLean et al., 2010). GREAT input is a set of genomic regions (in our case a set of transcription factor binding site events identified by peak calling analysis). As first step GREAT associates proximal and distal input genomic regions with their putative target genes, then it uses genes annotations from several ontologies to associate genomic regions with annotations. Finally, GREAT calculates statistical enrichments for associations between genomic regions and annotations (*Figure 2.8*).

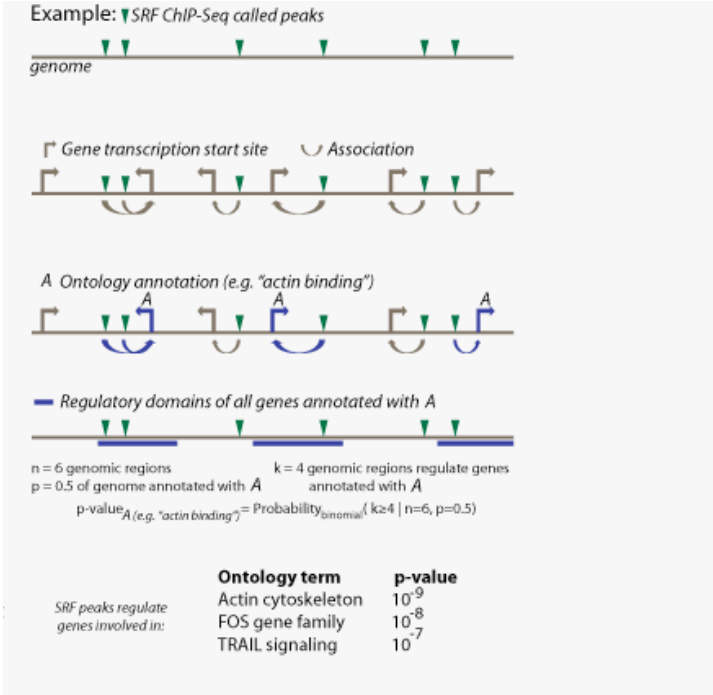


Figure 2.8: GREAT workflow (McLean et al., 2010)

2.6.3 MEME-chip

The MEME-ChIP web service is designed to analyze ChIP-seq peaks. MEME-ChIP represents motifs as position-dependent letter-probability matrices which describe the probability of each possible letter at each position in the pattern (Machanick et al., 2011). MEME-ChIP takes as input a group of DNA sequences (the training set) and outputs as many motifs as requested. MEME-ChIP automatically performs five types of analysis on ChIP-seq regions. (i) *Ab initio* motif discovery

identifies novel sequence patterns (motifs) in the ChIP-seq regions that may be due to TF binding sites. (ii) Motif enrichment analysis looks for enrichment of known TF DNA-binding motifs in the data. (iii) Motif visualization displays the relative locations and binding strengths of TF binding sites in the input regions. (iv) Motif binding strength analysis computes an estimate of the total DNA-binding affinity of each input region for the TF corresponding to each discovered motif. (v) Motif identification compares the *ab initio* motifs to known TF DNA-binding motifs. The output of MEME-ChIP is thus a multifaceted view of the identities, prevalence, DNA-binding patterns and potential interactions of the ChIP-ed TF and its regulatory partners. MEME-ChIP employs three motif discovery algorithms with complementary characteristics. The MEME (Bailey et al., 2006) algorithm uses expectation maximization (EM) to discover probabilistic models of DNA-binding by single TFs or TF complexes. MEME motifs can provide accurate thermodynamic models of TF binding. MEME is complemented by DREME (Bailey, 2011), which uses a simpler, non-probabilistic model (regular expressions) to describe the short binding motifs characteristic of single eukaryotic TFs. DREME is often able to detect very short motifs that are not found by MEME. MEME-ChIP also attempts to identify the motifs found by MEME and DREME by comparing them to a database of known TF motifs using the TOMTOM (Gupta et al., 2007) algorithm. Motif discovery thus identifies novel binding motifs and TFs that are regulatory

partners of the ChIP-ed TF. MEME-ChIP also implements CentriMo or Central Motif Enrichment Analysis that is a tool for inferring direct DNA binding from ChIP-seq data. CentriMo is based on the observation that the positional distribution of binding sites matching the direct-binding motif tends to be unimodal, well centered and maximal in the precise center of the ChIP-seq peak regions. CentriMo takes a set of sequences and plots the occurrence of motifs relative to the ChIP-seq peak. Motifs that occur exclusively at the peak provide good evidence of direct binding, while motifs that do not occur in a consistent position relative to the peak may not bind directly.

RESULTS

3.1 Meta-database of gene expression profiles of immune cells

Twenty-four series comprising 474 gene expression profiles were downloaded from GEO, and were organized in a proprietary meta-database using AMADMAN. Starting from these 474 samples, 303 were selected excluding samples that were related to pathological conditions. Samples were manually re-annotated and tagged based on the meta-information provided by GEO and by the original publications. Taking advantage of the references, we defined a list of 59 tags which were then associated to each sample in order to describe the most peculiar aspects that could influence the gene expression profile.

Tags can be grouped in some major categories, and these are:
- cellular type (e.g. monocyte, macrophage);
- purification method (e.g. immunomagnetic or adhesion purification);
- stimuli, either exogenous(LPS) and endogenous(IL4, INFγ);
- polarization (M1 or M2);
- quality control of the sample

In particular, we labelled 62 samples as untreated monocytes, and 46 and 20 samples as M1 and M2 activated monocytes/macrophages, respectively.

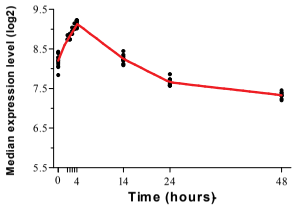
3.2 Analysis and validation of the inflammation model

3.2.1 Distinct gene signatures are identified during the inflammatory response

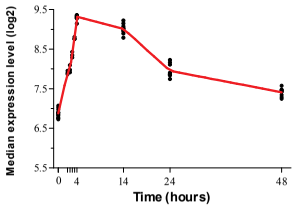
Transcriptomic analysis was performed on the data set of physiological inflammation. In particular, 60 gene expression profiles of the model were analyzed to identify specific genes ofeach stages of immune activation: fresh monocytes (time 0): early inflammation (2-4 h), late inflammation (14 h) (both corresponding to M1 polarization); early and late resolution (24 and 48 h) (different stages of M2c polarization). Genes showing statistically significant expression changes over time were identified using the microarray Significant Profiles method coded in the maSigPro R package. Since the time course was composed of nine points, we computed a regression fit for each gene using a polynomial with a degree of 3 (cubic regression model) and selected those regression models with an associated corrected p-value ≤0.05. P-values have been corrected for multiple comparisons using the false discovery rate procedure (FDR), i.e. setting the parameter Q=0.05 in the p.vector function. Once determined the statistically significant gene models, the regression coefficients have been used to identify genes showing statistically significant expression changes over time. To do this, a second model has been constructed using only significant genes and applying a variable selection strategy based on stepwise regression. Specifically, we selected the backward stepwise

regression and, at each iteration, retained those variables with a p-value ≤0.01 (i.e., set the T.fit parameters at step. method=backward and alfa=0.01). Finally, we generated the list of significant genes setting an additional selection criterion based on the R-squared value of the second regression model (i.e., set the get.siggenes parameters rsq=0.6 and vars=all). Results have been visualized clustering genes into nine groups (*Figure 3.1*) with similar expression profiles (k-means clustering method).

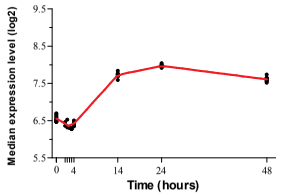
Cluster 1: inflammation (218 genes)



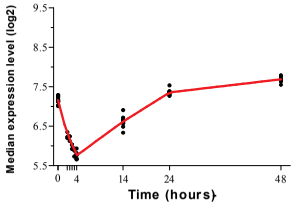
Cluster 2: inflammation (174 genes)



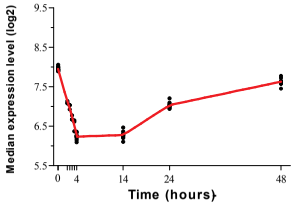
Cluster 3: early anti-inflammation (850 genes)



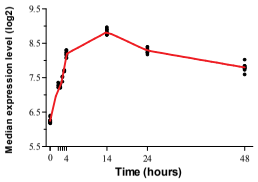
Cluster 4: anti-inflammation (445 genes)



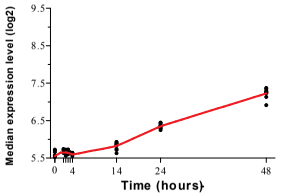
Cluster 5: anti-inflammation (576 genes)



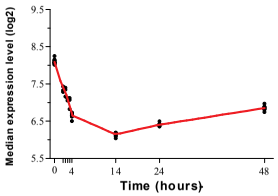
Cluster 6: inflammation-driven differentiation (457 genes)



Cluster 7: positive differentiation (214 genes)



Cluster 8: negative differentiation (680 genes)



Cluster 9: negative differentiation (381 genes)

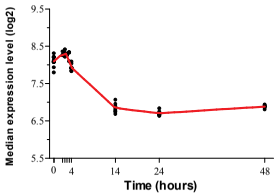


Figure 3.1: Median gene expression profiles of the 9 clusters identified by MaSigPro. Each cluster groups genes with a similar trend over time. Solid red lines have been drawn joining the median value of gene expression at each time point for each donor (dots).

In particular, we assembled i) genes of clusters 1 and 2 into inflammation functional group ii) genes of clusters 3, 4 and 5 into anti-inflammation group iii) genes of cluster 6 into inflammation driven differentiation group iv) genes of cluster 7 into positive differentiation and v) genes of clusters 8 and 9 in negative differentiation group. The inflammation phase corresponds to monocyte-to-M1 macrophages differentiation, and is associated with the modulation of 392 transcripts: 218 are transiently up-regulated during the first four hours of the inflammatory process (*Figure 3.1*, cluster 1), while 174 remain highly expressed during the late phases, i.e., until 14 h (*Figure 3.1*, cluster 2). In both clusters, transcriptional levels decrease during the resolution phase (24-48h). Genes included in these two groups are the typical effectors of classical activation, such as inflammatory cytokines (e.g., PPARG, IL6, TNFA IL1B, IL12B), chemokines (e.g., CCL4, CCL5, CCL20), extracellular mediators (e.g., PTX3, EDN1, APOL2), and enzymes (e.g., PTGS2, PLA1A). The early anti-inflammatory (*Figure 3.1*, cluster 3) and anti-inflammatory clusters (*Figure 3.1,* clusters 4 and 5) contain 850 and 1021 genes, respectively. The anti-inflammatory phase includes genes involved in metal homeostasis, detoxification, modulation of inflammation and control of the oxidative stress. The inflammation driven differentiation phase consists in 457 genes and which are

rapidly up-regulated upon the inflammatory reaction and then remain at elevated levels throughout all phases of the reaction (*Figure 3.1,* cluster 6). This behavior arises from the fact that this group comprises genes needed for the inflammatory response and also critical for the process of monocyte differentiation into deactivating and tissue-repairing macrophages. The Positive Differentiation group (*Figure 3.1*, cluster 7) includes 214 genes down-regulated in fresh monocytes and during the early inflammation phases, but progressively up-regulated during time with a transcriptional peak during late resolution. Conversely, the Negative Differentiation group (*Figure 3.1*, clusters 8 and 9) comprises a total of 1061 genes highly expressed in fresh monocytes and in early inflammation, and reduced during the subsequent phases.

