The immune system as a biomonitor : explorations in innate and adaptive immunity.

Benjamin Chain, Mahdad Noursadeghi, John Shawe-Taylor, Toyin Alabi, Judy Breuer, Guy Danon, Eleanor Gray, James Heather, Theres Matjeka, Gabriel Pollara, Elspeth Potton, Niclas Thomas, Nandi Simpson.

Division of Infection and Immunity, MRC Centre for Medical Molecular Virology, UCL, London , UK

The human immune system has a high complex multi-layered structure which has evolved to detect and respond to changes in the internal microenvironment of the body. Recognition occurs at the molecular or submolecular scale, via classical reversible receptor:ligand interactions, and can lead to a response with great sensitivity and speed. Remarkably, recognition is coupled to memory, such that responses are modulated by events which occur years or even decades before. Although the immune system in general responds differently and more vigorously to stimuli entering the body from the outside (e.g. infections) this is an emergent property of the system : many of the recognition molecules themselves have no inherent bias towards external stimuli but bind with targets found within the body and targets found externally. It is quite clear that the immune response registers pathophysiological changes in general . Cancer, wounding and chronic tissue injury are some obvious examples.

Against this background, the immune system “state” tracks the internal processes of the body, and is likely to encode information about possible current and former disease processes. Moreover the distributed nature of most immune responses (e.g. typically involving lymphoid tissue, non-lymphoid tissue, bone marrow , blood , extracellular interstitial spaces etc.) means that many of the changes associated with immune responses are manifested systemically, and specifically can be detected in blood. This provides a very convenient route to sampling immune cells.

We consider two different and complementary ways of querying the human immune “state” using high dimensional genomic screening methodologies, and discuss some of the potentials of this approach and some of the technological and computational challenges to be overcome.

The first approach focuses on innate immune responses. A growing number of receptors are being discovered which mediate recognition of molecular patterns independently of the classical antibody and T cell receptors on lymphocytes. Some of these are activated by molecules found only on microorganisms. Toll-like receptor 4 which recognises lipopolysacharide or TLR5 which recognises flagellin are examples. Others recognise very common macromolecules, but detect their presence in abnormal locations. The DNA and RNA sensors are examples of these classes. Yet another class, recognise molecules released from dying or damaged cells. Ligand binding to these receptors activates complex intracellular signalling pathways which in turn often trigger further release of singling molecules such as the interferons , cytokines or chemokines which propagate and amplify the initial recognition event. The many different signals are integrated by different cell types in different ways and give rise to specific patterns of gene transcription which can be evaluated using whole genome transcriptomics. We and others are exploring the possibility that features of the gene expression profile (or signature) of white blood cells may be mapped to underlying pathological events, and this information could therefore be used for diagnosis, and stratification of patients.

We have developed a preliminary pipeline for collection and analysis of such data sets using the Agilent microarray technology[1] . As a proof of concept of this approach, we tested the hypothesis that gene expression profiles in patients presenting with fever of unknown origin would be more diverse than the profiles of healthy volunteers.Patients with acute febrile illness were recruited from University College London Hospital, and whole blood samples were collected in PAXgene RNA tubes. Extracted RNA was hybridized to Agilent 4x44K Human Gene Expression Arrays. Leukocyte subset analyses were performed on blood samplescollected concurrently, to assess the relationship between gene expression signatures and changes in leukocyte composition.We found that gene expression profiles from patients with fever could be distinguished from healthy volunteer profiles. After normalisation for neutrophil counts, the data showed that patients are readily distinguished from healthy volunteers using unsupervised clustering, and crucially that there is greater variance in expression profiles from patients than from healthy volunteers. Furthermore, expression data from patients clustered in distinct groups. Surprisingly, components of the data with greatest variance were not necessarily genes with a known immunological function.

The analysis of these data sets pose very significant challenges. The data sets are very high dimensional (typically in excess of 40,000 signals per sample). The data dimensionality is typically far higher than the number of samples, and there is a complex pattern of correlation between different dimensions. The data is also noisy. All these features limit the efficiency of traditional classification approaches. Furthermore, model building is hampered by the fact that the function and interrelationships between most genes is still unknown. We will discuss some ongoing attempts at deriving meaningful functional signatures of gene response patterns by selecting optimum sparse inverse correlation matrices from the data.

An important challenge now is to build robust, platform independent, and easily queried repositories of these data sets. This will lead to increasing number of robust data sets with detailed clinical annotations, which can be integrated with the rapidly increasing body of knowledge regarding gene function and interaction. In turn this will yield computational models which can be used to reliably infer the pathological events which underlie the observed patterns in gene expression.

The gene expression profiles discussed above provide a window into the qualitative and quantitative response of the immune system at the time of sampling, and in the recent (hours or days) past. In contrast, the adaptive immune system profile can reflect cumulative changes over years or even decades. The classical approach to monitoring adaptive immunity has been to test for specific antigen responses, using antibody or T cell based assays. The ability to measure the frequency of every lymphocyte receptor in a sample of blood using high throughput parallel sequencing offers the opportunity to analyse the adaptive repertoire at a global level. Since the basis of adaptive immunity is the clonal expansion of lymphocytes with specific receptors (each of which is rare or even unique at a sequence level) , the frequency distribution of different TcR receptor sequences in a sample is one measure of the immune status of the individual.

