New tools for classification and monitoring of autoimmune diseases

Holden T. Maecker, Tamsin M. Lindstrom, William H. Robinson, Paul J. Utz, Matthew Hale, Scott D. Boyd, Shai S. Shen-Orr and C. Garrison Fathman

Abstract | Rheumatologists see patients with a range of autoimmune diseases. Phenotyping these diseases for diagnosis, prognosis and selection of therapies is an ever increasing problem. Advances in multiplexed assay technology at the gene, protein, and cellular level have enabled the identification of 'actionable biomarkers'; that is, biological metrics that can inform clinical practice. Not only will such biomarkers yield insight into the development, remission, and exacerbation of a disease, they will undoubtedly improve diagnostic sensitivity and accuracy of classification, and ultimately guide treatment. This Review provides an introduction to these powerful technologies that could promote the identification of actionable biomarkers, including mass cytometry, protein arrays, and immunoglobulin and T-cell receptor high-throughput sequencing. In our opinion, these technologies should become part of routine clinical practice for the management of autoimmune diseases. The use of analytical tools to deconvolve the data obtained from use of these technologies is also presented here. These analyses are revealing a more comprehensive and interconnected view of the immune system than ever before and should have an important role in directing future treatment approaches for autoimmune diseases.

Maecker, H. T. *et al. Nat. Rev. Rheumatol.* **8**, 317–328 (2012); doi:10.1038/nrrheum.2012.66

Introduction

Biomarkers—biological characteristics that can be objectively evaluated as indicators of a biological or pathological state—are being sought for many diseases. Biomarkers have the potential to transform our basic understanding and clinical management of a wide range of human illnesses. We have coined the term 'actionable biomarkers' to describe biomarkers that can inform clinical practice—that is, biomarkers upon which clinicians can act (Figure 1).

Actionable biomarkers are already used in the clinical management of certain diseases, most notably cancer. A prime example is the *BCR–ABL1* fusion gene of t(9;22) chromosomal translocations, which, in the correct clinical context, can be used to identify patients with chronic myelogenous leukaemia who are likely to respond to therapy with drugs that target the activity of the tyrosine protein kinase ABL1.[1] Likewise, overexpression of the receptor tyrosine-protein kinase erbB2 (also known as HER2) characterizes the subset of patients with breast cancer who are likely to respond to treatment with a monoclonal antibody that targets the erbB2 receptor.[2] These two success stories illustrate how molecular characteristics that are linked to disease pathogenesis, rather than clinical characteristics (which are generally a disease epiphenomenon), are most likely to serve as actionable biomarkers. In these examples a single biomarker suffices; in other cases, however, a panel of multiple biomarkers is more useful as it can yield a more comprehensive picture (termed a molecular signature) of a disease and its subtypes.[3–6] In fact, in rheumatic diseases, only profiling using multiple biomarkers has so far proven useful.

One potential use for actionable biomarkers is in diagnosing disease. First, by casting a wide net, combinations of biomarkers might be identified that improve both the sensitivity and specificity of disease detection and classification. Second, by revealing a molecular signature of disease before the onset of definitive, characteristic symptoms, biomarkers might enable earlier diagnosis and therefore earlier institution of therapeutic, or even preventive, interventions. For example, biomarkers that can distinguish individuals with early-stage rheumatoid arthritis (RA) from patients with undifferentiated arthritis—or better yet from asymptomatic individuals who are genetically predisposed to develop RA—would be invaluable as evidence suggests that early intervention with existing drugs could prevent RA progression.[7]

As illustrated earlier, a potential use for actionable biomarkers is in predicting how an individual's disease will develop. As all known rheumatic diseases are heterogeneous, they do not manifest identically in all patients, nor do all patients respond to treatment in the same way. For example, RA ranges from mild and self-limiting to severe and progressive. In our opinion, stratification into subtypes is important for the clinical management of a disease and we propose that actionable biomarkers could aid this subtyping. Stratification of disease could help

Competing interests
S. D. Boyd declares an association with the following company: ImmuMetrix LLC. See the article online for full details of the relationship. The other authors declare no competing interests.

Department of Microbiology and Immunology (**H. T. Maecker, M. Hale**), Department of Medicine (**T. M. Lindstrom, W. H. Robinson, P. J. Utz, C. G. Fathman**), Department of Pathology (**S. D. Boyd**), Stanford University School of Medicine, Stanford, CA 94305, USA. Department of Immunology, Rappaport Faculty of Medicine, Technion-Israel Institute of Technology, Haifa 32000, Israel (**S. S. Shen-Orr**).

Correspondence to: C. G. Fathman cfathman@stanford.edu

Key points

- Antigen arrays are valuable for profiling autoantibodies in diverse rheumatic autoimmune diseases and can be composed of most biomolecules including proteins, peptides, protein complexes, sugars, nucleic acids and lipids
- High-throughput DNA sequencing enables the tracking of disease-associated clones of T cells and B cells in autoimmune diseases; changes in populations of these cells can be correlated with therapeutic response
- The analysis of peripheral blood cells following cellular activation might be important in identifying clinically actionable biomarkers
- New technologies enable analysis of gene and protein expression in whole blood samples; deconvolution of datasets reveals which immune-cell subset underlies a change without isolating or manipulating the cells

clinicians determine whether an individual's condition is likely to progress, and therefore whether aggressive intervention is needed, as well as select and establish an effective treatment strategy. For example, less than two-thirds of all individuals with RA have an adequate response to anti-TNF therapy.[8] Using appropriate biomarkers might enable identification of non-responders before TNF-inhibitor therapy is initiated, thereby lowering costs and preventing unwanted complications associated with a therapy that was not going to be effective. Emerging reports of autoantibody profiles that can predict disease progression in so-called incomplete lupus,[9] predict which patients will develop RA,[10] or predict which patients with RA will respond to anti-TNF therapy,[11] suggest that biomarker-based predictive tests will become as much a mainstay in the management of rheumatic diseases as they currently are in cancer.