Overall, maSigPro results revealed that a total of 3995 genes were differentially expressed during the different phases of the inflammatory reaction, and during the concomitant monocyte-to-macrophage differentiation. The nine clusters generated by maSigPro have been merged into five major functional groups of genes to reflect the different phases of inflammation (*Figure 3.2*). In particular, genes were organised into five major functional groups characterizing the different phases of inflammation in this experimental setting: inflammation, early anti-inflammation and anti-inflammation, inflammation driven differentiation, positive and negative differentiation.

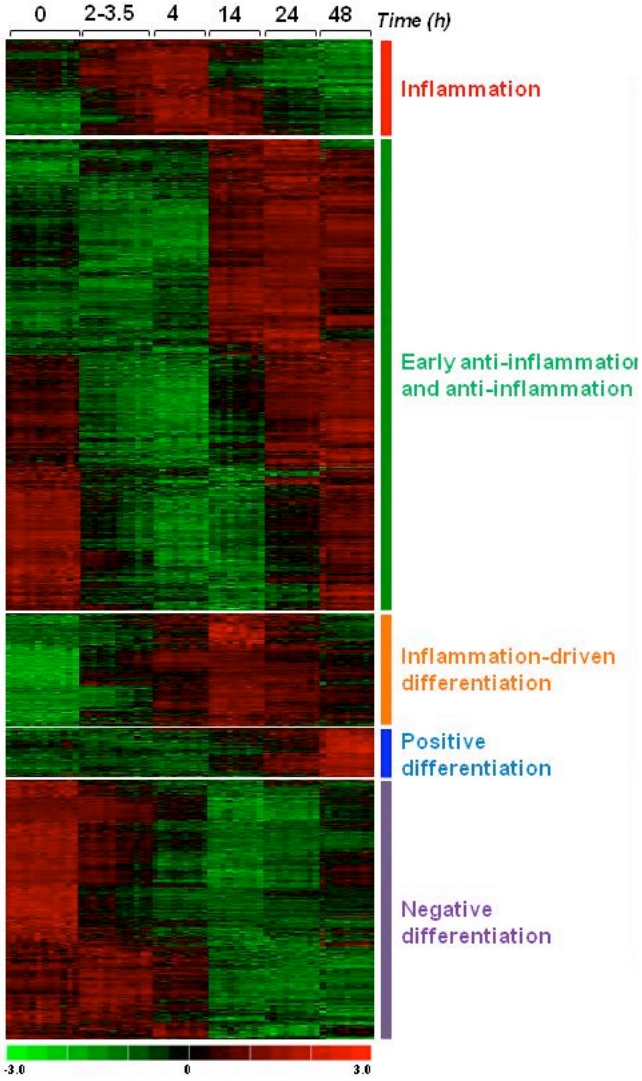


Figure 3.2: Differential gene expression during the inflammation. Heat-map showing the fold-expression levels of the genes that were identified by maSigPro as coherently down-regulated (green) or up-regulated (red) within the experimental set of 60 samples.

3.2.2 Pathway analysis reveals relationship between activation and differentiation

To investigate the biologic role of the genes differentially expressed during the development of the inflammatory response, each cluster was subjected to an over-representation analysis for statistical associations between expression profiles of distinct groups and other known gene signatures characteristic of various pathways or cellular processes described in KEGG, BIOCARTA and REACTOME databases. Over-representation analysis was performed using Gene Set Enrichment Analysis software (GSEA) and the gene sets of the Broad Institute Molecular Signatures Database (www.broadinstitute.org/gsea/msigdb/index.jsp). GSEA was applied on log2 expression data of the entire time course. The median expression profile of the 9 groups of genes identified by maSigPro was used as continuous phenotype labels and Pearson's correlation as the metric to select gene sets with expression patterns resembling those encoded in the phenotype labels. As gene sets we used KEGG, BIOCARTA, and REACTOME lists of the C2 curated gene sets collection. We identified a total of 155, 358, 55, 149, and 66 pathways most strongly associated with the median expression profile of the inflammation, early anti- and anti-inflammation, inflammation driven differentiation, positive differentiation and negative differentiation clusters, respectively. We found that most gene sets associated to the inflammation clusters are

classical inflammatory pathways involved in innate immune activation (*Table 3.1*) and, specifically, involved in type I inflammatory response carried out by M1 macrophages (such as NF-kB, MAPK and JAK-STAT signaling, NOD-like receptor and Toll-like receptor signaling, cytokine/chemokine receptor interaction, and the IL-1 receptor pathway). The Early Anti-Inflammation and Anti-Inflammation clusters are enriched in pathways associated to metabolism and regulation of gene expression (*Table 3.1*). The modulation of genes involved in cellular metabolic activities is characteristic of M2 macrophage polarization/differentiation (Martinez et al., 2011), because it occurs during the phases of resolution and repair, when major rearrangements of cellular functions are required. The inflammation driven differentiation group is associated to signaling cascades that are in common with both inflammatory and anti-inflammatory phases, while pathways enriched in the positive differentiation and negative differentiation clusters are similar to those found during the anti-inflammatory phase (data not shown). Globally, the functional enrichment analysis indicates that genes involved in inflammatory response and monocytes activation present transcriptional profiles that are statistically similar to those of genes involved in the control of the different cellular processes (as cell growth/proliferation and metabolism) during the monocyte-to-macrophage differentiation in vitro. These results establish a transcriptional link between monocyte activation and differentiation,

inflammation and metabolism on one side and inflammation, resolution and cell differentiation on the other.

Functional groups	FDR q-val
Inflammation	
BIOCARTA_NFKB_PATHWAY	0.003
BIOCARTA_IL-1R	0.010
BIOCARTA_IL-10_PATHWAY	0.013
BIOCARTA_INFLAM_PATHWAY	0.021
BIOCARTA_CD40_PATHWAY	0.033
BIOCARTA_CYTOKINE_PATHWAY	0.044
KEGG_MAPK_SIGNLING	0.000
KEGG_CYTOKINE_CYTOKINE_RECEPTOR_SIGNALING	0.000
KEGG_NOD_LIKE_RECEPTOR_SIGNALING	0.000
KEGG_ECM_RECEPTOR_INTERACTION	0.001
KEGG_CELL_ADHESION_MOLECULES_CAMS	0.005
KEGG_PATHWAY_IN_CANCER	0.008
KEGG_JAK_STAT_SIGNALING	0.011
KEGG_NOTCH_SIGNALING	0.015
KEGG_TOLL_LIKE_RECEPTOR_SIGNALING	0.037
REACTOME_CHEMOKINE_RECEPTORS_BIND_CHEMOKINES	0.000
REACTOME_GPCR_LIGAND_BINDING	0.000
Early Anti- and Anti-Inflammation	
KEGG_OXIDATIVE_PHOSPHORYLATION	0.000
KEGG_RNA_DEGRADATION	0.001
KEGG_FATTY_ACID_METABOLISM	0.014
REACTOME_BRANCHED_CHAIN_AMINO_ACID_CATABOLISM	0.000
REACTOME_ELECTRON_TRANSPORT_CHAIN	0.000
REACTOME_INTEGRATION OF ENERGY METABOLISM	0.000
REACTOME_METABOLISM_OF_CARBOHYDRATES	0.000
REACTOME_PYRUVATE_METABOLISM_AND_TCA_CYCLE	0.000
REACTOME_METABOLISM_OF_PROTEIN	0.000
REACTOME_DIABETES_PATHWAYS	0.000
REACTOME_METABOLISM_OF_RNA	0.003
REACTOME_FORMATION_AND_MATURATION_OF_MRNA_TRANSCRIPTS	0.004
REACTOME_MRNA_SPLICING	0.009
REACTOME_METABOLISM_OF_MRNA	0.011
REACTOME_GENE_EXPRESSION	0.016
REACTOME_MICRORNA_BIOGENESIS	0.040
REACTOME_CELL_CYCLE_MITOTIC	0.000
REACTOME_G1_S_TRANSITION	0.000
REACTOME_G2_M_CHECKPOINTS	0.031

Table 3.1: Most representative gene sets associated with the inflammation, early anti-Inflammation and anti-inflammation functional groups.

3.2.3 The M1 inflammatory signature develops into M2 during resolution

To verify if the in vitro model of inflammation can also recapitulate the transition from M1 to M2 phenotype polarization, we made a comparison between the polarized macrophages of the inflammation model and the polarized macrophages of the meta-database previously described. To this end, samples labeled as unstimulated monocytes, M1 or M2 polarized macrophages (*Table 3.2*) have been extracts from meta-data sets of monocytes, macrophages, and dendritic cells. In particular we selected and extracted gene expression data for 62 fresh monocyte, 46 M1 (treated with LPS/TNF-α or IFN-γ) and 20 M2 samples (M2c; treated with glucocorticoids, IL-10 or TGF-β).

