Making Sense of Molecular Signatures in The Immune System

Nicholas J. Davies*, Mahlet G.T. Tadesse#, Marina Vannucci#, Hugh Kikuchi*, Victor Trevino*, Donatella Sarti*, Ilaria Dragoni+, Andrea Contestabile§, Edward Zanders~ and Francesco Falciani*¶

**School of Biosciences, the University of Birmingham, Birmingham B15 2TT, UK, #Department of Statistics, Texas A&M University, College station, TX 77843-3143, ~Purely Proteins Ltd, 254 Milton Road, Cambridge CB4 0WE, UK, +Novartis Institute for Molecular Sciences, London, UK, §Department of Human and General Physiology, University of Bologna, P.zza Porta S. Donta 2, 40127 Bolgogna, Italy*

Abstract: The development of Functional Genomics technologies has opened new avenues to investigate the complexity of the immune system. Microarray technology has been particularly successful because of its relatively low cost and high genome coverage.

Consequently to our ability to monitor the expression of a significant proportion of an organism genome, our understanding of the molecular dynamics behind cell differentiation and cell response has greatly improved. Molecular signatures associated to immune cells have provided important tools to investigate the molecular basis of diseases and have been often associated to diagnostic and prognostic markers. The availability of such large collection of data has stimulated the application of complex machine learning techniques in the attempt to link molecular signatures and cell physiology. Here we review the most recent developments in the analysis of molecular signatures in the immune system.

Keywords: Microarray, Gene Expression Profiling, Immune System, Lymphocytes, Bioinformatics.

1. INTRODUCTION

In the last few years we have seen the development and diffusion of Functional Genomics technologies. These techniques allow monitoring the expression and the interaction of thousands of genes, proteins and metabolites in a single experiment. Among these, gene expression profiling has so far made the greatest impact in Biology. With a relatively small investment it is possible to monitor a substantial amount of the transcriptional capacity of an organism. As a consequence large amounts of information have become available in the public domain creating issues of data management and analysis previously unseen in Biology.

Microarray technology and other large scale functional genomics technologies have provided the means to characterise the molecular state of cells and tissue to an unprecedented level of detail. This has contributed to improve our understanding of how cells develop through interactions with their environment and how, by altering their molecular state, they develop new functions. Understanding and characterising the global molecular state of a cell not only helps answering basic questions in biology but also provide a path towards understand diseases.

A key issue in Biology has always been to understand how cells differentiate from a multi-potent progenitor cell. So far, the identification of key genes in the differentiation process and the use of antibodies against key differentiation markers have provided a valid experimental approach to dissect the developmental process leading to a mature cell. Although very powerful this approach is limited by the number of markers that can be measured in single experiments. A Genome-wide picture of developing cells can be presumably much more revealing and could contribute to identify more comprehensive signatures associated to specific developmental stages.

Microarray technology has reinforced the concept that cell differentiation is just a transition between relatively stable cell states and that as differentiation progresses the repertoire of genes expressed is reduced. When a mature cell is stimulated (for example with a cytokine) it undergoes a transformation that can be reversible or it can lead to a new stable state. A given cell type can reveal surprising heterogeneity when stimulated. Microarray technology has demonstrated to be an extremely suitable experimental approach to dissect and characterise these high level dynamics. The identification of molecular signatures of quiescent and activated immune and stromal cells has contributed to increase our understanding of the pathology of many diseases and in some cases it has provided useful clinical markers.

In this paper we will review how expression profiling has contributed to modify our view of cell differentiation in the immune system and how molecular signatures associated to specific "cell states" can be predictive of physiological and clinical parameters. We will also show some applications of probabilistic modelling techniques to the analysis of microarray data in the immunological settings.

2.0 DEFINING CELL IDENTITY

2.1 Stem Cells

Cell specification starts from the very early stages of development from multi-potent embryonic stem (ES) cells that have the ability to differentiate along all lineages. At more advanced stages of development and in the adult organisms' stem cells, such as haematopoietic stem cells

*Address correspondence to this author at the School of Biosciences, the University of Birmingham, Birmingham B15 2TT, UK; E-mail: f.falciani@bham.ac.uk

1386-2073/04 $45.00+.00

(HSCs) still retain the potential to differentiate within a specific lineage. Conventional techniques have identified a number of genes such as Oct3/4, FGF4, Sox2, and genes involved in the Leukaemia Inhibitory Factor (LIF) signal transduction (reviewed in [1, 2]) that appear to be important in maintaining an undifferentiated state. Many of these factors are not only at the basis of cell differentiation but are also important in the process of tissue regeneration.

Despite the importance of these genes many more factors and pathways are likely to play a role in the differentiation process. Moreover, our understanding of the differentiation process itself rely on our ability to identify homogeneous cell types that we interpret as intermediate steps in the maturation of fully differentiated cells. These cell types have been traditionally defined and characterised with a relatively small set of molecular markers.

Since Microarray technology offers the possibility to characterise the molecular profile of a cell at a genome level it is reasonable to expect that it would provide a much clearer image of the differentiation process. Once cells are characterised at the genome level would we still see differentiation as a series of discrete transitions between defined states? Is response to cytokines such as TNF-α a simple transient response to stimulation or it lead to a proper change in cell identity, perhaps in part similar to the differentiation process itself?

These and other questions have been the drive for many of the application of microarray technology over the last few years.