Actionable biomarkers can also be used to monitor a patient's response to specific therapies. Such pharmacodynamic biomarkers can accelerate clinical trials by serving as early surrogate markers of the efficacy and safety of an investigational drug as well as guide clinicians as to when a given therapy should be initiated.

Systems immunology

The nascent field of systems immunology, a branch of systems biology, uses computational mathematical modelling to characterize the immune system and predict its response when a specific component is affected. New technological approaches that can generate vast multiplex datasets have enabled the development of this field. Indeed, more than 40,000 mRNA transcripts from the human genome can now be routinely measured in a single microarray (a technology that provides details on which genes are expressed in a tissue or cell of interest). Multiplexed Luminex TM (Luminex Corporation, Austin, TX, USA) assays can quantitate 50 or more proteins that are involved in inflammation (that is, cytokines and chemokines) in a single small sample of tissue or blood; protein arrays can measure many more.

Additionally, new flow-cytometric methods are now available to simultaneously analyze the expression of 30 or more surface and intracellular proteins in individual cells. This technology promotes the identification and enumeration of the various peripheral blood cells in addition to revealing, for instance, which signalling pathways are activated in the different cell types.

A successful systems immunology study requires that the assays employed are as comprehensive as possible, and that they also possess sufficient resolution to distinguish the changes that accompany differential outcomes. Investigators of systems immunology are increasingly measuring a plethora of signals in response to an experimental intervention, such as a vaccine.[12] Complex signatures emerging from such studies can act as biomarkers, and also provide clues to the mechanistic pathways that lead to specific outcomes, such as protection from disease. In addition, sufficient assay standardization and sample handling, including standardization of processing and storage protocols, are essential for a study to achieve reproducible results over time. This standardization is particularly important for studies in human immunology, which often involve longitudinal sampling, collection of specimens from multiple sites, and/or subject recruitment that can span multiple years.

New approaches in systems immunology

New immunological technologies provide novel types of highly multiplexed readouts, with the potential to measure the activation induced *in vitro* by a given intervention, as well as resting immune phenotypes of cells (Figure 1). For example, individual differences in activation-induced signalling, but not in resting expression levels of certain phosphoproteins, correlate with disease outcome in acute myeloid leukaemia.[13] Therefore, measuring changes in activation-induced signalling in rheumatic autoimmune diseases, using a flow cytometry based technique, might lead to changes in the clinical management of these diseases.

Mass cytometry

Cytometry by time-of-flight (CyTOF) mass cytometry uses multiple antibodies, each tagged with multiple copies of an individual heavy metal ion, and measures their binding to cells by mass spectrometry.[14] By contrast, fluorescence cytometry is used to measure the binding of antibodies tagged with a fluorophore. The advantage of mass cytometry is that many more antibodies can be used in combination to assay a single sample (such as whole blood or single-cell suspensions from tissues), without the inherent spillover between fluorescence spectra that is inherent in optical fluorescence systems.[15] Such a system has already been used to quantitate differences in cellular constitution and drug responses of individual cells in a complex mixture of cells such as bone marrow.[16]

In one of the authors' laboratories, 36 different metal ions have been chelated to polymers that have then been conjugated to antibodies, DNA dyes, or other markers (H. T. Maecker, unpublished work). In most cases, the resolution and sensitivity of mass cytometry are comparable to those of fluorescence flow cytometry, although generating a sufficiently sensitive reagent has so far not been possible for a few cellular markers. As experience with this approach increases, and with the availability of pre-made heavy metal ion–antibody conjugates for mass cytometry, this problem should be resolved. Moreover, as the number of mass cytometry systems in use increases,