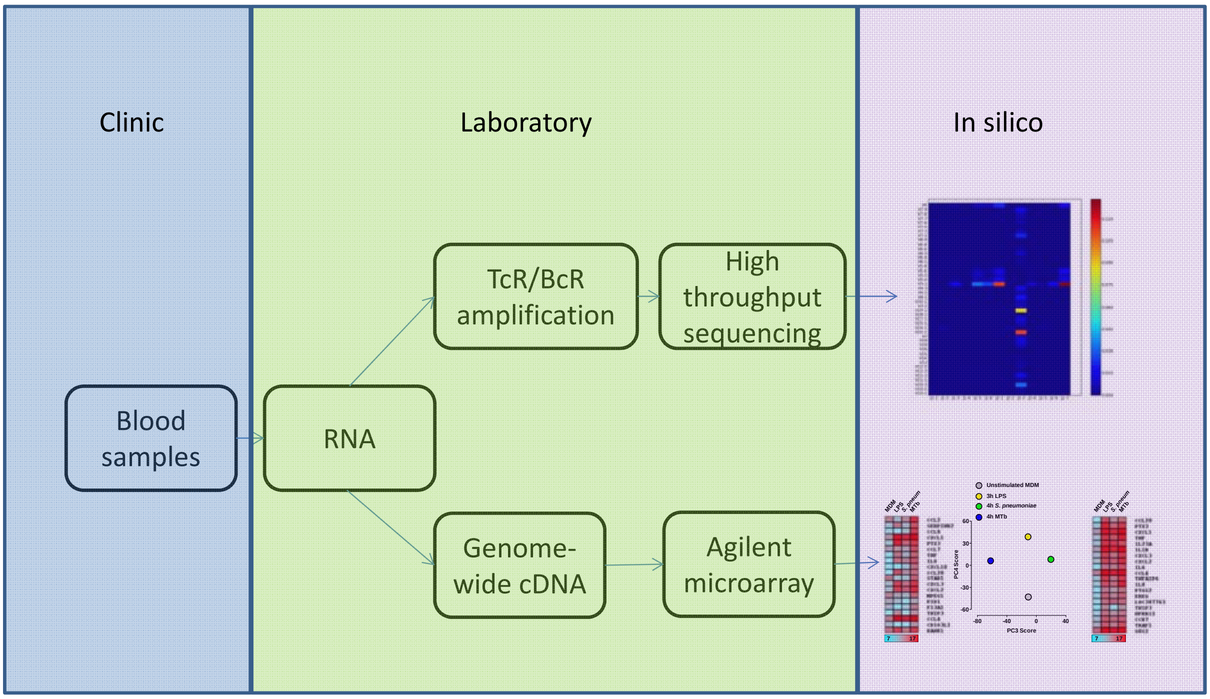
We are developing a pipeline for this type of analysis, focusing predominantly on the T cell repertoire, although easily extendible to B cell responses. An RNA sample is collected exactly as for gene expression profiling above, and in practice the same sample can easily be analysed for both expression profiling and TcR frequency profiling. The TcR(both alpha and beta chains) are amplified using unbiased PCR, with a constant region primer in combination with RACE. The expanded library is gel purified, and then analysed using either Illumina “sequencing by synthesis” or Roche 454 pyrosequencing. The relative advantages and disadvantages of the data produced by these two technologies will be discussed. We are currently developing a pipeline for analysis of these large data sets. Efficient algorithms (based on the Aho-Corasick string search[2]) for assigning V and J segment classification to individual reads, and identifying and cataloguing deletions and additions at the V and J boundaries are being developed and will be discussed. As for the transcriptomics data discussed above, the utility of these data sets in diagnosis, prognosis and hence patient management will depend on the establishment of easily queried robust databases, with accurate and reliable clinical annotation.

In conclusion,we discuss two complementary powerful tools for querying the immune status of an individual,by genome wide transcriptomics and by T or B cell antigen receptor frequency profiling of peripheral blood cells. Implementation of robust standardised laboratory and computational protocols, and establishment of secure and easily queried databases linked to informative clinical data will be the crucial factors if these novel technologies are exploited effectively for improved patient care.

References

1: Chain B, Bowen H, Hammond J, Posch W, Rasaiyaah J, Tsang J, Noursadeghi M.Error, reproducibility and sensitivity: a pipeline for data processing of Agilentoligonucleotide expression arrays.BMC Bioinformatics.2010;11:344.

2. Aho, Alfred V.; Margaret J. Corasick (June 1975). "Efficient string matching: An aid to bibliographic search". Communications of the ACM 18 (6): 333–340.



Pipeline for leukocyte transcriptomics analysis