In this context Microarray technology has been applied to characterise different stem cell populations in the attempt to define a *molecular signature defining a totipotent cell*. This has been done both with human and murine cell lines [3, 4]. By comparing the expression profiles of murine embryonic, neural and haematopoietic stem cells (HSCs) with those of

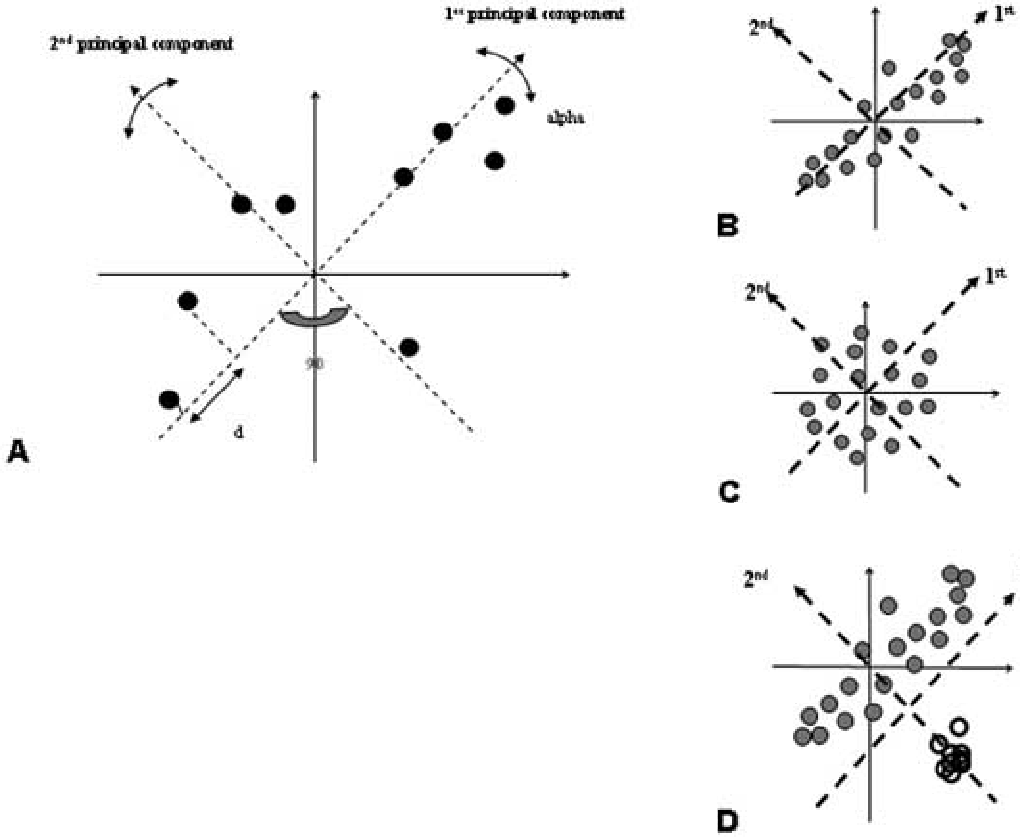

Fig. (1). Principal Component Analysis. This series of cartoons summarise the basic principals behind Principal Component analysis. (**A**) This panel display in a graphical format a simple two dimensional dataset (nine genes measured in two conditions). The objective of PCA in this case is to represent the relative similarity in the expression profile of the nine genes using a single axis instead of the two required to describe the dataset in full. The approach taken is projecting the individual genes on a new axis that is chosen to maximise the distance **d** between the projected points. This new axis (first principal component) therefore explains the majority of the information that can be represented in on a single axis. A second component can then be generated orthogonal to the first one. Clearly the approach is not useful on a two dimensional dataset but its application to complex microarray experiments with hundreds of dimensions is extremely useful. (**B**) This is a schematic example of a situation where PCA is very effective in describing the overall information contained on the data. The first component in this case represents the large majority of the information. (**C**) In this situation the data are distributed symmetrically around the centre. In this case, the first component explains as much variation as the original axis making the application of PCA not very effective. (**D**) In this case the second component summarizes important information. In fact, it allows the identification of a small cluster of genes (open circles) far from the main data cloud.

differentiated cells Ramalho-Santos [3] have identified transcripts enriched in totipotent cells that are potential candidates to sustain stem cells in their undifferentiated state. They defined a "stem cell expression signature" containing ESTs of unknown function as well as genes downstream of the JAK/ STAT, TGF-β, Yamaguchi sarcoma kinase and Notch signalling pathways; growth hormone and thrombin receptors; genes that are capable of interacting with the extra cellular matrix and genes associated with stress resistance. They used Cluster analysis (a data mining technique used to graphically represent the similarity of the transcriptional profile of different biological samples) to compare progenitor cells with their more mature counterparts. The aim of their analysis was to identify at what stage in the differentiation process cells begin to demonstrate a profile that is significantly different from a multi-potent cell state.

The HSCs most closely resembled the main population of bone marrow cells suggesting that when ES cells differentiate into HSCs they already acquire many of the lineage traits that will characteristic of more differentiated cells. Interestingly, neural stem cells more closely resembled ES cells than samples from lateral ventricles of the brain, which suggests that the acquisition of neuronal phenotypes occurs at later stages of the lineage.