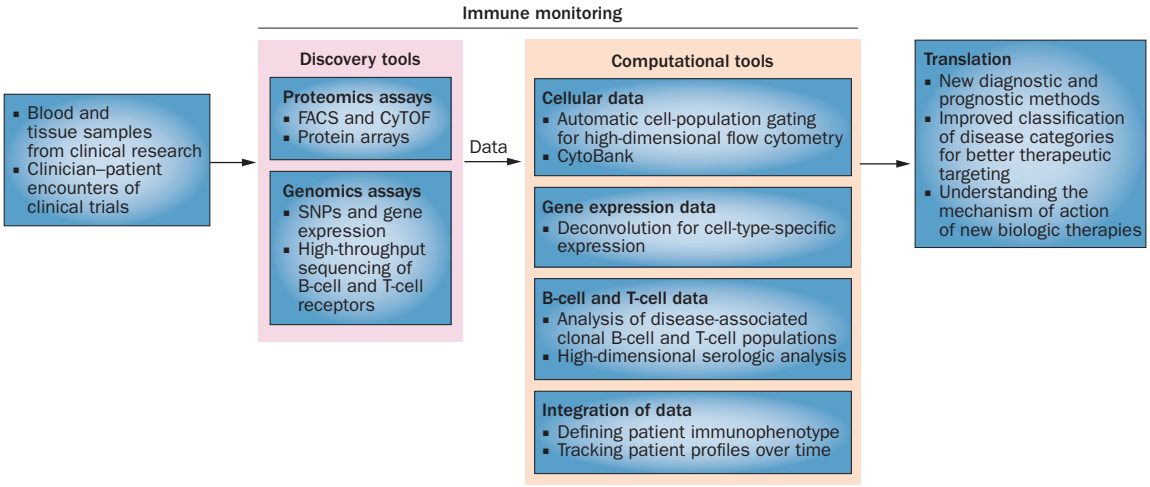

Figure 1 | Application of new immune-monitoring technologies to rheumatology. Samples for biomarker discovery can be generated during clinical research and actual clinical trials. For comprehensive immune monitoring, these samples are subjected to multiple assays at the proteomic and genomic level. Moreover, computational tools are applied to organize and better analyze the complex data sets that are generated, as well as to integrate heterogeneous data types. The end result should be the discovery of new actionable biomarkers, which aid disease diagnosis, prognosis, therapeutic targeting and contribute knowledge to the mechanism of action of a specific therapy. Abbreviations: CyTOF, cytometry by time of flight; FACS, fluorescence-activated cell sorting; SNP, single nucleotide poymorphism.

mass cytometry is likely to become the preferred method for initial multi-parameter flow-cytometric analysis, especially as the cost per marker analyzed is similar to that of fluorescence systems.

Analyzing complex flow cytometry datasets

A number of new analysis platforms such as HyperCyt® (IntelliCyt Corporation, 9620 San Mateo Blvd NE, Alberquerque, NM 87113, USA)[17] and CyTOF[16] are vastly increasing the sample throughput and number of independent proteomic parameters that can be measured at the single cell level. The data collected in a single day, if reviewed by conventional methods, would require viewing many thousands of bivariate plots. This approach is not only inefficient, but also results in an incomplete understanding of the multidimensional relationships present in the underlying data. Effective automated gating and specialized tools for visualizing high-dimensional flow cytometry data are crucial areas of development.

In 2009, two automated gating methods—flow analysis with automated multivariate estimation (FLAME)[18] and density-based merging (DBM)[19]—were developed, both of which are highly promising but that use very different approaches (Figure 2). By contrast, spanning-tree progression analysis of density-normalized events (SPADE),[20] a tool developed for visualizing high-complexity flow cytometry data, foregoes traditional gating and bivariate plots altogether.

Automated gating methods—FLAME and DBM

FLAME is based on the assumption that a sample of flow cytometry data can be modelled as a heterogeneous mixture of populations of cells (known as clusters) in which each cluster can be described by a skewed Student's t distribution (skew-t distribution).[18] The skew-t distribution better fits asymmetrical populations than traditional gating approaches that are based on Gaussian mixture modelling. FLAME is designed to create an optimal number of clusters by comparing the average scale-free intracluster distance with the average scale-free intercluster distance. If the optimal number of populations has been assigned, the average scale-free intracluster distance will be smaller than the average scale-free intercluster distance (Figure 2).

FLAME seems to be effective when the populations can be distinguished by surface markers whose expression is binary. However, certain combinations of markers, such as those used in the study of cell cycle and differentiation, have staining patterns that are too irregular to be well-approximated by the skew-t distribution. These combinations include distributions with concave perimeters or distributions with 'U', 'L', or 'S' shapes. Fortunately, DBM uses the density contours of the data itself to define the gates for each population and is better-suited for irregularly shaped distributions than FLAME.[19] DBM detects inflection points in the data, much as experienced immunologists do when gating manually. Unlike FLAME, DBM becomes computationally inefficient beyond three dimensions.

FLAME and DBM are marked advances in automated cell-population gating, which is of great importance for complex datasets that can require the gating of a large number of distinct cell populations across each biological sample in the dataset. However, manually reviewing all of the automatically assigned gates to confirm that they have been properly applied can be time consuming.

Visualizing flow cytometry data—SPADE

As an alternative to automated gating approaches that attempt to approximate manual gating, SPADE[16,20] is a visualization tool that organizes clusters into a 2D tree representation on the basis of their similarities across all