GEO series	Platform	GEO samples
		Untreated monocytes
GSE5099	HG-U133A	GSM115051; GSM115046; GSM115047; GSM115048; GSM115049; GSM115050
GSE7807	HG-U133 Plus2.0	GSM189447; GSM189448; GSM189449; GSM189450
GSE8286	HG-U133A	GSM205587; GSM205588; GSM205590; GSM205591; GSM205592; GSM205594
GSE8658	HG-U133 Plus2.0	GSM214749; GSM214734; GSM214737; GSM214738; GSM214739; GSM214740; GSM214741; GSM214742; GSM214743; GSM214744; GSM214745; GSM214746
GSE9080	HG-U133Av2	GSM230145; GSM230149; GSM230147
GSE9988	HG-U133 Plus2.0	GSM252476; GSM252478; GSM252479; GSM252480; GSM252481; GSM252484; GSM252485
GSE11393	HG-U133Av2	GSM287664; GSM287665; GSM287666
GSE11430	HG-U133 Plus2.0	GSM257664; GSM257666; GSM257668; GSM257670; GSM257672
GSE11864	HG-U133 Plus2.0	GSM299556; GSM299557; GSM299561; GSM299562
GSE12108	HG-U133 Plus2.0	GSM305434; GSM305436; GSM305438; GSM305440; GSM305430; GSM305432
GSE12837	HG-U133A	GSM15431; GSM321582; GSM15430
GSE13762	HG-U133 Plus2.0	GSM346564; GSM346577; GSM346553
		M1 activation
GSE5099	HG-U133A	GSM115055; GSM115057
GSE9988	HG-U133 Plus2.0	GSM252423; GSM252424; GSM252425; GSM252427; GSM252428; GSM252429; GSM252431; GSM252432; GSM252433; GSM252434; GSM252435; GSM252436; GSM252437; GSM252438; GSM252439; GSM252440; GSM252441; GSM252442; GSM252443; GSM252444; GSM252445; GSM252447; GSM252448; GSM252449; GSM252450; GSM252451; GSM252453; GSM252454; GSM252455; GSM252456; GSM252457; GSM252458; GSM252459; GSM252460; GSM252461; GSM252462; GSM252463; GSM252464; GSM252430; GSM252426
GSE14419	HG-U133Av2	GSM360141; GSM360145; GSM360184; GSM360188
		M2 activation
GSE7568	HG-U133 Plus2.0	GSM183464; GSM183465; GSM183466; GSM183467; GSM183482; GSM183483; GSM183484; GSM183485; GSM183486; GSM183487; GSM183217; GSM183305; GSM183306; GSM183315; GSM183316; GSM183392; GSM183393; GSM183394; GSM183462; GSM183463

Table 3.2: Complete list of 128 samples labeled as untreated monocytes and as M1 and M2 activated monocytes and their sources.

Gene expression signals of selected samples of the meta-database were generated using the *Virtual-Chip* approach, which allows integrating raw expression data (i.e., CEL files) obtained from different Affymetrix arrays. Specifically, expression values were generated from intensity signals using

the combined HG-U133A/HG-U133Av2/HG-U133 Plus2.0 *virtual CDF* file, the custom definition files for human GeneChips based on GeneAnnot, and the transformed virtual-CEL files. Intensity values for a total of 12167 meta-probesets were background-adjusted, normalized using quantile normalization, and gene expression levels calculated using median polish summarization (RMA algorithm). Such matrix of gene expression profiles was analyzed with statistical method Significance Analysis of Microarray method (SAM), coded in the samrR package (http://cran.r-project.org/web/packages/samr/index.html), to identify differentially expressed genes in the comparisons between subsets of monocytes tagged as untreated, M1, and M2 (128 samples). Specifically, in the comparison between untreated monocytes and samples labelled as M1 (or as M2), we used the two-class procedure, estimated the percentage of false positive predictions with 1000 permutations, and selected those transcripts whose q-value (i.e., False Discovery Rate, FDR) was equal to 0. This selection was further refined setting the lower limit for fold change induction (or reduction) to 5 and 8, when considering the comparison between untreated monocytes and samples M1 or untreated monocytes and samples M2, respectively. The statistical comparison returned that monocyte-to-M1 differentiation is associated with modulation of 98 genes, of which 85% are highly expressed in M1 and 15% in monocytes (*Figure 3.3*), while monocyte-to-M2 differentiation resulted in the modulation of 107 genes, 62%

highly expressed in M2 and 38% in monocytes (*Figure 3.4*). Transcripts that are up-regulated in M1 cells vs. monocytes included cytokines and chemokines, while those up-regulated in M2 cells included enzymes and extracellular mediators.

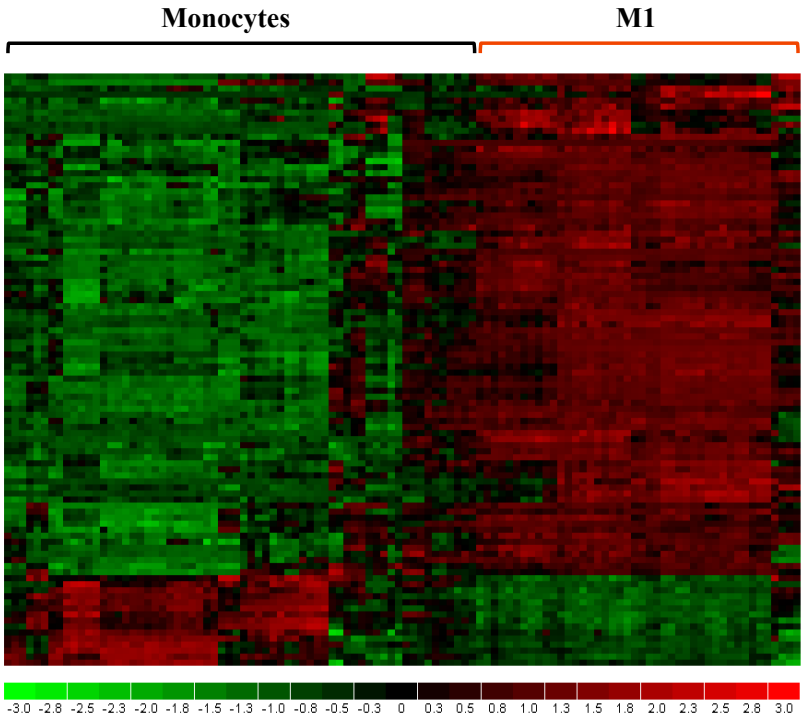


Figure 3.3: Differentially expressed genes in M1 macrophages vs. monocytes. Heat-maps representing the fold-expression levels of gene lists identified by SAM as statistically down-regulated (green) or up-regulated (red) in M1 samples compared to fresh unstimulated monocytes.

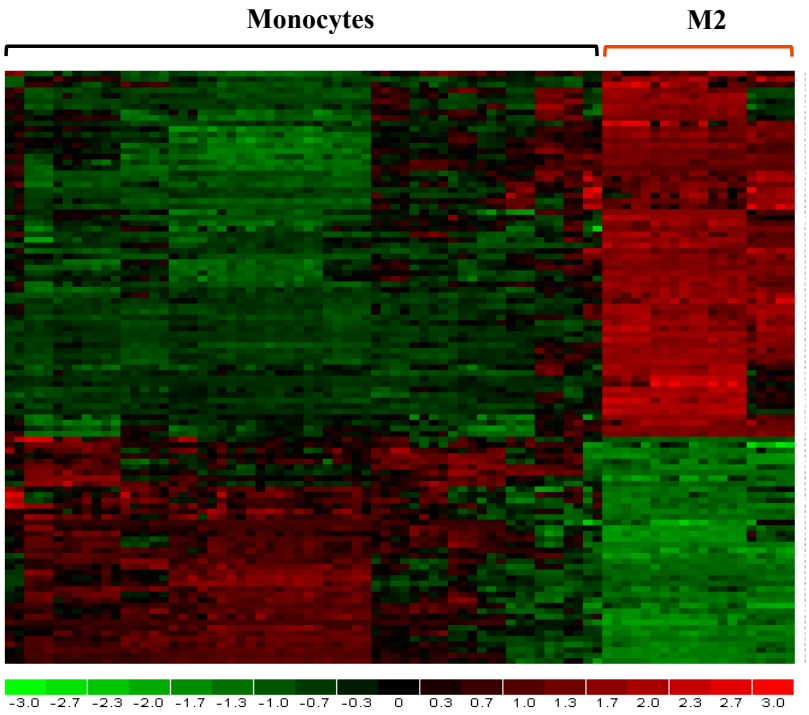


Figure 3.4: Differentially expressed genes in M2 macrophages vs. monocytes. Heat-maps representing the fold-expression levels of gene lists identified by SAM as statistically down-regulated (green) or up-regulated (red) in M2 samples compared to fresh unstimulated monocytes.

The two signatures of M1 and M2 polarization were used to cluster samples of our in vitro model of inflammation. As shown in the *Figure 3.5*, fresh monocytes (recruitment phase) showed a gene expression profile fully overlapping with that of unstimulated monocytes in the meta-database, then they presented a M1-like expression profile during the inflammatory phases, to return to a monocyte-like profile in the resolution

phase. When considering the gene set that discriminates monocytes from M2 cells, fresh monocytes (recruitment phase) showed the same profile as the untreated monocytes of the meta-database, and this profile gradually changed during the progression of inflammation, to become similar to that of M2 macrophages at the end of resolution phase (*Figure 3.6*). When comparing the list of genes differentially expressed during the inflammation process (*Figure 3.1*, Clusters 1 and 2) with the list of genes differentially expressed in monocytes vs. M1, a large number of genes expressed in M1 cells (34%) belongs to the Inflammation group. Conversely, 21% of genes expressed in M2 cells belong to the Positive Differentiation group (*Figure 3.1*, cluster 7) and are expressed only during the resolution phase. In the monocytes vs. M1 comparison, a large part of genes expressed in fresh monocytes belongs to the Anti-Inflammation group (26%, Figure x, clusters 4 and 5), while in the monocytes vs. M2 comparison 51% of genes expressed in monocytes are in the Negative Differentiation group (*Figure 3.1*, clusters 8 and 9). Among genes common to both M1 and M2 polarization, several belong to the Inflammation Driven Differentiation group (14% and 20%, respectively, *Figure 3.1*, Cluster 6). *Table 3.3* shows some representative genes identified in these comparisons.

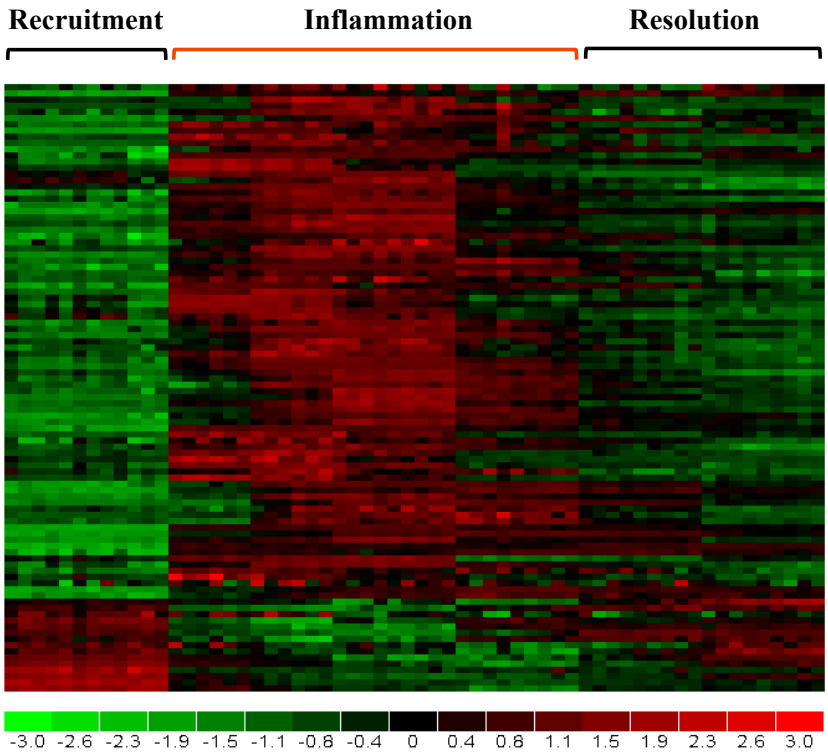


Figure 3.5: Fold-expression levels of the 98 monocyte-to-M1 genes assessed in the 60 samples of our in vitro model of inflammation.