Further insights into the development of the Haematopoietic linage has been emerging from a comparison of the transcriptional profiles of four cell populations at different stages of haematopoietic differentiation [5]. FACS-purified populations of haematopoietic stem cells (HSCs), non-self-renewing mutlipotential precursors (MPPs), common lymphoid precursors (CLPs) and common myeloid precursors (CMPs) were isolated from C57B6-J mice, ensuring the same genetic background [5]. Lineage restricted CMPs and CLPs can differentiate into cells of a myeloid lineage, such as neutrophils and monocytes, or lymphoid cells, such as B, T and natural killer cells, respectively. Where possible the differentially expressed genes identified through microarray analysis were classified according to whether they were associated with haematopoietic or non-haematopoietic tissues, such as brain, liver and kidney. This study demonstrated that as cells become more dedicated to a particular lineage, the more that they become restricted with respect to their gene expression. Indeed, the transcriptional profile of HSCs contained a large proportion of non-haematopoietic-affiliated genes. This is in keeping with reports suggesting that murine HSC enriched bone marrow samples can give rise to nonhaematopoietic tissues [6, 7]. Other transcripts identified in the HSC populations were associated with self-renewal, cell growth arrest, immortalization of cells and cell commitment. These genes were moderately expressed in MPPs, but were strongly downregulated in both the CLPs and CMPs, as would be expected because these precursors have become quite limited with respect to differentiation. MPPS have the potential to acquire cells of both the lymphoid and myeloid lineage and as expected the MPP-representative genes included both myeloid- and lymphoid-specific genes, as well as other haematopoiesis-related genes and genes associated with cell cycle progression and proliferation. There was a great difference in the genes found to be representative for CLP and CMP cells. A lot of the CMP genes were quite highly expressed in MPPs and HSCs, but were not expressed in CLPs, whereas the CLP-representative genes were barely expressed in the other cell types.

Utilising the potential provided by microarray technology to characterise cells at the genome level has provided new insights into the mechanisms that are involved in maintaining a multi-potent cell state and has proved the validity of this approach to study the dynamics of cell differentiation.

2.2 The Characterization of Lymphocyte Populations

One of the earliest applications of microarray technology was to identify signatures representative of different lymphocytes subsets [8]. Using a microarray developed to be representative of a lymphocyte transcriptional capacity. The authors profiled a number of samples representative of activated and resting T and B cells. Not surprisingly, T cell receptor (TCR) genes and genes downstream of the TCR signalling pathway were identified as components of a T cell signature [9]. A Germinal Centre (GC) B cell populations was also discovered to have a characteristic transcriptional signature including genes encoding proteins involved in cell cycle progression, DNA synthesis and protein translation [8, 10].

In the development of an inflammatory reaction B cells further differentiate in plasma cells (PCs). These cells, specialized in the production of large quantities of Immunoglobulin, represent the end-point of B cell differentiation. There are three different subtypes of PC. PCs found in the lymph nodes develop from naïve B cells, are short-lived and secrete unmutated immunoglobulins (Ig). B cells expressing Ig with high specificity for antigen can be selected in the germinal centre to differentiate into short-lived PCs secreting high levels of Ig upon persistent antigen exposure. B cells with mutated Ig, of the germinal centre can migrate to the bone marrow where the microenvironment can support them as PCs [11, 12]. In order to describe the molecular relationships of these subpopulations of cells, Tarte *et al.* [13] have used a popular analysis technique called principal Component Analysis (PCA). The objective of this technique is to represent the molecular state of biological samples in a two or three dimensional plot while retaining the majority of the information contained in the original microarray profile (an overview of this technique can be found in Fig. (**1**). Using this technique Tarte *et al.* were able to identity three distinct subtypes of plasma cells. The technique allow to identifying gene subsets that are associated to a defined Component. Using this approach the authors were able to demonstrate that the first Principal Component is primarily associated to B cell development whereas the second component seemed to be primarily associated to activation or proliferation.

2.3 Response to Stimulation

Similar approaches to the one described in the previous paragraph can be used to study the response of mature immune cells to relevant stimuli. The activation of mature T lymphocytes is a critical step in the activation of the immune system. The process requires a main signal, associated to the activation of the T cell receptor complex, and a secondary is through co-receptors such as CD28 [14]. Unstimulated T cells are relatively quiescent but upon activation they rapidly proliferate, become highly motile and produce mediators of the immune response, such as the pro-inflammatory cytokine IL-2. Eventually, inflammation clear and the majority of T cells undergo apoptosis.

In order to characterise the transcriptional phenotype of stimulated T cells Teague *et al.* [15] have used a mouse model of antigen stimulation. Mice were injected with superantigen and T cells were subsequently purified and samples processed for microarray analysis. The expression signature of quiescent T cells included genes whose products are associated with the suppression of the cell cycle [15]. This suggests that resting T cells actively maintain their quiescent state and are not just dormant. There was little difference observed in the number of genes expressed in T cells extracted eight or forty eight hours after exposure to antigen. The profiles of eight hour exposed T cells expressed genes associated with cell division, cytokines and cytokine receptors. The signature of T cells at forty eight hours after exposure to antigen closely resembled that of resting, unstimulated T cells, but genes associated with the induction of apoptosis, cell cycle progression and metabolism were upregulated. This study has demonstrated that T cell stimulation is a transient state and that T cells that do not undergo apoptosis at the end of the immune response appear to return to a quiescent state similar to that before their exposure to antigen.