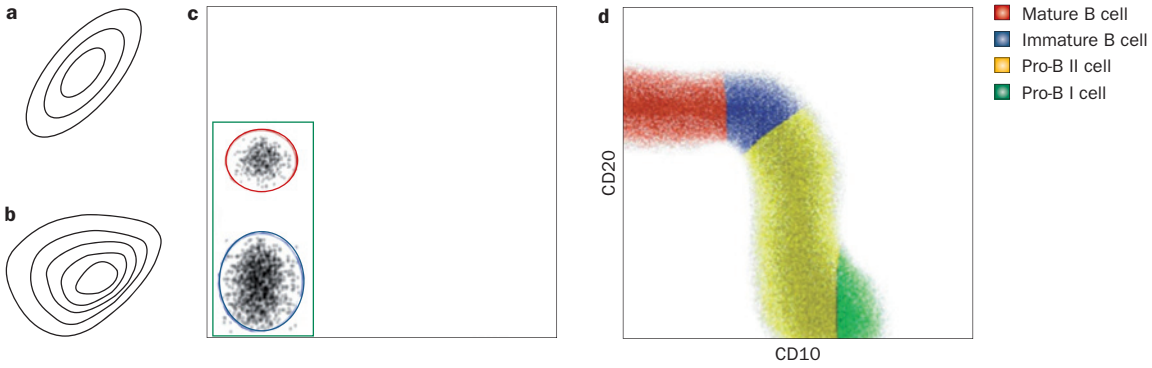

Figure 2 | Alternative analysis approaches for high-complexity flow cytometry data. **a** | Example of a bivariate Gaussian distribution as used in Gaussian mixture modelling. **b** | Example of a bivariate skew-t distribution as used in FLAME. **c** | Comparison of average intracluster distance and average intercluster distance. The average distance between events within the green gate (intracluster distance) is very large so it is likely to be composed of multiple distinct populations. The average distance between events within the red gate or within the blue gate (intracluster distance) is much smaller than the average distance between events in the red and blue gates (intercluster distance). **d** | Illustration of flow cytometry data showing normal human B cell development in bone marrow.[95] Continuous distributions such as this poorly fit with Gaussian mixture modelling, FLAME, or DBM, but the phenotypic relationships are well-visualized by SPADE. Abbreviations: DBM, density-based merging; FLAME, flow analysis with automated multivariate estimation; SPADE, spanning-tree progression analysis of density-normalized events.

markers selected by the user. By displaying clusters in a 2D tree structure, and using size and colour to denote cell density and marker expression, SPADE enables users to rapidly review large, high-dimensional datasets (Figure 3). Importantly, the density-dependent down-sampling and agglomerative clustering employed by SPADE can prevent rare cellular phenotypes from being 'drowned out' by more highly represented cell types.[21]

One caveat of SPADE is that the user must specify the number of clusters to be found in the dataset, rather than have the number of clusters be driven by the data itself. In our experience, the user must specify that SPADE find a large number of clusters in order to ensure that rare cellular phenotypes are represented in the ensuing SPADE trees. This requirement causes SPADE to overcluster the data. We think, therefore, that SPADE needs to implement a formal methodology for determining when a single cluster cannot be further subdivided on the basis of the data being analyzed. This methodology should, at a minimum, take into consideration the empirically determined resolution limit of the detection platform, whether it be CyTOF or conventional fluorescence-based flow cytometry. If all differences between cells in a cluster fall below this resolution limit, then no further division into subclusters would be permitted. In addition, SPADE should enable groups of files to be compared using the same tree structure (such as comparing patients with healthy controls in which the tree structure is defined by data from the healthy controls). Currently, groups of files can only be compared if all data files are submitted to the program at one time, and no group-level statistical comparisons are available.

Protein and peptide microarrays

Microscope-slide-based linear antigen arrays were developed over a decade ago and have proven particularly useful for studying antibody responses to a large panel of different antigens in autoimmune, rheumatologic, and allergic diseases.[22] The initial methodology was simple and involved printing purified or recombinant peptides or proteins on glass microscope slides coated with materials such as poly-L lysine, epoxy, and nitrocellulose to enhance noncovalent binding of the printed target peptides to the slide surface.[22–25] Printing was, and still is, usually performed using contact printing and standard robotic microarrayers, but has evolved to include delivery using piezoelectric arrays, among other methods. Array content for the characterization of autoantigens has also progressed to include arrays of proteins, peptides, carbohydrates, and even lipids.[26–28]

Many groups still construct their own custom microarrays for individual diseases and applications. Investigators who lack the instrumentation or expertise to set up an array facility can purchase commercially available large-scale arrays containing over 10,000 recombinant proteins.[29,30] The majority of array methodologies employ fluorescence or chemiluminescence for detection; new technologies for detection include multiplexed surface plasmon resonance,[31] Raman spectral measurement,[32,33] and magnetic particles on giant magnetoresistive sensors.[34] If antigen array techniques are to alter the clinical practice of rheumatology, they will most likely do so in clinical laboratories or even at point-of-care using sophisticated sensors to read out the array data.

Rheumatology has several factors that make it particularly well-suited to the use of protein array technology. First, many rheumatic diseases are characterized by the presence of serum autoantibodies that predate development of clinical disease. These proteins are useful for diagnosis and prognosis, and, as some of them can be directly pathogenic, offer important clues for understanding disease pathogenesis.[35] Second, a large