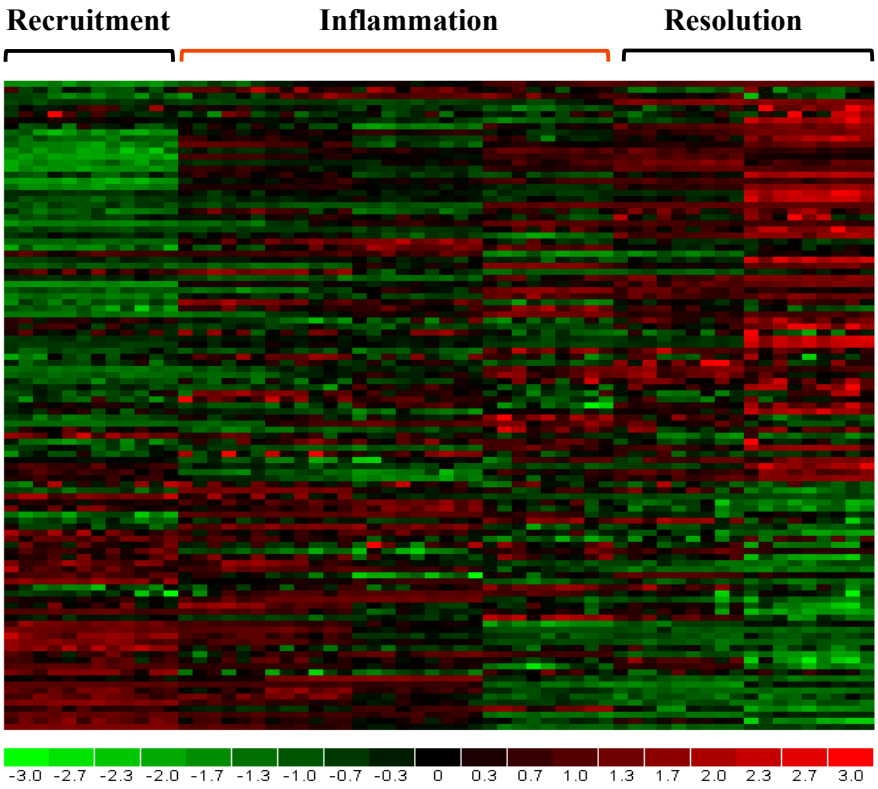


Figure 3.6: Fold-expression levels of the 107 monocyte-to-M2 genes assessed in the 60 samples of our in vitro model of inflammation.

Gene Symbol	Functional Groups
Genes uprgulated in M1 polarization	
IL12B, PTX3, CCL4, IL1RN, TNF, IL6, CCL20, IL1A, ICAM1, NFKB1, TRAF1, SERPINB9, IL1F9, MAFF	*Inflammation*
CXCL1, DRAM, TNIP3, CCL2, SLAMF7, CCR7, TNFAIP6	*Inflammation Driven Differentiation*
Genes downregulated in M1 polarization	
P2RY5, FGL2, CD1D	*Anti-Inflammation*
Genes upregulated in M2 polarization	
TREM2, A2M, NUPR1, C1QA, MS4A4A, APOE, APOC1, ADORA3	*Positive Differentiation*
ADAMDEC1, CD59, TFPI, CCL3	*Inflammation Driven Differentiation*
Genes downregulated in M2 polarization	
FCER1A, LGALS2, PF4, CD69, CD93, NR4A2, VCAN, CD62L, ICAM3, NLRP3, ERG1	*Negative Differentiation*

Table 3.3: Correlation between M1/M2 polarization and functional groups.

3.2.4 Validation of gene expression by real-time PCR

In order to quantitatively validate the microarray results, a total of ten genes were examined by real-time PCR, employing the same RNA used to hybridize the Affymetrix arrays. A subset of ten genes was assessed by qRT-PCR, five transcription factors chosen as markers of monocyte differentiation, and five inflammation-related factors as markers of monocyte activation, selected within each functional group of *Figure 3.2*. The qPCR results confirmed the expression patterns observed by microarray analysis (*Figure 3.7*). Genes belonging to the Inflammation group (PPARG, IL6, TNFA) were up-regulated

during the early phase, while IL7R was over-expressed during the late phase of inflammation. CD163 (Early Anti-Inflammation) was highly up-regulated at the beginning of resolution, possibly induced by IL-10, while the transcription factor CEBPA (Anti-Inflammation) was overexpressed during late resolution, possibly induced by TGF-β. Expression of PPARD (Inflammation Driven Differentiation) increased during late inflammation and remained elevated, while MAFB and MMP9 genes (Positive Differentiation) were up-regulated during resolution. Finally, expression of KLF4 (Negative Differentiation) was high in fresh monocytes and decreased thereafter.

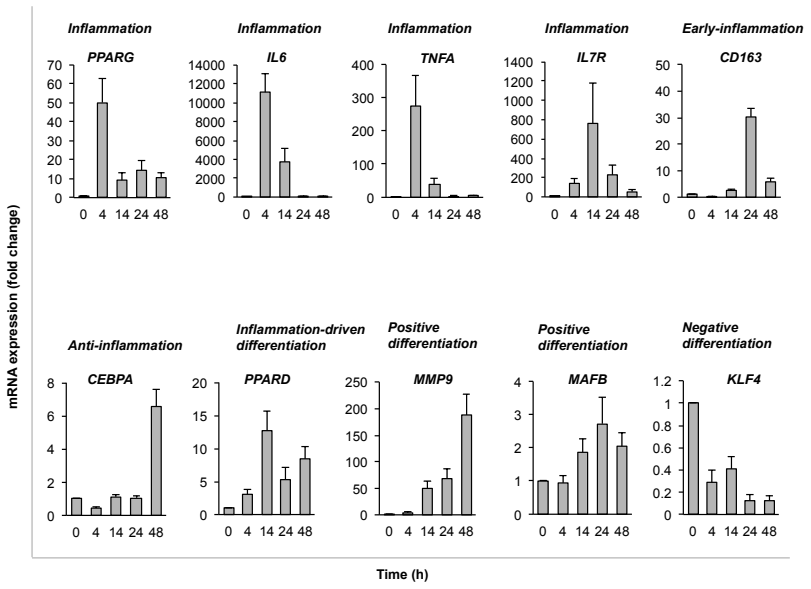


*Figure 3.7:Fold-expression levels determined by qPCR for the 10 genes selected. The mean expression values ± SEM from six different donors are reported. Statistical significance was calculated with ANOVA followed by Fisher's test for significant differences between two consecutive experimental time points. * P < .05 ; ** P < .001 ; *** P < .0001*

3.3 Reconstruction of module networks during physiologic inflammation

Once established that the in vitro model of inflammation accurately recapitulates the development of the human inflammatory reaction, we decided to use it to reconstruct gene regulatory network. However, extracting new biological insight from high-throughput genomic studies of human diseases is a challenge, limited by difficulties in recognizing and evaluating relevant biological processes among huge quantities of expression data. To overcome this problem, it could be very helpful to analyze the genes differentially expressed at a higher level, the gene sets. We decided to analyse changes in monocytes/macrophages gene expression patterns during the inflammatory phase in order to better elucidate the mechanisms underlying systemic inflammatory responses. To this end, we extracted from the original matrix of gene expression profile of the in vitro model, a submatrix containing only the genes that showed an "inflammation profile" identified by maSigPro, as previously described (*Figure 3.1*, cluster 1 and 2). GSEA was applied on log2 expression data of the

entire time course. The median expression profile of the submatrix of the inflammation related genes identified by maSigPro was used as continuous phenotype labels, and Pearson's correlation as the metric to select gene sets with expression patterns resembling those encoded in the phenotype labels. The algorithm rank genes based on the correlation between their expression and the continuous labels using Pearson correlation, then estimates the false discovery rate (FDR) to control the probability that each reported result is a false positive; this permit to determine the specificity and the significance of the analysis. The output of the analysis is a ranked list of enriched pathways compared to the genes contained in the matrix. A low FDR correspond to a high significance of the analysis; in case of time course experiments, FDR must be lower than 0.05. Using these parameters we obtained 6 gene sets strongly associated with the given profile and strongly associated with inflammatory reaction. In particular, the analysis of the inflammation cluster of maSigPro resulted in the enrichment of the following pathways: interleukin-1 receptor (IL-1R), NF-kB activation, Cytokines and Inflammatory Response, the nuclear factor-kB (NF-kB) pathway (*Table 3.4*), Cytokine Network, and TNFR2 Signaling Pathway.

Gene Set Name	FDR q-val
BIOCARTA IL-1R PATHWAY	0.046
BIOCARTA NTHI PATHWAY	0.029
BIOCARTA INFLAMMATION PATHWAY	0.034
BIOCARTA NF-Kb PATHWAY	0.044
BIOCARTA CYTOKYNE PATHWAY	0.049
BIOCARTA TNFR PATHWAY	0.050

Table 3.4: GSEA analysis results obtained using the median expression profile from maSigPro inflammation clusters.

Since IL-1R pathway is one of the major pathways that lead the inflammation reaction and also one of the most modulated pathway in our model (based on previous analysis) we decided to focus our attention on the IL-1R pathway. IL-1R family detects microbial components and triggers complex signaling pathways that result in increased expression of multiple inflammatory genes; on the other hand, an aberrant activation of IL-1R signaling can promote the onset of inflammatory and autoimmune diseases. The signal transduction cascade utilized by IL-1 receptor results in NF-kB activation. Understanding which are the main protagonists of these mechanisms could identify attractive targets for anti-inflammatory drug discovery, because their inhibition may impair a subset of noxious inflammatory signals impinging on NF-kB, while sparing its normal physiological activation. In particular, we focused on those genes constituting the core enrichment of IL-1 receptor pathway in GSEA analysis. The

core enrichment is the subset of genes of the pathway that contributes the most to the enrichment result and contains the genes in the experiment data that show the highest correlation with the given profile. Through the clustering visualization of genes that form the core enrichment pathway of IL1-R, it can be seen how the inflammation-related events impact the expression changes of the IL1-R pathway (*Figure 3.7*).