In the absence of co-stimulation activation is sub-optimal and may lead to anergy or apoptosis [16]. For these reason Diehn *et al.* [17] have used an *in vitro* model of T cell activation to investigate the transcriptional response of primary and secondary activation signals. T cells were stimulated using anti-CD3 and anti-CD28 antibodies. These are activating antibodies that mimic the physiological activation of the TCR via antigen presentation and through the CD28 co-stimulatory pathway. The transcriptional profile of *in vitro* TCR stimulation alone was very similar to that of co-stimulation, it was just less intense. The signal transduced through CD28 appears to synergise with TCR

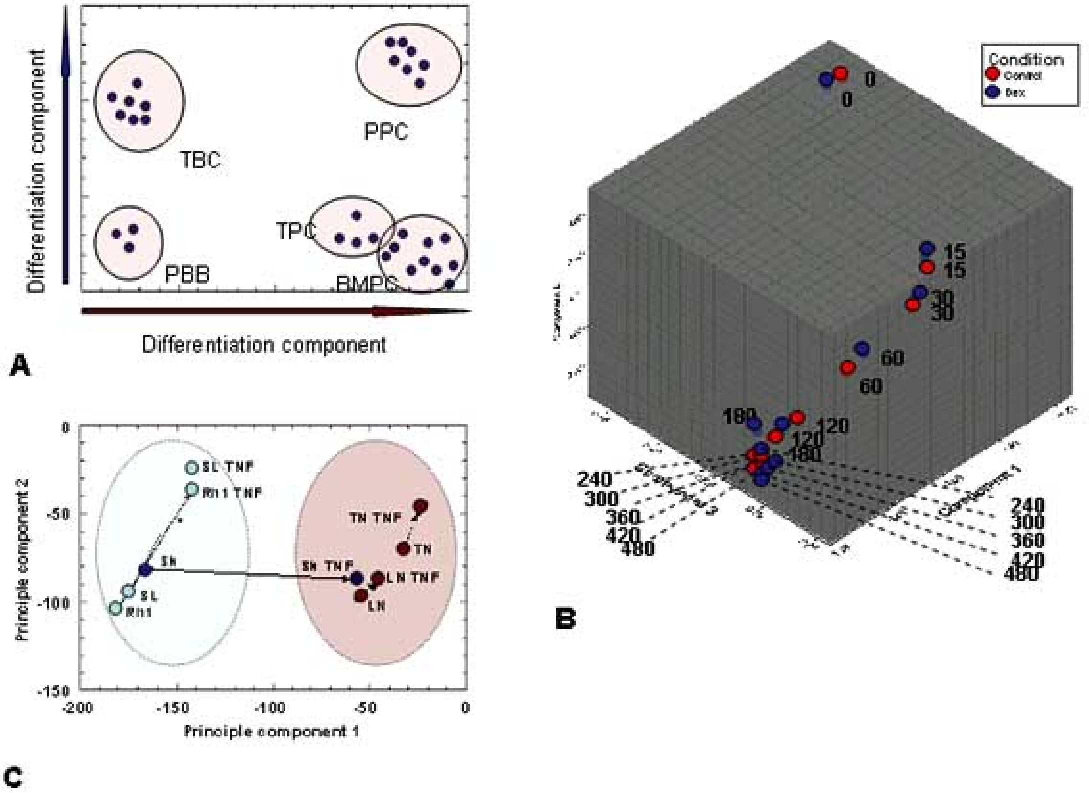

Fig. (2). (**A**) Semi-schematic representation of principal component analysis representing the molecular state of different B cell populations. The analysis clearly separates four groups of cells and reveals that genes that contribute to the first component represent differentiation, whereas genes associated to the second component are primarily representing genes involved in proliferation. (TBC indicates tonsil B cells; PBB, peripheral blood B cells; TPC, tonsilar plasma cells; BMPC, bone marrow plasma cells; PPC, polyclonal plasma cells). (**B**). Principal component analysis plot representing the response to TNF-α and IL-1 of human fibroblast cells derived from the synovial tissue of RA patients. Red circles represent samples taken from cells treated with TNF-α and IL-1 (control) whereas blue circles represent cells treated with these cytokines in combination with the anti-inflammatory drug dexamethasone. Numbers associated to circles represent time after treatment in minutes. (**C**). Semi-schematic representation of principal component analysis of fibroblast cells derived from different anatomical sites and treated with TNF-α. The blue oval represents an area of the PC space that is associated to cells derived from non-lymphoid organs. The red oval represents an area of the PC plot that is associated to fibroblast derived from lymphoid organs. Arrows represent cell state transitions in response to cytokine treatment.

signalling to produce optimal T cell activation. As it may be expected the most up-regulated genes in the co-stimulation experiment included IL-2, IL-2 receptor α and GM-CSF, which are targets of NFAT [18]. Their findings were supported by the fact that blocking of NFAT phosphorylation largely alleviated the synergistic effect of CD28 co-stimulation. Their results show that combined activation of TCR and CD28 co-stimulatory pathways modulate the response of target of the TCR pathway, presumably through NFAT.

2.4 The Identity of Stromal Cells

Immune cells such as B and T cells perform their functions in the context of the tissue microenvironment. Consequently, interaction with other cells is therefore crucial to determine the outcome of inflammation. Fibroblast cells produce a variety of factors that are contributing to establish the tissue micro-environment.

Fibroblast from different anatomical sites or even different subpopulations of fibroblasts from the same site however shows different morphologies, molecular markers and phenotypic behaviour. In the context of an inflammatory reaction stromal cells are involved in regulating the recruitment and survival of lymphocytes and play a major role in determining the intensity and outcome of chronic inflammatory reactions. In chronic inflammatory diseases such as rheumatoid arthritis (RA) stromal cells play an important role in sustaining the inflammatory process. In this context microarray technology can help to answer a number of important questions. What is the dynamics of response of these cells to important cytokines such as TNF alpha? Can we define a stable molecular state that defines chronically activated fibroblasts? Do fibroblast cells, derived from different anatomical sites, display a different response to cytokines?