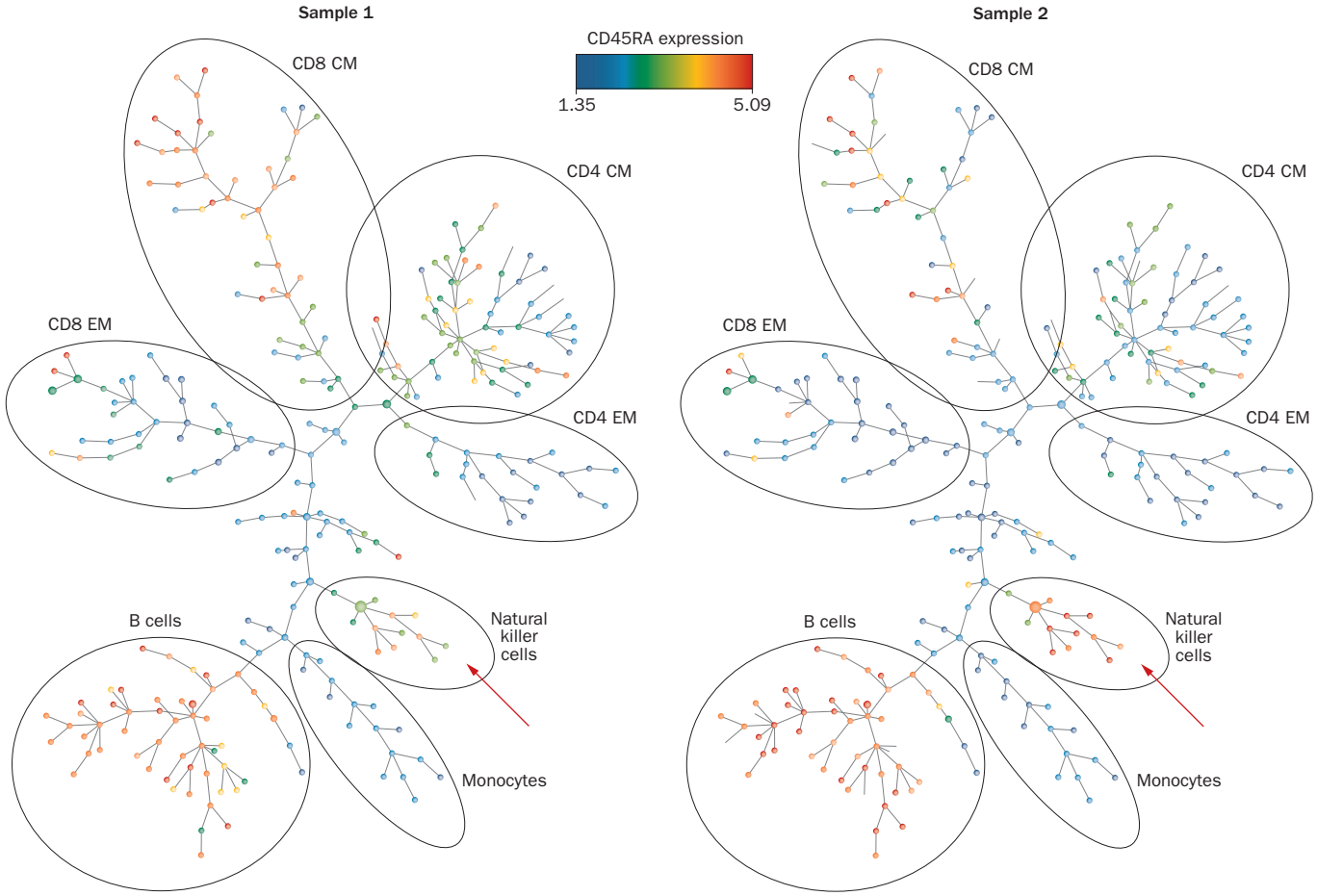

Figure 3 | Example of a SPADE representation of CyTOF data from analysis of peripheral blood mononuclear cells from two healthy individuals. The SPADE algorithm was used to perform unsupervised clustering of cells according to their expression of 23 cell surface markers. The algorithm then arranged the clusters into a consensus 'tree' structure, to show which clusters are most related to one another. Annotation of major cell lineages was added manually, based on the observed expression of known lineage markers in each 'branch' of the tree. Cluster size is proportional to cell number in the sample analyzed. Colouring shows relative CD45RA staining intensity in each cluster. Note the difference in CD45RA expression on the surface of natural killer cells in the two different individuals (arrows). SPADE is thus a powerful way to visualize differences between samples, without the bias introduced by traditional flow cytometry gating and enables a much more defined subset analysis of cells. Abbreviations: CM, central memory; CyTOF, cytometry by time-of-flight; EM, effector memory; SPADE, spanning-tree progression analysis of density-normalized events.

number of rheumatic and other inflammatory diseases are thought to be autoimmune in nature, yet the target antigen(s) have yet to be identified. Third, autoantibody identification might prove useful for development of antigen-specific therapies[36,37] or for selecting treatment modalities, such as belimumab or other biologic therapeutics, that are known to reduce levels of autoantibodies in treated patients.

New approaches in SLE

Systemic lupus erythematosus (SLE) is a model autoimmune disease that has been extensively studied using multiplex assays. SLE is characterized by multisystem organ involvement and the production of high-titre, highly specific autoantibodies directed against molecules found in the nucleus (anti-nuclear antibodies).[38] SLE is an extremely heterogeneous disease and, as such, is poorly understood, has few good biomarkers, and had no approved therapeutics until 2011. A striking finding in SLE and SLE-related diseases, including dermatomyositis, polymyositis, and systemic sclerosis, is that a majority of prominent autoantigens exist as particles containing one or more polypeptides that are associated with nucleic acids, such as RNA and DNA.[39] Antigen arrays, whether spotted onto microscope slides or developed as bead-based arrays, have been used to simultaneously measure antibodies directed against all of the particles, individual polypeptides from the particles, and even linear epitopes modelled on each polypeptide, for both SLE and SLE-related diseases.[22,40–43]

Peripheral blood mononuclear cells (PBMCs) from a large subset of patients with SLE contain what has been referred to as an interferon biosignature.[36,44] Several groups have demonstrated that mRNA transcript profiles from this SLE subset are highly similar to mRNA transcript profiles from PBMCs from healthy individuals that

are exposed, *in vitro*, to type I interferons (IFN-α and IFN-β).[44,45] This observation led to the hypothesis that defects in type I interferons and/or interferon-related signalling pathways could underlie the disease a large subset of patients who develop SLE, and could lead to therapies targeting this pathway.[46,47]