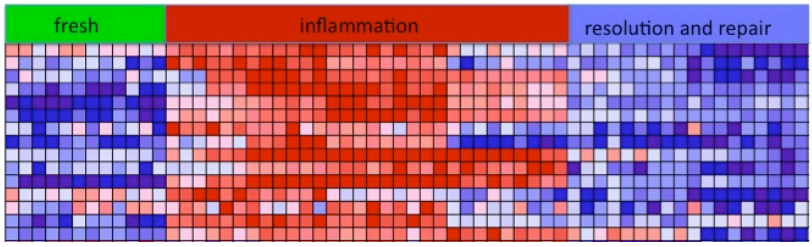


Figure 3.8: Cluster of genes that form the core enrichment pathway of IL-1R. This analysis shows that all the members of the IL-1R pathway are highly upregulated during inflammation phase and downregulated during resolution and repair phases.

Since our bioinformatics analysis allowed the characterization of several members of interleukin-1 family as master regulators of the inflammation process, we can consider those genes constituting the core enrichment of IL-1R pathway from GSEA as regulators of the IL-1R pathway in our model. Once we selected potential regulators that mediate the observed transcriptional response, the major goal of our work was to understand how they interact together and which are the transcriptional modules regulated by the gene network of

selected regulators. The last step of the pipeline includes the use of software GENOMICA. GENOMICA has been applied to human in vitro model of physiological inflammation after GSEA analysis to look for the transcriptional regulation of genes involved in the immune response. Genomica is an analysis and visualization tool for genomic data, using this software it is possible to reconstruct module networks of sets of genes that are coregulated. As input Genomica needs a matrix containing gene expression values and a list of putative regulators formed, in this case, by the core enrichment genes of GSEA analysis. As parameters we specified 0 as lookahead depth and 5 as maximum tree depth. The cluster below the regulation tree contains regulated genes, i.e, those genes whose expression profiles are correlated to the regulators.

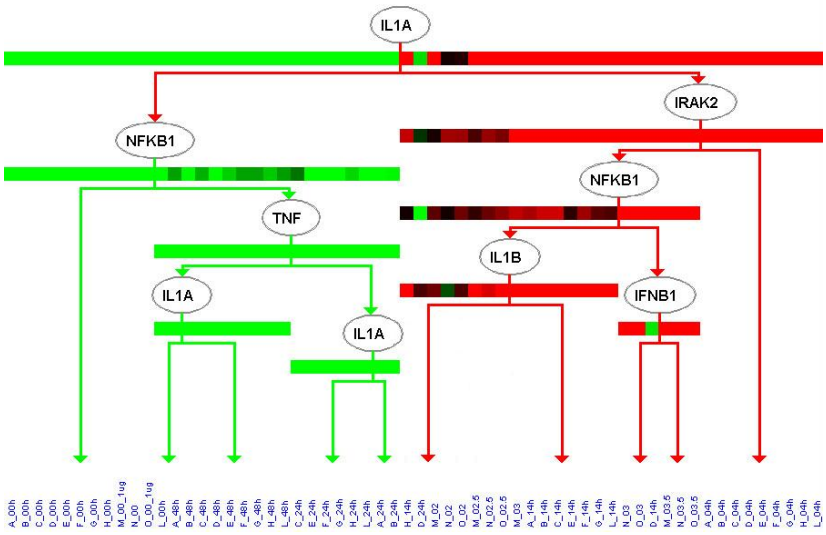


Figure 3.9: module 1 of IL-1R pathway. Regulation tree of IL-1R pathway obtained with Genomica. Red branches show what happens when genes are up-regulated instead green branches indicate genes down-regulation. A regulation tree or program can represent the different modes of regulation described above. Each node in the tree consists of a regulatory gene (for example, 'Activator') and a query on its qualitative value, in which an red arrow denotes the query "is gene up-regulated?" and a green arrow denotes the query "is gene down-regulated?". Right branches represent instances for which the answer to the query in the node is 'true'; left branches represent instances for which the answer is 'false'. The expression of the regulatory genes themselves is shown below their respective node.

One of the genes recognized as main regulator by the algorithm is interleukin-1A (IL-1A) that is one of the IL-1R ligands; the other genes in the tree trace the pathway (Figure 3.8). The left arm of the tree describes what happens when the first regulator is down-regulated and this occurs at time 0 and at end of the experiment (24 and 48 hours time points). This is confirmed by the fact that a pathway involved in the activation of inflammatory response must be turned off at the beginning of the experiment and in the final stages, which correspond to the resolution of inflammation. The right arm of the tree instead illustrates the effects derived from first regulator up-regulation that occur in the central phases of the experiment (from 2 to 14 hours time points) where the inflammatory process takes place. The module regulated by this gene network consists of 183 genes up-regulated during inflammation phase (listed in *table 6.1*).The comparison between Genomica result and the validated pathway downloaded from BIOCARTA website (www.BIOCARTA.com) denotes the efficacy of the method,

since our gene network has almost the same structure of the BIOCARTA pathway illustrated in *figure 3.9*.

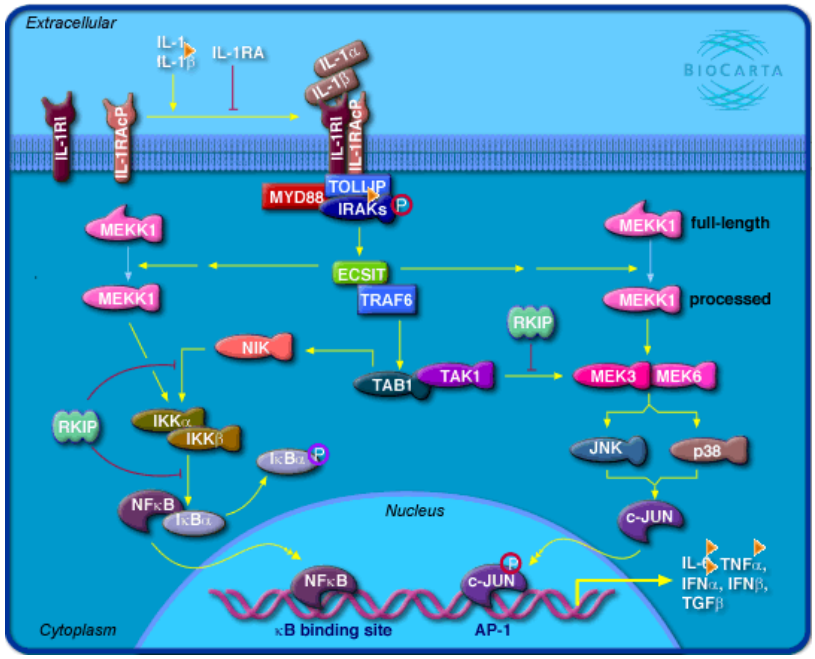


Figure 3.10: validated IL-1R pathway from BIOCARTA website.

Moreover, Genomica tool identified several gene regulatory networks showing IL-1R pathway activation in which the signal is mediated via classical NF-kB signal members (*Figure 3.10* and *Figure 3.11*), and in which both genes forming the regulation tree and those belonging to the underlying module (modulated by regulation tree) are transcriptionally activated during the inflammation phase. In particular in *figure 3.10* is shown a gene regulatory network in which IL-1R, Irak2, IFNB1

and MAPK8 are master genes regulating the up-regulation of 187 target genes contained in that specific module.

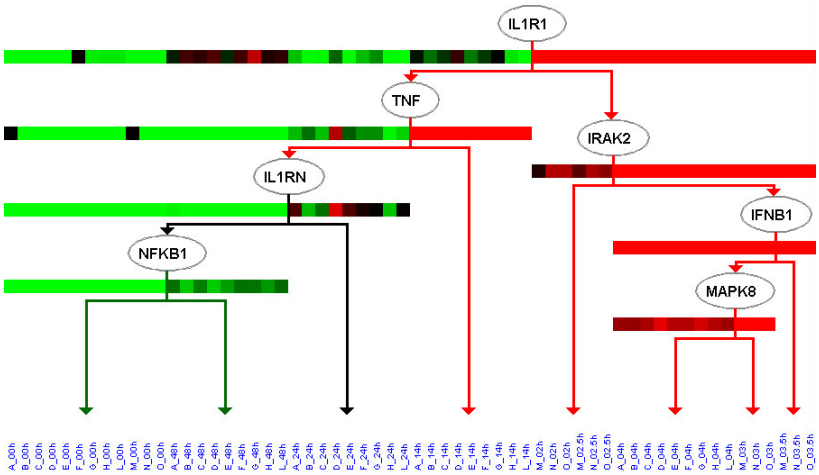


Figure 3.11: module 2 of IL-1R pathway. The second regulation tree of IL-1R pathway obtained with Genomica.

In *figure 3.12* there is another IL-1R network identified by GENOMICA, that regulates a module containing 214. This module summarizes the IL-1R pathway activation, in which IL-1R stimulation leads the expression of TNF via NF-kB signaling cascade.

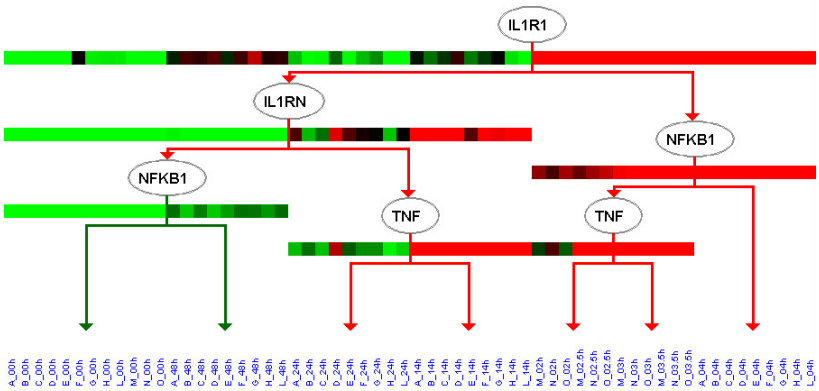


Figure 3.12: module 3 of IL-1R pathway. The third regulation tree of IL-1R pathway obtained with Genomica.

3.4 GSEA enrichment

GSEA has been used to functionally characterize Genomica modules. Genes belonging to the modules previously described has been analyzed with the “analyze gene sets” tool of the MSigDB web pages. Those genes resulted in an enrichment of cascades or cellular functions referring to the immune system and metabolic process. *Tables 3.5*, and *3.6* report the enrichment result using genes from the first and from the second Genomica modules previously showed. In particular the first and the second modules are significantly enriched in pathways related to the inflammatory process while the third seems to be made up genes involved in metabolic process and regulation of transcription (*data not shown*).