In order to address the first question our group have used microarray technology to profile fibroblast cells derived from the joints of RA patients after treatment with TNF-α and IL-1. Fig. (**2B**) shows that response to cytokine treatment is accompanied to dramatic molecular transformations in the first 15 minutes of response. In the subsequent two hours, cells converge to a new stable state characterised by the expression of high levels of proteases, cytokines and chemokines (data not shown). Interestingly, we observed that although treatment with the anti-inflammatory drug dexamethasone reduces the expression of some of these key inflammation genes, it fails to revert the overall dynamics of response.

Parsonage *et al.* [19] have used molecular similar profiling techniques to characterise the response to cytokines of fibroblast cells derived from a variety of lymphoid and non lymphoid organs. Their two time point analysis was designed to observe the transition from the resting state to a stable activated state we have just described. Fig. (**2C**) shows the results of a Principal Component Analysis describing the response of fibroblast cells from Skin, RA joints and from two lymphoid organs (lymph nodes and tonsils) to TNF-α. This analysis first reveal that the molecular profile of cells derived from non lymphoid organs is very different from the one derived from non lymphoid organs. Of particular interest is the observation that skin fibroblast are extremely similar to synovial fibroblast derived from RA patients, in resting state, and that these are clearly separated from fibroblasts derived from lymphonode and tonsils. The analysis of the dynamics of response to TNF-α however, reveals profound differences. Skin and RA fibroblasts in fact respond with completely uncorrelated transcriptional programs. Interestingly TNF-α treated skin fibroblasts become almost identical to lymph node fibroblasts. This analysis reveals the great potential of stromal cells to modulate the production of cytokines and chemokines and other inflammatory mediators and how this potential is determining the tissue micro-environment where immune effector cells operate. Using co-culture experiments Parsonage *et al.* have then demonstrated that heterogeneity at the molecular level corresponds to differential effects on the biology of lymphocytes proving the importance of stromal cells in determining the tissue-specificity of the inflammatory reaction.

3.0 UNDERSTANDING THE MOLECULAR BASIS OF IMMUNE DISEASES

3.1 Stem Cell Signatures in Acute Myeloid Leukaemias

Chromosomal translocations are common in acute myeloid leukaemias (AMLs). They can generate fusion proteins that are responsible for the priming of progenitor cells with the potential to develop into leukaemic cells. Invariably one component of these AML-associated fusion proteins is a transcription factor involved in the regulation of cell differentiation [20]. A reasonable expectation is that formation of these fusion proteins may result in the dysregulated gene expression, especially the transcriptional regulation of genes involved in myeloid differentiation and possibly proliferation. Alcalay *et al.* [21] tested this hypothesis profiling an haematopoietic cell line expressing the three most commonly occurring AML-fusion proteins. A large number of genes were identified as being differentially regulated by at least two of the fusion proteins. Of these 163 genes were regulated by all three and 265 were found to be significantly upregulated by two fusion-proteins, but down regulated by the third. The majority of the genes fell into two categories, those of metabolism, or growth signalling, which includes growth factors and signalling molecules. Importantly, twenty-one known genes were identified whose products are involved in the regulation of differentiation, fifteen of which have been implicated in haematopoiesis. Of these genes, those that were induced by the fusion proteins are associated with stem cells or early progenitors, whereas the repressed genes are mainly expressed at later stages of differentiation.

Their results fit well with the observation that a build up of haematopoietic precursors blocked at a particular stage of development occurs in AMLs. Interestingly a significant number of genes down-regulated by the AML-associated fusion proteins were involved in DNA repair, implying a reduced ability to repair damaged DNA.

3.2 Signatures of Immune Cells are Predictive of Disease Outcome

In the previous paragraph we have seen how expression profiling can reveal important clues on the molecular events that may be at the onset of AML. Similarly, molecular signatures characteristic of mature immune cells can be used to investigate the molecular and cellular bases of inflammatory diseases.

Alizadeh *et al.* [8] applied this principal by profiling patients affected by follicular lymphomas (FL), chronic lymphoblastic leukaemia (CLL) and diffuse large B cell lymphomas (DLBCL). They first realized that patients affected by DLBCLs, the most common form of non-Hodgkins lymphoma were showing a marked molecular heterogeneity. Samples extracted from these patients were clustering into two distinct groups. One subgroup clustered closely to germinal centre B cell samples whilst the other closely resembled the signature of activated B cells.

Follicular lymphomas were found to cluster very closely with the germinal centre B cells, which is in line with the suggestion that transformation of B cells to FL occurs whilst the cell is within the germinal centre [22]. Chronic lymphoblastic leukaemias (CLLs) clustered closely to resting B cells as well. The identification of two distinct sub-clusters for DLBCL suggested a new definition for this disease: a GC-like DLBCL and an activated B-like DLBCL. The significance of this classification was confirmed by the different clinical outcome of these patients. Activated B-like classified patients showed significantly worst survival rates confirming the validity of gene signature based classification.

One of the major achievements of Alizadeh *et al.* was in showing that the complexity of a clinical sample could be understood as a combination of elementary gene expression signatures and that these are predictive of clinical outcome.