Multiplexed protein measurements have now been used to broadly characterize serum analytes; patients with SLE who possess the interferon biosignature were identified as part of the Autoimmune Biomarkers Collaborative Network,[44] to test the hypothesis that, just as interferon-inducible transcript profiles in PBMCs are strongly associated with SLE, interferon-inducible serum cytokine and chemokine expression can be found in blood from patients with SLE. Bauer *et al.*[48] used a method called rolling circle amplification to compare protein levels of a panel of 160 cytokines, chemokines, growth factors, and soluble receptors in patients with SLE with those in healthy controls.[48] The same analytes were also measured in supernatants prepared from PBMCs from healthy donors that had been stimulated for varying periods of time with IFN-α. Surprisingly, ~30 circulating factors were markedly upregulated in blood from patients with SLE, many of them interferon-inducible. This striking observation provided early biochemical evidence that the interferon biosignature was not just an epiphenomenon, but rather was directly linked to the biology of the underlying disease. Importantly, these findings seem to be clinically actionable, as measurement of just three of the chemokines (namely CCL2, CCL19 and CXCL10, performed using a high-throughput method chemiluminescent assay) accurately predicted disease activity and clinically meaningful disease flares over a 1-year period in a cohort of 267 patients with SLE.[49] In fact, measurement of these three chemokines proved to be superior to standard clinical rheumatology assays including those that measure C3, C4, double-stranded DNA, erythrocyte sedimentation rate, and C-reactive protein level.[49] Taken together, these results provide a rationale for multiplexed measurement of cytokines and chemokines in other autoimmune diseases, including RA, in which a subset of cytokines have been shown to be elevated and associated with aggressive disease,[50] and multiple sclerosis, in which a multiplexed bead-based assay demonstrated that IL-17F levels were elevated in patients with multiple sclerosis who failed to respond to IFN-β treatment.[51]

Are autoantibody profiles associated with the interferon signatures described above? The research group of one of the authors (P. J. Utz) has used arrays containing over 100 antigens to analyse the same serum samples used by Bauer *et. al.*,[48] and demonstrated a strong association with autoantibodies directed against particles associated with RNA and DNA; this association has now been replicated in two additional SLE cohorts (P. J. Utz, unpublished work). We hypothesize that immune complexes composed of these RNA-containing and DNA-containing antigens are internalized by B cells and dendritic cells, at which point the RNA and DNA moieties dissociate from the immune complexes and activate proinflammatory Toll-like receptors including TLR3, TLR7, TLR8 and TLR9.[52]

Autoantibody profiles have been used by other groups to study cohorts of patients with SLE, RA, and multiple sclerosis. Multiple ongoing studies by one of the authors (P. J. Utz) are focused on characterizing antibody profiles in patients who are exposed to investigational drugs, with the goal of identifying predictive biomarkers.[53] Although beyond the scope of this Review, antigen arrays have been extremely useful in studying mouse models of lupus, particularly mice lacking genes encoding interferon signalling molecules, retrogenic mice, and mice with altered MHC molecules.[37,54–57]

Clearly, multiplexed protein measurements will be crucial for elucidating pathogenic mechanisms in rheumatic diseases. Newer methods, such as high-throughput immunophenotyping using transcription (HIT) and Intel® (Intel Corporation, Santa Clara, CA, USA) peptide arrays synthesized using photolithography on the surface of silicon wafers, will enable more rapid and accurate measurement of serum analytes than ever before.[58,59]

High-throughput DNA sequencing

Immunoglobulin and TCR profiling

Prior to the development of 'next-generation' DNA sequencing instruments in the first decade of the 21st century, sequencing costs limited the characterization of B-cell receptor (BCR) and T-cell (TCR) populations. The experimental landscape has changed with the commercialization of several sequencing technologies that now make it possible to obtain thousands to millions of TCR or immunoglobulin sequences at a relatively low cost.[40,60–68] Currently, the major issues are: how best to prepare immune-receptor-sequence libraries, which sequencing technologies to use, how to analyze the data, and how to relate sequence data with functional activities of the immunoglobulin or TCR complexes.

One can break down the kinds of analysis enabled by high-throughput DNA sequencing of TCR or immunoglobulin rearrangements into three main categories. First, this method can be used to measure overall repertoire features, including: V, D and J segment usage frequencies (Figure 4); junctional properties, such as exonuclease digestion and non-templated base addition; the pattern of amino acid usage in the CDR3 region; evidence of receptor editing; heavy-chain isotype usage and hypermutation of rearranged gene segments (in the case of antibodies); and the number of distinct sequences present, which can be used to estimate repertoire diversity. Second, the receptors expressed by clonally expanded B cells or T cells (Figure 4) can be detected and characterized, whether or not one knows the antigen specificity or other functional features of the expanded clones. Third, B-cell or T-cell clones of interest that have previously been identified and correlated with known function can be tracked. Each of these kinds of analysis can yield insights into lymphocyte populations but the features of T-cell and B-cell repertoires that distinguish autoimmune disease patients from healthy individuals have not yet been fully explored.