GENE SET NAME	FDR q-value
REACTOME_IMMUNE_SYSTEM	4.65E-05
REACTOME_CYTOKINE_SIGNALING_IN_IMMUNE_SYSTEM	2.26E-04
REACTOME_PLATELET_ACTIVATION_SIGNALING_AND_AGGREGATION	2.26E-04
REACTOME_HEMOSTASIS	2.26E-04
REACTOME_INTEGRIN_CELL_SURFACE_INTERACTIONS	1.50E-03
REACTOME_ANTIVIRAL_MECHANISM_BY_IFN_STIMULATED_GENES	1.17E-02
REACTOME_INTERFERON_SIGNALING	2.99E-02
REACTOME_METABOLISM_OF_LIPIDS_AND_LIPOPROTEINS	3.40E-02
REACTOME_SIGNALING_BY_ILS	4.97E-02

Table 3.5: Enrichment results using genes from the module 1 of IL-1R reconstructed with Genomica.

GENE SET NAME	FDR q-value
REACTOME_CYTOKINE_SIGNALING_IN_IMMUNE_SYSTEM	5.80E-06
REACTOME_IMMUNE_SYSTEM	4.22E-05
REACTOME_SIGNALING_BY_ILS	7.49E-04
REACTOME_SIGNALING_BY_TGF_BETA_RECEPTOR_COMPLEX	7.49E-04
REACTOME_TRANSCRIPTIONAL_ACTIVITY_OF_SMAD2_SMAD3_SMAD4_HETEROTRIMER	1.69E-03
REACTOME_IL1_SIGNALING	1.69E-03
REACTOME_SIGNALING_BY_NOTCH	4.71E-03
REACTOME_INTERFERON_ALPHA_BETA_SIGNALING	9.10E-03
REACTOME_SMAD2_SMAD3_SMAD4_HETEROTRIMER_REGULATES_TRANSCRIPTION	1.03E-02
REACTOME_SIGNALING_BY_NOTCH1	1.03E-02
REACTOME_AMINO_ACID_TRANSPORT_ACROSS_THE_PLASMA_MEMBRANE	1.38E-02
REACTOME_INTERFERON_SIGNALING	2.10E-02
REACTOME_INNATE_IMMUNE_SYSTEM	3.82E-02
REACTOME_AMINO_ACID_AND_OLIGOPEPTIDE_SLC_TRANSPORTERS	4.23E-02

Table 3.6: Enrichment results using genes from the module 2 of IL-1R modules reconstructed with Genomica.

3.5 Pscan analysis

Transcription is modulated by the interaction of transcription factors with their corresponding binding sites on the DNA

sequence. To asses which motifs are significantly over-represented in the promoters of genes comprise in the modules, we performed an analysis using Pscan. Pscan is a software tool that scans promoter sequences from co-regulated or co-expressed genes, looking for the most represented motifs describing the binding specificity of known TFs. We performed the analysis defining the promoter region as 450 bp upstream and 50 bp downstream the transcription start sites (TSSs) of the genes and using transcription binding sites matrices in JASPAR database. As expected, the most over-represented TFs in modules obtained with Genomica analysis was NF-kB. Surprisingly we found that many modules are also enriched in genes that show a binding site motif for Sp1 despite the gene regulatory network that governs the modules has several genes that belong to NF-kB signal transduction pathway as mainly signal transductors. Moreover Sp1 doesn't belong to the core enrichment of the IL-1R pathway. The protein encoded by Sp1 gene is a zinc finger transcription factor that binds to GC-rich motifs of many promoters. The encoded protein is involved in many cellular processes, including cell differentiation, cell growth, apoptosis, immune responses, response to DNA damage, and chromatin remodeling. Our analysis suggests that there could be a cooperative action of the two TFs during activation of inflammation. Previous studies confirm our hypothesis, data from literature have shown that the activation of genes by NF-κB transcription factors may be modulated by synergistic or

antagonistic interactions with other promoter-bound transcription factors. For example, Sp1 sites are often found in NF-κB-regulated genes, and Sp1 can activate certain promoters in synergism with NF-κB through nonoverlapping binding sites (Hirano et al., 1998). In particular, we have found that in all three modules examined, Sp1 is significantly overrepresented, with a p-value equal to 0.00044, 0.00025, 0.00055 for module 1, 2 and 3 respectively.

3.6 ChIP-seq analysis

In order to show that the genes found in Genomica analysis are true targets of NF-kB, Sp1, or both, we analyzed some ChIP-seq data sets. ChIP-seq is a technique that allows to combine chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins. In particular, we used two ChIP-seq data set: one in which immunoprecipitation was performed with specific antibodies against Sp1 and one with specific antibodies against a subunit of NF-kB (p65), both after stimulation with LPS. We retrieved genomic regions with significant ChIP-seq peaks and, through an analysis with MEME-ChIP web service, we confirmed that the most over-represented motifs in the peaks, were those associated to the immunoprecipitated factors. Therefore, we associated peaks with genes using GREAT peaks annotation tool. We identified as target genes of Sp1 or NF-kB all these genes that are

comprised between -10000 and 2000 bp from peaks identified by Homer and annotated with GREAT. To asses if there is a significant overlap between genes belonging to genomica modules and TFs targets identified with ChIP-seq (Figure 3.12), we performed an enrichment test using the hypergeometric distribution and corrected p-values for FDR (BH correction). All three considered modules display a highly significant enrichment in both Sp1 and p65 targets (*Table 3.9*).

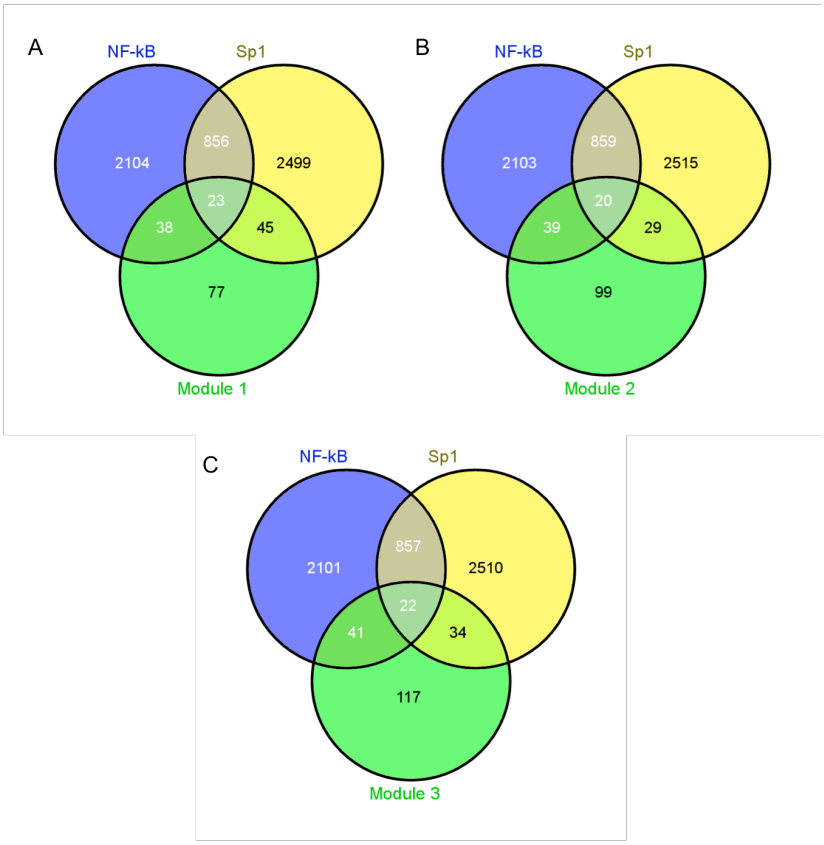


Figure 3.13: Venn diagrams show the overlaps between genes obtained from ChIP-seq analysis and genes belonging to (A) module 1, (B) module 2 and (C) module 3.

Genes in common between the results of ChIP-seq analysis of NF-kB, Sp1and genes in the module 1
DCP1A, BACH1, CRK, ITGB1, RNF19A, NINJ1, SQSTM1, SLC39A13, TXNL4B, G3BP1, PPAP2A, DEGS1, TRIO, KPNA4, IFNGR2, PANK3, RELB, TEX10, AZIN1, PIM3, CPD, TCP1, RAB8B
Genes in common between the results of ChIP-seq analysis of NF-kB, Sp1and genes in the module 2
DGAT2, NR4A3, MTHFD1L, TNFSF14, TNFAIP3, HIVEP1, NFAT5, PTGER2, REL, RAPH1, NFKBIZ, OSGIN1, BTG2, SLC7A1, DENND4A, ARRDC4, INSIG1, FAM126A, NFKBID, IL23A
Genes in common between the results of ChIP-seq analysis of NF-kB, Sp1and genes in the module 3
FEM1C, SERPINB8, CTNNB1, ITGB1, HMGA1, EDEM1, STK40, HOMER1, SPAG9, MAPK6, SLC35F5, CEBPG, NDRG1, TRIB3, SERTAD1, ZCCHC6, HNRNPL, PIM3, VRK2, CRY1, NUMB, CHIC2

Table 3.7: Genes in common between genes obtained from ChIP-seq analysis and genes belonging to module 1, module 2 and module 3 .

Module N°	Sp1 q-val	NF-kB q-val
1	5.03E-09	1.60E-08
2	3.83E-03	1.76E-07
3	2.60E-03	7.95E-07

Table 3.8 Benjamini–Hochberg corrected P-values representing enrichment of module genes

3.7 Validation of IL-1B up-regulation during inflammatory phase

In order to validate the enrichment results, one of the most important protagonist of IL-1R pathway, i.e. cytokine IL-1β,

was examined by real-time PCR, employing the same RNA used to hybridize the Affymetrix arrays, and its protein production was assessed by ELISA analysis. As show in *figure 3.13* we obtained in vitro the same results of in silico analysis i.e. the exposure to inflammatory stimuli (LPS and TNF-α) induced an early (4 h) increase in gene expression of IL-1β, while it was down-regulated during the late phase of inflammation (at 14 h, after the addiction of IFN-γ in culture). The protein production reported as the velocity of production (pg or ng/hr/million cells), confirmed the inflammatory role of IL-1β highlighting its abundant presence during the full (14 h) development of the inflammatory response, with a significant decrease in the later phases.