CLL has a variable prognosis depending upon whether the rearranged immunoglobulin (Ig) genes are somatically mutated or not, with unmutated Ig genes having a worse prognosis [23, 24]. It has been possible to build a CLL gene signature irrespective of the Ig mutation using microarrays by comparing CLL samples to other lymphomas and lymphocytes [8, 25]. The genes used to classify CLL are highly expressed in germinal centre or peripheral blood B cells. They also included a set of genes that can be used to distinguish resting B cells from activated and germinal centre B cells. Unsupervised clustering was used to try to identify genes that may be used as predictors of Ig mutational status [25].

This approach was unable to discern the difference between Ig V_H mutated and unmutated samples. Using 56 predictor genes that were identified as being significantly expressed only in one group it was possible to differentiate between the two groups. CLL subtype predictors based on the three most differentially expressed genes (ZAP-70, activation-induced C-type lectin and IM286077) were found to be able to perform with 100% accuracy.

3.3 Molecular Heterogeneity in Chronic Inflammatory Diseases

Rheumatoid Arthritis (RA) is a chronic inflammatory disease affecting the synovial tissue in joints. Clinically it is a heterogeneous disease mediated by both immune and non-immune cellular systems [26, 27]. Transcriptional profiling of synovial biopsy samples from RA patients has been used to identify key pathways or genes that are differentially expressed between mild and more severe cases [28]. Cluster analysis of the RA samples generated three subgroups that differed according to the levels of inflammatory genes. The high-inflammation (RA^{hi}) subgroup has elevated expression of genes associated with T/B cell and antigen presenting cells (APCs). The low-inflammation (RA^{lo}) subgroup has repressed expression of T/ B cell genes and variable levels of APC genes. The third group was an intermediate-inflammation group that had slightly elevated T/B cell genes and variable APC genes. Clinically and genetically RA is a heterogenous disease. One reason for this may be in the expression of genes involved in the signal transducer and activator of transcription 1 (STAT-1) pathway, elements of which were identified through a comparison of the transcriptional profiles of RA^{hi} with RA^{lo} samples. STAT-1 plays an important role in controlling cell cycle progression and apoptosis.

4.0 TOWARDS MODELLING GENE EXPRESSION AT THE GENOME LEVEL

The studies we have described above have contributed to improve our understanding of the immune system. In most cases however the data have been analysed using descriptive approaches. These techniques (e.g. Clustering, Principal Component Analysis) are extremely powerful but do not provide any robust prediction of the behaviour of a Biological system. Moreover, the identification of "interesting" genes has been often driven by biological knowledge rather than from objective statistical criteria. In order to overcome these limitations a number of approaches based on relatively complex machine learning techniques have been proposed (reviewed in [29]). Modelling the immune system is not a prerogative of Functional Genomics. Mathematical approaches to simulate the behaviour of the immune response have been in fact developed in the past twenty years (reviewed in [30]). Modelling complex biological systems using expression profiling data however present many unusual challenges. Most important of all is the extremely large number of variables (genes) that need to be examined. In order to be effective each individual set of genes needs to be evaluated as a whole in order to determine if the combination of the selected genes is effective. Although this concept is intuitively obvious it is computationally very demanding. The simple task of selecting a 10 gene signature from a dataset of 25 would in fact require evaluating 3268760 combinations of genes. It is easy to realize that, since a typical microarray experiment involve several thousands of genes, an exhaustive search is computationally impossible.

With a strategy that allows the identification of reasonably small gene sets it is reasonable to assume that the interpretation of the models would be feasible and that hypothesis on the molecular mechanisms behind the predictive associations could be made.

Our group has developed a Bayesian variable selection strategy [31, 32] and has applied this to identify molecular signatures predictive of disease stage in a population of RA patients. Our model appears to be extremely effective in predicting disease stage but also suggest interesting hypothesis. The best predictive model (5% error) was based on eight genes. Despite the fact that these genes were selected from a large microarray five of these genes were part of a multi-protein complex involved in cytoskeleton remodelling and adhesion. In addition, Connexin 40, Notch 4 and the Adenosine Receptor were suggesting cells in the peripheral blood of late RA patients could be characterised by a reduced ability to express the pro-inflammatory cytokine IL-2 in response to activation. Moreover, an important gene in the model was JunB, a transcription factor involved in the control of Proliferation and programmed cell death.

In addition to develop models predictive of a defined class of disease it is possible to construct models that are predictive of the expression of a marker gene.

In order to explore the potential role of JunB in the synovial tissue of RA patients we have developed a regression model to identify gene signatures predictive of the level of JunB. Our analysis has been performed using a modelling technique that uses a similar variable selection strategy to the one described above but coupled to a regression model [33 and references therein]. Our regression models were able to predict the expression of JunB with reasonable accuracy (62% of the variability was explained) from an expression signature based on 21 genes (table 1). The inclusion of cell cycle and apoptosis genes (6) is consistent with the role of this gene in controlling these crucial cellular functions. The fact that such association is constructed from data derived from the synovial tissue of a population of RA patients strongly suggests that JunB may have a primary role in maintaining inflammation in the joints. A number of genes involved in signalling and with a demonstrated effect in the immune system were also included in the model. Interestingly we discovered that Connexin 40 and Calcineurin A1 (two genes involved in the model predicting early versus late stage RA from peripheral blood) were also important in the prediction of JunB expression.

More information can be acquired from modelling the evolution of a Biological system in a time domain. Using advanced modelling techniques it is in fact possible to reverse engineer gene networks from functional genomics data. Bayesian reverse engineering techniques have been applied to model time course gene expression profiling data [34]. [35-37] have been the first to apply reverse engineering approaches to a problem of immunological interest: the activation of human T lymphocytes. Their models infer a number of known molecular pathways in the activation of human mature T lymphocytes suggesting the validity of the approach.