Sequencing methodologies

The key variables in high-throughput DNA sequencing are read length, throughput, accuracy, and cost. Although this technology is rapidly developing, most published work on high-throughput sequencing of immunoglobulin and TCR to date has used either the 454 platform (Roche, Basel, Switzerland), owing its long read lengths (~450 bases) and moderate throughput (1 million reads per run), or the Illumina platform (Ilumina, San Diego, CA, USA) with its higher throughput (tens to hundreds of millions of reads per run) for comparable cost, but shorter read lengths (up to 150 bases from each end of a DNA molecule). The 454 instrument can capture a full immunoglobulin heavy chain V(D)J sequence in a single read, which is very helpful when studying patterns of hypermutation in clonally related IgH.[40,43,61,64,68–70] TCR sequences can be captured by shorter reads covering the V(D)J junction, and can take advantage of the Illumina platform throughput.[62,65,71]

The number of sequences that must be measured to provide meaningful data depends on the biological question being asked. Features of the immune repertoire such as segment usage, junctional nucleotides, hypermutation rates, and clonality can be analysed with thousands to tens of thousands of sequences. Deeper sequencing can detect progressively rarer populations. Typically, the detection of very rare sequences will only be meaningful if one has a prior reason for being interested in them, such as knowing the binding activity of these sequences, having previously observed clonally-related sequences in the same individual, or having seen similar sequences in other individuals. In addition, the finite rate of sequencing errors or PCR errors in a deep-sequencing experiment leads to the generation of artifactual sequence variants that can complicate estimation of the true diversity of an immunoglobulin or TCR library, particularly if the number of input B cells or T cells is not known, or if conservative filtering and replicate sample sequencing steps are not taken.[63,72]

For library preparation, multiplexed PCR reactions using large numbers of primers specific to the families of genes that encode the V and J segments have the advantage of relatively efficiently capturing sequences for amplification, but are difficult to optimize and usually confer amplification bias to some sequences. Heavily hypermutated immunoglobulin sequences are expected to be under-represented in all datasets owing to mutations in primer binding sites. The use of a variety of primer sets, including primers located in the relatively less-mutated leader regions of genes encoding the V segment, can alleviate this problem.[73] An alternative strategy requires using a protocol involving rapid amplification of complementary DNA ends (5' RACE), which does not rely on gene segment-specific primers. Our current knowledge of human variation in immunoglobulin and TCR germline loci is incomplete, and copy number variants (both deletions and amplifications), allelic variants, and other germline locus features might affect detection strategies.[74,75] Choice of template can also affect data interpretation, as genomic DNA is normalized to one copy of a V(D)J rearrangement per cell, and replicate libraries generated from genomic DNA aliquots give information about distinct cell populations. As mRNA is present in multiple copies, sequencing from cDNA actually limits the ability to distinguish between expanded clonal populations compared with high levels of mRNA expression by a single cell.