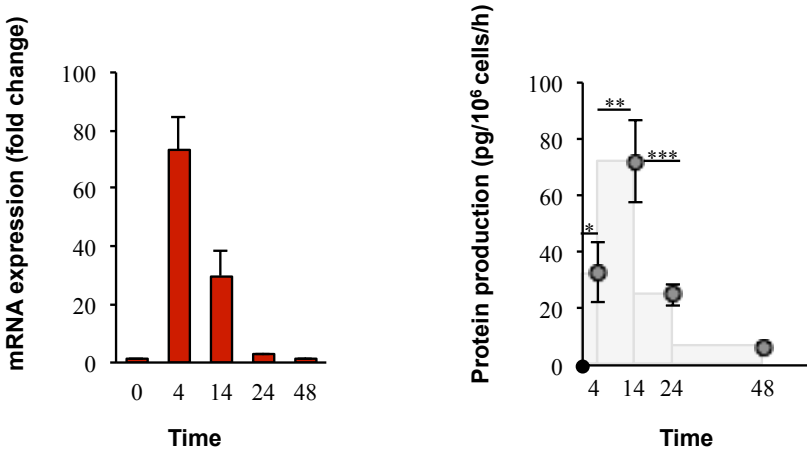


Figure 3.14: Gene expression and protein production of IL-1B during the different phases of inflammatory reaction. Soluble protein recovered in the supernatant is reported in terms of velocity of

production, (the amount of protein produced per one million cells per hour). Statistical significance was calculated with ANOVA followed by Fisher's test for significant differences between two consecutive experimental time points. P < 0.05 ; ** P < 0.001 ; ***P < 0.0001.*

DISCUSSION

The continuing use of high-throughput assays to investigate cellular responses to infection is providing a large repository of information. Most of these data is organized in public databases and their meta-analysis represents an enormous opportunity to study the mechanisms of immunological processes. However, their meta-analysis is a complex challenge for bioinformatics. Collection and analysis of public gene expression remains a process which is long and easily affected by errors, further complicated by the heterogeneity of microarray technologies and the characterization of samples. These latter aspects are the most critical problems affecting the meta-analysis approaches, in fact different design of microarray platforms complicate the integration of already incomplete, and often inadequately characterized samples, limiting the robustness of the statistical analysis. Many software and web applications have been proposed for the recovery, the organization and meta-analysis of microarray expression data obtained and deposited in public databases. However, none of them offers the possibility to download and organize locally the raw data and related meta-information, re-annotate samples using meta-data, group the samples according to the user's choices and integrate and standardize data from different platforms. To overcome the problem of incomplete and inadequate characterization of the samples, we constructed and annotated a meta-database of gene

expression profiles of monocytes, macrophages and dendritic cells, obtained from public databases in AMADMAN. Moreover, to overcome the problem of the integration of different platforms we proposed and implemented a process of integration and standardization of meta-analysis, called *Virtual-Chip*. The innovative aspect of this approach is that the integration of different platforms is performed before the generation of signal directly at the level of the probes on the microarray, in practice we normalize the fluorescence signals and quantify the level of expression as if all the different samples were obtained with the same type of platform. AMADMAN coupled with the *Virtual-Chip* procedure were applied to create, annotate and analyze a gene expression meta-database of monocytes, macrophages and dendritic cells, derived from 24 different experiments stored in GEO. The meta-database was used to validate an owner experiment (including 60 samples). Specifically, the experiment is an in vitro model of physiological inflammation that consists in a 48 hour culture of primary human monocytes isolated from 9 healthy donors. During the culture monocytes are exposed sequentially to a series of stimuli that mimic conditions in a simplified microenvironment that develops in inflamed tissue. The samples were taken at time 0, 4, 14, 24 and 48 hours and processed for analysis of transcriptional profiles with microarray. Using the expression matrix of the model and the tool maSigPro, we found distinct clusters of genes that are differentially regulated during the different phases of

inflammation. By comparing the genes differentially expressed between monocytes vs. M1 and vs. M2, it is evident that monocytes in our model show the M1 transcriptome in the inflammatory phase and an M2 profile during resolution. Thus, that the genes involved in inflammatory activation belong to the same biological pathways involved in cellular processes of monocyte-to-macrophage differentiation establishes a transcriptional connection between monocyte activation and differentiation, inflammation and metabolism. Therefore, resolution of inflammation is strictly connected to macrophage differentiation in the tissue. However, measurement of mRNA levels can provide only a partial view of the regulatory state of a cell during inflammation. At present, unfortunately, there remain major technical difficulties to elevate the analysis from gene lists to signaling pathways and transcriptional networks. In particular, many tools have been developed for the reconstruction of molecular mechanism in cancer (Agnelli et al., 2011) while there is a lack of methodologies for unveiling these mechanism in inflammatory diseases. With the aim of characterizing the functions of specific genes, the relationships among these genes, their regulation and coordination during inflammatory reaction we setup an *ad-hoc* pipeline. Specifically, in a first step we used several computational methods to identify those network components that have a critical role in the innate immune system and in a second step we adopted concepts from the Bayesian network theory to reconstruct gene regulatory network and their associated

modules. Using these methods in the appropriate order we have been able to identify “master regulator” genes, i.e., genes that, although not necessarily associated with any specific phenotype, nevertheless play a fundamental role in the web of transcriptional interactions. Then we used these genes to reconstruct regulatory networks that take places during physiological inflammation. This analysis permitted to reconstruct transcriptional modules regulating monocytes differentiation during physiological inflammation. The designed pipeline has allowed to reproduce the pathway of IL-1R and to decipher the cascade of genes that derived from this process. In particular, on these genes, we performed an enrichment analysis through which it was possible to assess that the genes of the modules are enriched in inflammation genes and genes related with metabolic process. We performed also an enrichment analysis to asses which motifs are significantly over-represented in the promoters of genes comprised in the modules. We found that most of genes show a binding site motif for NK-kB, as we expected since the stimulation of IL-1R leads to an activation of NF-kB signalling cascade. However, also putative Sp1 targets were over-represented in our modules. This may suggests that there is a cooperative action of the two TFs during activation of inflammation. Data from literature (Hirano et al., 1998) have shown that in most vertebrates NF-κB factors interact with Sp1. To verify if genes in the modules are true targets of NF-kB and Sp1, we performed a comparison between data from the physiological

inflammation model and data obtained from ChIP-seq analysis. This comparison showed that there is an actual contribution to the expression of genes in the three modules from the two TFs. So, through our pipeline, it is possible to identify genes that characterize the various steps of the inflammatory process, understand which are the pathway enriched in the different steps of inflammation, and, through the extraction of the core enrichment of a pathway of interest, understand which genes actually govern its activation. Finally it is possible to investigate the genes modulated by the gene network built with the core enrichment, to find new targets of a pathway of interest or to make new hypotheses about gene interactions.

CONCLUSIONS

Here, we report the setup of a physiological inflammation model, a bioinformatics approach to analyze it and an *ad-hoc* pipeline to reconstruct module networks from these data.

We demonstrated that i) the *in vitro* system based on primary human cells can allow us to describe the kinetic development of cell reactivity and its modulation during the entire course of the inflammatory response in a robust and reliable fashion, ii) there is a transcriptional link between monocyte activation and differentiation, inflammation and metabolism on one side and inflammation, resolution and cell differentiation on the other, iii) monocytes can shift from M1 to M2 polarization upon microenvironmental changes.

The fact that the same monocyte population goes through all the phases of the inflammatory process by adapting its phenotype and function to the evolution of microenvironmental conditions was already suggested by studies in mouse models, but never previously shown for human cells.

Finally, in order to reconstruct gene regulatory modules based on the transcriptional data of the inflammation model, I developed *ad-hoc* bioinformatics pipeline. This approach allowed the characterization of several member of the interleukin-1 receptor pathway as master regulators of specific regulatory modules linked to the inflammation process and, through scanning analysis of transcription factor (TF) binding site motifs, we identified genes putatively regulated by the cooperation of the transcription factors nuclear factor kappa-light-chain-enhancer of activated B cells (NF-kB) and specificity protein 1 (Sp1). Finally,I was able to validate the results using a ChIP-seq data sets.

Overall I demonstrate that it is possible to investigate the genes modulated by the gene network built with the core enrichment, to find new targets in a complex biological system such as the immune system and to make new hypotheses about gene interactions

REFERENCES

Agnelli, L., Forcato, M., Ferrari, F., Tuana, G., Todoerti, K., Walker, B.A., Morgan, G.J., Lombardi, L., Bicciato, S., Neri, A., 2011. The reconstruction of transcriptional networks reveals critical genes with implications for clinical outcome of multiple myeloma. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. 17, 7402–7412.

Bailey, T.L., 2011. DREME: motif discovery in transcription factor ChIP-seq data. Bioinforma. Oxf. Engl. 27, 1653–1659.

Bailey, T.L., Williams, N., Misleh, C., Li, W.W., 2006. MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res. 34, W369–373.

Barish, G.D., Yu, R.T., Karunasiri, M., Ocampo, C.B., Dixon, J., Benner, C., Dent, A.L., Tangirala, R.K., Evans, R.M., 2010. Bcl-6 and NF-kappaB cistromes mediate opposing regulation of the innate immune response. Genes Dev. 24, 2760–2765.

Bisognin, A., Coppe, A., Ferrari, F., Risso, D., Romualdi, C., Bicciato, S., Bortoluzzi, S., 2009. A-MADMAN: annotation-based microarray data meta-analysis tool. BMC Bioinformatics 10, 201.

Bisognin A, Mazza EMC, Ferrari F, Forcato M, Pizzini S, Bortoluzzi S, Bicciato S. The Virtual-Chip: a new approach for the integration of different oligonucleotide arrays. RECOMB 2010 – Fourteenth International Conference on Research in Computational Molecular Biology, August 12-15 2010, Lisbon, Portugal

Bolstad, B.M., Irizarry, R.A., Astrand, M., Speed, T.P., 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinforma. Oxf. Engl. 19, 185–193.

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G.G., Oezcimen, A., Rocca-Serra, P., Sansone, S.-A., 2003. ArrayExpress--a public repository for

microarray gene expression data at the EBI. Nucleic Acids Res. 31, 68–71.

Castagna, A., Polati, R., Bossi, A.M., Girelli, D., 2012. Monocyte/macrophage proteomics: recent findings and biomedical applications. Expert Rev. Proteomics 9, 201–215.

Clark, E.A., 1997. Regulation of B lymphocytes by dendritic cells. J. Exp. Med. 185, 801–803.

Conesa, A., Nueda, M.J., Ferrer, A., Talón, M., 2006. maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. Bioinforma. Oxf. Engl. 22, 1096–1102.

Dai, M., Wang, P., Boyd, A.D., Kostov, G., Athey, B., Jones, E.G., Bunney, W.E., Myers, R.M., Speed, T.P., Akil, H., Watson, S.J., Meng, F., 2005. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res. 33, e175.

Dempster et al., 1977] Dempster, A., Laird, N., Rubin, D. & Others (1977). Journal of the Royal Statistical Society. Series B (Methodological)

Edgar, R., Domrachev, M., Lash, A.E., 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 30, 207–210.

Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J.J., Gardner, T.S., 2007. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. Plos Biol. 5, e8.