Reverse engineering techniques are extremely powerful but are still experimental. One of the more severe limitations so far is the need of large amounts of replicated data to infer causal relationships. Simulation studies have shown [38] that beyond the theoretical limitation of the techniques a 50 gene network can be inferred by a 500 replication experiment. These types of experiments are possible with current technologies but the cost of producing data to model a larger number of cellular components would be extremely prohibitive.

CONCLUSION

The use of microarray technology in immunology has contributed to change our understanding of many basic concepts. The most relevant impact has been redefining the concept of cell identity. Our perception of a cell has in fact changed dramatically since we have been able to observe the transcriptional state of a significant proportion of an organism genome. The analysis of lymphocyte and stromal cell populations has for example revealed unexpected dynamics of response to cytokines with implications on the way that these cells interact in the tissue micro-environment. The analysis of clinical samples has revealed that it is possible to associate molecular signatures of cell populations to generate hypothesis on the molecular basis of the disease it self. Such signatures have also demonstrated to be very powerful in predicting clinical outcome giving a new spin to clinical research.

The development of statistical models predictive of phenotypic variables has also been demonstrated to be a useful approach to formulate hypothesis on mechanisms of disease progression and more generally on the molecular basis of cell physiology.

Although the use of microarray technology is rapidly becoming a standard technique in laboratories, there are still many problems that remain open. More suitable data analysis approaches are required to face the many challenges in the analysis of large scale expression data. Of these, statistical modelling techniques will be key for making sense of these data. Ultimately our ability to characterise the molecular state of a cell and the dynamics of its response create an extremely complex problem. Integrating existing biological knowledge with this new perception of a cell molecular state is a great challenge. Ultimately we must be able to merge all this information and model the behaviour of complex systems to, for example predict the effects of an increase of cytokine levels. It seems that microarray technology has stimulated a renewed interest in the modelling and machine learning fields but it has also increased the complexity of the problem.

ACKNOWLEDGEMENTS

We thank Prof Christopher Buckley and Prof Mike Salmon (MRC Centre for immunoregulation, Birmingham, UK) for their enthusiastic support and for the many scientific discussions that lead to the ideas described in this paper. We also thank Prof. David Wild (Keck Graduate Institute, CA, USA) and Dr Chris Bunce (School of Biosciences, Birmingham, UK) for his enthusiastic support in the course of the last few years.

REFERENCES

[1] Burdon, T.; Smith, A.; Savatier, P. *Trends. Cell. Biol.,* **2002,** *12,* 432.

[2] Niwa, H. *Cell Struct. Funct.*, **2001,** *26*, 137.

[3] Ramalho-Santos, M.; Yoon, S.; Matsuzaki, Y.; Mulligan, R. C.; Melton, D. A. *Science,* **2002,** *298,* 597.

[4] Sperger, J. M.; Chen, X.; Draper, J. S.; Antosiewicz, J. E.; Chon, C. H.; Jones, S. B.; Brooks, J. D.; Andrews, P. W.; Brown, P. O.; Thomson, J. A. *Proc. Natl. Acad. Sci. USA,* **2003,** *100,* 13350.

[5] Akashi, K.; He, X.; Chen, J.; Iwasaki, H.; Niu, C.; Steenhard, B.; Zhang, J.; Haug, J.; Li, L. *Blood,* **2003,** *101,* 383.

[6] Graf, T. *Blood,* **2002,** *99,* 3089.

[7] Lagasse, E.; Shizuru, J. A.; Uchida, N.; Tsukamoto, A.; Weissman, I. L. *Immunity,* **2001,** *14,* 425.

[8] Alizadeh, A. A.; Eisen, M. B.; Davis, R. E.; Ma, C.; Lossos, I. S.; Rosenwald, A.; Boldrick, J. C.; Sabet, H.; Tran, T.; Yu, X.; Powell, J. I.; Yang, L.; Marti, G. E.; Moore, T.; Hudson, J. Jr.; Lu, L.; Lewis, D. B.; Tibshirani, R.; Sherlock, G.; Chan, W. C.; Greiner, T. C.; Weisenburger, D. D.; Armitage, J. O.; Warnke, R.; Levy, R.; Wilson, W.; Grever, M. R.; Byrd, J. C.; Botstein, D.; Brown, P. O.; Staudt, L. M. *Nature,* **2000,** *403,* 503.

[9] Alizadeh, A. A.; Staudt, L. M. *Curr. Opin. Immunol.,* **2000,** *12,* 219.

[10] Ma, C.; Staudt, L. M. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **2001,** *356,* 83.

[11] O'Connor, B. P.; Cascalho, M.; Noelle, R. J. *J. Exp. Med.,* **2002,** *195,* 737.

[12] Slifka, M. K.; Antia, R.; Whitmire, J. K.; Ahmed, R. *Immunity,* **1998,** *8,* 363.

[13] Tarte, K.; Zhan, F.; De Vos, J.; Klein, B.; Shaughnessy, J. Jr. *Blood,* **2003,** *102,* 592.

[14] Salomon, B.; Bluestone, J. A. *Annu. Rev. Immunol,* **2001,** *19,* 225.

[15] Teague, T. K.; Hildeman, D.; Kedl, R. M.; Mitchell, T.; Rees, W.; Schaefer, B. C.; Bender, J.; Kappler, J.; Marrack, P. *Proc. Natl. Acad. Sci. USA,* **1999,** *96,* 12691.

[16] Van Parijs, L.; Abbas, A. K. *Science,* **1998,** *280,* 243.

[17] Diehn, M.; Alizadeh, A. A.; Rando, O. J.; Liu, C. L.; Stankunas, K.; Botstein, D.; Crabtree, G. R.; Brown, P. O. *Proc. Natl. Acad. Sci. USA,* **2002,** *99,* 11796.

[18] Macian, F.; Garcia-Rodriguez, C.; Rao, A. *EMBO J.,* **2000,** *19,* 4783.

[19] Parsonage, G.; Falciani, F.; Burman, A.; Filer, A.; Ross, E.; Bofill, M.; Martin, S.; Salmon, M.; Buckley, C. D. *Thromb. Haemost.,* **2003,** *90,* 688.

[20] Alcalay, M.; Orleth, A.; Sebastiani, C.; Meani, N.; Chiaradonna, F.; Casciari, C.; Sciurpi, M. T.; Gelmetti, V.; Riganelli, D.; Minucci, S.; Fagioli, M.; Pelicci, P. G. *Oncogene,* **2001,** *20,* 5680.

[21] Alcalay, M.; Meani, N.; Gelmetti, V.; Fantozzi, A.; Fagioli, M.; Orleth, A.; Riganelli, D.; Sebastiani, C.; Cappelli, E.; Casciari, C.; Sciurpi, M. T.; Mariano, A. R.; Minardi, S. P.; Luzi, L.; Muller, H.; Di Fiore, P. P.; Frosina, G.; Pelicci, P. G. *J. Clin. Invest.,* **2003,** *112,* 1751.

[22] Bahler, D. W.; Levy, R. *Proc. Natl. Acad. Sci. USA,* **1992,** *89,* 6770.

[23] Damle, R. N.; Wasil, T.; Fais, F.; Ghiotto, F.; Valetto, A.; Allen, S. L.; Buchbinder, A.; Budman, D.; Dittmar, K.; Kolitz, J.; Lichtman, S. M.; Schulman, P.; Vinciguerra, V. P.; Rai, K. R.; Ferrarini, M.; Chiorazzi, N. *Blood,* **1999,** *94,* 1840.

[24] Hamblin, T. J.; Davis, Z.; Gardiner, A.; Oscier, D. G.; Stevenson, F. K. *Blood,* **1999,** *94,* 1848.

[25] Rosenwald, A.; Staudt, L. M. *Semin. Oncol.,* **2002,** *29,* 258.

[26] Kraan, M. C.; Haringman, J. J.; Post, W. J.; Versendaal, J.; Breedveld, F. C.; Tak, P. P. *Rheumatology (Oxford),* **1999,** *38,* 1074.

[27] Tak, P. P.; Smeets, T. J.; Daha, M. R.; Kluin, P. M.; Meijers, K. A.; Brand, R.; Meinders, A. E.; Breedveld, F. C. *Arthritis Rheum.,* **1997,** *40,* 217.

[28] van der Pouw Kraan, T. C.; van Gaalen, F. A.; Kasperkovitz, P. V.; Verbeet, N. L.; Smeets, T. J.; Kraan, M. C.; Fero, M.; Tak, P. P.; Huizinga, T. W.; Pieterman, E.; Breedveld, F. C.; Alizadeh, A. A.; Verweij, C. L. *Arthritis Rheum.,* **2003,** *48,* 2132.

[29] Szabo, A.; Boucher, K.; Carroll, W. L.; Klebanov, L. B.; Tsodikov, A. D.; Yakovlev, A. Y. *Math Biosci.,* **2002,** *176,* 71.

[30] Morel, P. A., *Front Biosci,* **1998,** *3,* d338.

[31] Sha, N.; Vannucci, M.; Brown, P. J.; Trower, M.; Amphlett, G.; Falciani, F. *Comp. Funct. Genom.,* **2003,** *4,* 171.

[32] Sha, N.; Vannucci, M.; Tadesse, M. G.; Brown, P. J.; Dragoni, I.; Davies, N.; Roberts, T. C.; Contestabile, A.; Salmon, M.; Buckley, C.; Falciani, F. *Biometrics,* **2004,** *In Press.*

[33] Brown, P. J.; Vannucci, M.; Fearn, T. *J. Roy. Stat. Soc. B.,* **2002,** *64,* 519.

[34] Kim, S. Y.; Imoto, S.; Miyano, S. *Brief Bioinform,* **2003,** *4,* 228.

[35] Beal, M. J.; Falciani, F.; Ghahramani, Z.; Rangel, C.; Wild, D. *Submitted to Bioinformatics,* **2004**.

[36] Rangel, C.; Angus, J.; Ghahramani, Z.; Lioumi, M.; Sotheran, E.; Gaiba, A.; D., W.; Falciani, F. *Bioinformatics,* **2004,** *In Press.*

[37] Rangel, C.; Gharamani, Z.; Gaiba, A.; Wild, D.; Falciani, F. Proceedings of the International Conference on System biology, November **2001**.

[38] Rangel, C.; Angus, J.; Ghahramani, Z.; Wild, D. L. In *Applications of probabilistic modelling in medical informatics and bioinformatics,* Husmeier, D.; Roberts, S.; Dybowski, R. Ed. Springer-Verlag: New York, **In Press**.

Received: 03 March, 2004 **Accepted: 26 March, 2004**