Fallarino, F., Volpi, C., Fazio, F., Notartomaso, S., Vacca, C., Busceti, C., Bicciato, S., Battaglia, G., Bruno, V., Puccetti, P., Fioretti, M.C., Nicoletti, F., Grohmann, U., Di Marco, R., 2010. Metabotropic glutamate receptor-4 modulates adaptive immunity and restrains neuroinflammation. Nat. Med. 16, 897–902.

Feng, J., Liu, T., Qin, B., Zhang, Y., Liu, X.S., 2012. Identifying ChIP-seq enrichment using MACS. Nat. Protoc. 7, 1728–1740.

Ferrari, F., Bortoluzzi, S., Coppe, A., Sirota, A., Safran, M., Shmoish, M., Ferrari, S., Lancet, D., Danieli, G.A., Bicciato, S., 2007. Novel definition files for human GeneChips based on GeneAnnot. BMC Bioinformatics 8, 446.

Ganguly, D., Haak, S., Sisirak, V., Reizis, B., 2013. The role of dendritic cells in autoimmunity. Nat. Rev. Immunol. 13, 566–577.

Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., Noble, W.S., 2007. Quantifying similarity between motifs. Genome Biol. 8, R24.

Hirano, F., Tanaka, H., Hirano, Y., Hiramoto, M., Handa, H., Makino, I., Scheidereit, C., 1998. Functional interference of Sp1 and NF-kappaB through the same DNA binding site. Mol. Cell. Biol. 18, 1266–1274.

Iglesias, M.J., Jesus Iglesias, M., Reilly, S.-J., Emanuelsson, O., Sennblad, B., Pirmoradian Najafabadi, M., Folkersen, L., Mälarstig, A., Lagergren, J., Eriksson, P., Hamsten, A., Odeberg, J., 2012. Combined chromatin and expression analysis reveals specific regulatory mechanisms within cytokine genes in the macrophage early immune response. Plos One 7, e32306.

Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B., Speed, T.P., 2003. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. 31, e15.

Italiani P., Mazza E.M.C., Cifola I., Lucchesi D., Mandriani B., Bicciato S., Battaglia C., Boraschi D.: Reprogramming of human monocytes activation during inflammation, kinetical profiling of response by integrated in vitro analysis. Toll 2011 Meeting - Decoding Innate Immunity; Riva del Garda, 4-7 Maggio 2011

Johnson, W.E., Li, W., Meyer, C.A., Gottardo, R., Carroll, J.S., Brown, M., Liu, X.S., 2006. Model-based analysis of tiling-arrays for ChIP-chip. Proc. Natl. Acad. Sci. U. S. A. 103, 12457–12462.

Kuska, B., 1998. Beer, Bethesda, and biology: how “genomics” came into being. J. Natl. Cancer Inst. 90, 93.

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinforma. Oxf. Engl. 25, 1754–1760.

Machanick, P., Bailey, T.L., 2011. MEME-ChIP: motif analysis of large DNA datasets. Bioinforma. Oxf. Engl. 27, 1696–1697.

Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., Califano, A., 2006. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 7 Suppl 1, S7.

Martinez, F.O., 2011. Regulators of macrophage activation. Eur. J. Immunol. 41, 1531–1534.

Matzinger, P., 2007. Friendly and dangerous signals: is the tissue in control? Nat. Immunol. 8, 11–13.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., Bejerano, G., 2010. GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. 28, 495–501.

Park, P.J., 2009. ChIP-seq: advantages and challenges of a maturing technology. Nat. Rev. Genet. 10, 669–680.

Ruggeri F., Faltin F. & Kenett R., Encyclopedia of Statistics in Quality & Reliability, Wiley & Sons (2007).

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., Friedman, N., 2003. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat. Genet. 34, 166–176.

Sales, G., Calura, E., Martini, P., Romualdi, C., 2013. Graphite Web: Web tool for gene set analysis exploiting pathway topology. Nucleic Acids Res. 41, W89–97.

Sica, A., Mantovani, A., 2012. Macrophage plasticity and polarization: in vivo veritas. J. Clin. Invest. 122, 787–795.

Song, J.S., Johnson, W.E., Zhu, X., Zhang, X., Li, W., Manrai, A.K., Liu, J.S., Chen, R., Liu, X.S., 2007. Model-based analysis of two-color arrays (MA2C). Genome Biol. 8, R178.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P., 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. 102, 15545–15550.

Tusher, V.G., Tibshirani, R., Chu, G., 2001. Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. U. S. A. 98, 5116–5121.

Zambelli, F., Pesole, G., Pavesi, G., 2009. Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. Nucleic Acids Res. 37, W247–252.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., Liu, X.S., 2008. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 9, R137.

PUBLICATIONS

Pallotta MT, Orabona C, Volpi C, Vacca C, Belladonna ML, Bianchi R, Servillo G, Brunacci C, Calvitti M, Bicciato S, Mazza EMC, Boon L, Grassi F, Fioretti MC, Fallarino F, Puccetti P, Grohmann U. Indoleamine 2,3-dioxygenase is a signaling protein in long-term tolerance by dendritic cells. Nat Immunol. 2011 Jul 31;12(9):870-8.

Italiani P, Mazza EMC, Lucchesi D, Cifola I, Gemelli C., Grande A, Battaglia C, Bicciato S, Boraschi D. Transcriptomic profiling of the development of the inflammatory response in human monocytes in vitro, Plos One 2013, under revision

Bessede A, Gargaro M, Pallotta M, Matino D, Servillo G, Brunacci C, Bicciato S, Mazza EMC, Macchiarulo A, Vacca C, Iannitti R, Tissi L, Volpi C, Belladonna ML, Orabona C, Bianchi R, Lanz T, Platten M, Della Fazia MA, Piobbico D, Zelante T, Funakoshi H, Nakamura T, Gilot D, Denison MS, Guillemin GJ, DuHadaway JB, Prendergast GC, Metz R, Geffard M, Boon L, Romani L, Veyret B, Grohmann U, Puccetti P, Fallarino F. Endotoxin tolerance as a disease tolerance defense pathway involving tryptophan catabolism and the aryl hydrocarbon receptor. Nature 2013, under revision

Gemelli C., Zanocco Marani T., Bicciato S., Mazza EMC., Boraschi D., Parenti S., Selmi T., Tagliafico E. Ferrari S. and Grande A. MafB is a downstream target of the IL-10 / STAT3 signaling pathway involved in the regulation of macrophage de-activation. BBA Molecular Cell Research 2013, under-revision

Zoso A, Mazza EMC, Bicciato S, Mandruzzato S, Bronte V, Serafini P, Inverardi L. Molecular and cellular characterization of Cord Blood derived, IDO expressing human fibrocytic Myeloid Derived suppressor Cells. Blood, under submission

ABSTRACTS & POSTERS

Gemelli C., Zanocco Marani T., Bicciato S., Mazza E.M.C., Parenti S., Vignudelli T., Tagliafico E.,Grande A., Ferrari S.: Identification of novel MAFB primary response genes in phorbol 12-myristate13-acetate (PMA) dependent macrophage differentiation. AIBG 2010 - XII National Congress of the Italian Association of General and Molecular Biology and Genetics (AIBG); Trento, 8-9 October 2010.

Boraschi D., Italiani P., Mazza E.M.C., Gemelli C., Cifola I., Lucchesi D., Rossi R.,Bicciato S., Grande A., Mauri P., Battaglia C.: Modulation of macrophage activation programming during inflammation. Integrated profiling in an in vitro kinetical model of human monocyte reactivity. 14th International Congress of Immunology; Kobe,Giappone, 22-27 Agosto 2010.

Bisognin A., Mazza E.M.C., Ferrari F., Forcato M., Pizzini S., Bortoluzzi S., Bicciato S.: The Virtual Chip: a new approach for the integration of different oligonucleotide arrays. Fourteenth International Conference on Research in Computational Biology; Lisbona, Portogallo, 12-15 Agosto 2010.

Bisognin A., Mazza E.M.C., Ferrari F., Forcato M., Bicciato S., Bortoluzzi S.: Metaanalysis of Microarray raw data using a virtual integrated platform. BITS 2010 – VII Annual Meeting of the Bioinformatics Italian Society; Bari, 14-16 Aprile 2010.

Italiani P., Mazza E.M.C., Cifola I., Lucchesi D., Mandriani B., Bicciato S., Battaglia C., Boraschi D.: Reprogramming of human monocytes activation during inflammation, kinetical profiling of response by integrated in vitro analysis. Toll 2011 Meeting – Decoding Innate Immunity; Riva del Garda, 4-7 May 2011.

Boraschi D, Italiani P., Lucchesi D., E.M.C. Mazza, Bicciato S., Cifola I.,Battaglia C., Costantino L.: Exploring the interaction of nanomedicines with the human immune system. Assessing immunosafety and investigating preventive/therapeutic

targeting. Fourth European CLINAM-Conference for Clinical Nanomedicine. 22-25 May 2011, Basel, Switzerland, Europe

Italiani P, Costantino MD, Cifola I, Mazza EMC, Bicciato S, Battaglia C, Boraschi D. IL-1 family cytokines in the inflammatory response of human monocytes. IL-1 family members and the inflammasome. Dublin, September 15-16 2011

Gemelli C., Grande A., Zanocco Marani T., Bicciato S., Mazza E.M.C., Ferrari S. MafB is a transcriptional regulator of IL-10 mediated macrophage polarization process." XIV National Congress of the Italian Association of General and Molecular Biology and Genetics (AIBG). Assisi 28-29 September 2012

Mazza E. M. C., Italiani P., Valsoni S., Battaglia C., Boraschi D., Bicciato S. A bioinformatics approach to elucidate networks of molecular interactions in a human model of physiological inflammation; Rimini, 11-13 October 2012

Mazza EMC, Italiani P, Valsoni S, Cifola I, Boraschi D, Battaglia C, Bicciato S. A bioinformatics pipeline to uncover regulatory modules and their condition-specific regulators in human monocyte-to-macrophages differentiation and polarization. Front Immunol 22 Aug 2013

Mazza EMC, Italiani P, Valsoni S, Cifola I, Boraschi D, Battaglia C, Bicciato S. A bioinformatics pipeline to uncover regulatory modules and their condition-specific regulators in human monocyte-to-macrophages differentiation and polarization. 15th International Congress of Immunology (ICI). Milan, 22-27 Aug 2013

ACKNOWLEDGEMENTS

I would like to thank Prof. Silvio Bicciato for the guidance, encouragement and advice during these years. I thank Prof. Cristina Battaglia for the suggestions she has provided during my PhD and for her patient support. I want to thank all the people with whom I have collaborated for my thesis work in particular: Dr. Diana Boraschi, Dr. Paola Italiani, Prof. Alexis Grande and Dr. Claudia Gemelli.

This work was supported by funding from PRIN 2007 Y84HTJ and with a PhD fellowship from AIRC Special Program Molecular Clinical Oncology '5 per mille' grant.