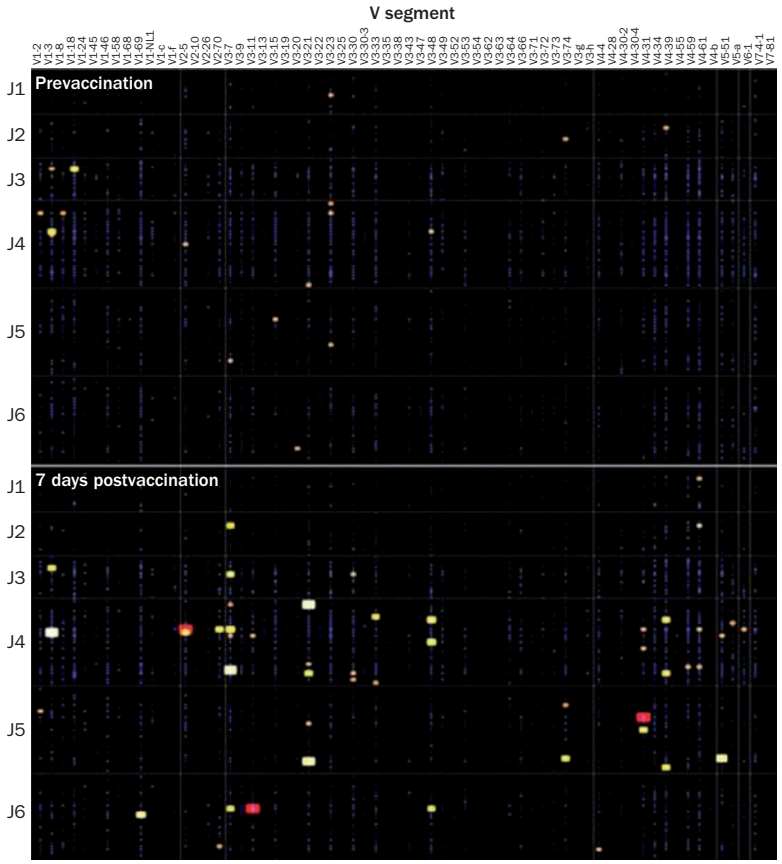

Figure 4 | The use of high-throughput DNA sequencing of immunoglobulin or T-cell receptor gene rearrangements to detect dynamic changes in lymphocyte repertoire and clonal expansions. In this example, the data show the response of a healthy individual to vaccination with a meningococcal polysaccharide vaccine, with the upper panel showing the peripheral blood B-cell repertoire prevaccination, and the lower panel showing the clonal B-cell response stimulated by the vaccine at day 7 postvaccination. Immunoglobulin heavy-chain V(D)J rearrangements were PCR-amplified from peripheral blood B cells from each sample, in sixfold replicate, using genomic DNA as the PCR template. Approximately 2,000–3,000 V(D)J rearrangements were sequenced from the libraries generated from each sample. If sequences with the same V, D, and J segments and junctions are detected in more than one replicate library from a sample, it provides evidence of a clonally expanded B-cell population. Expanded B-cell clones are displayed as squares of progressively larger size and warmer-spectrum (yellow, orange, red and white) colour. Clones detected in two replicates are shown by a small yellow square; clones detected in all six replicates are shown by a large white square. Small blue dots indicate VDJ combinations for which sequences were found in only a single replicate. The x-axis indicates the V segment used for a particular V(D)J rearrangement. The large y-axis rows show the J segment. The fine y-axis rows within each J segment row indicate the D segment. This method can be used to detect expanded clonal populations with a sensitivity limited mainly by the amount of sample available, and by the depth of sequencing carried out. Application of this approach to study the clonal populations of B cells and T cells in rheumatologic disorders should enable detailed tracking of lymphocyte populations that are correlated with disease activity and with therapeutic responses.

BCR and TCR rearrangements in autoimmunity

The initiating events of human autoimmune disorders are uncertain, and, despite clear evidence that adaptive immune responses have an important role in disease pathogenesis, it remains unknown whether T cells or B cells, or neither, are the site of primary dysregulation leading to immune-mediated damage of host tissues. Studies of the overall repertoire may shed light on abnormal selection processes for T cells and B cells in patients with autoimmune disease, as suggested by reports of alterations in the receptor repertoire following negative selection of self-reactive B cells, and impairment of selection checkpoints in patients with SLE.[42,76–78] DNA sequence-based understanding of the underlying immunoglobulin and TCR repertoires, and of the receptors expressed by expanded clonal B-cell and T-cell populations in patients, might offer important new information for classification and monitoring of these diseases.

Will it transpire that the overall repertoires of immunoglobulin or TCR gene rearrangements in patients with autoimmune diseases are pathognomonic in gene segment usage or detailed sequence features, or that they have any other distinguishing parameter when compared with the repertoires of healthy individuals? The answer is currently unknown. It is possible that public TCR or immunoglobulin rearrangements (that is, identical receptors used to respond to the same antigen in more than one person, despite the huge diversity of possible receptors) could be essential pathologic features of some autoimmune diseases. However, aberrant immune responses in different patients with the same diagnosis apparently target multiple self-antigens, different subsets of self-antigens, and multiple epitopes on those antigens, decreasing the likelihood that a particular immunoglobulin or TCR rearrangement will be a highly specific or sensitive disease marker. Indeed, phage display of human single-chain variable antibodies has shown that many distinct sequences can bind the same antigen; over 1,000 distinct immunoglobulin heavy-chain rearrangements result in molecules that bind human B-lymphocyte stimulator (BLyS, also known as TNF ligand superfamily member 13B), with little overall stereotyping of this repertoire.[79] Nevertheless, a high-throughput DNA sequencing study of monozygotic twins showed that an individual's germline genomic DNA sequence might be the strongest determinant of the usage of V, D and J segments in the immunoglobulin repertoire, providing a potential mechanism for some of the heritable predisposition to developing autoimmune disorders.[80] Other results have highlighted that extensive public rearrangements contribute to immunoglobulin light-chain repertoires.[81]

If autoimmune disease-specific public TCR or BCR signatures prove difficult to identify, tracking of clonally expanded (and presumably antigen-stimulated) B-cell or T-cell populations over the course of disease and treatment could act as a filter, to identify clones of cells that are likely to be involved in disease pathogenesis in a particular patient. Persistence of particular clones of B cells or T cells, and their correlation with disease activity, response to therapy, and likelihood of relapse, could guide immunosuppressive medication regimens. Studies of lupus nephritis demonstrated that the T cells in renal infiltrates are relatively oligoclonal, and that related clone members can also be detected in blood samples.[82–84] In one study, a clonal $CD8^+$ T-cell lineage found in blood and renal tissue samples from a patient with lupus nephritis was still detectable in a subsequent renal biopsy sample taken 6 years later, suggesting that persistent and long-lived clones are a relevant feature of this disease.[84] Further investigation of these topics will be greatly enhanced by the use of high-throughput DNA sequencing, by the more comprehensive measurement of TCR or immunoglobulin rearrangements present in a given blood or tissue sample, as well as by establishing age-adjusted normal-range measurements of the clonality of T cells and B cells in healthy individuals. Elderly individuals have high rates of oligoclonal and frequently cytomegalovirus-specific T-cell populations in the blood, particularly in the $CD8^+$ compartment.[85,86] Ensuring that such persistent clonal expansion of these T cells are interpreted with caution is an important factor in studies of autoimmunity.[85,86] Tracking of clonally related B cells and T cells in patient samples over time, particularly if functional data have been obtained to identify pathologically important cell lineages, might offer the best hope of monitoring disease in a patient-specific fashion. This approach might be challenging, given the imperfect correlation or lag between the presence of both B cells and T cells that express autoreactive sequences, or the detection of autoantibodies in the serum, and the development of disease signs or symptoms in the patient.[35,77]

In summary, high-throughput sequencing of immunoglobulin and TCR sequences offers a number of opportunities to expand our knowledge of human autoimmune biology. Global signatures might be present in some autoimmune diseases, but even in the absence of such signatures, tracking of B-cell and T-cell clones in individual patients could be used to monitor disease status and responses to therapy. We predict that the pairing of immunoglobulin or TCR sequencing with other experimental methods (such as selection of antigen-specific cells, or sorting of phenotypic lymphocyte populations of interest) should be particularly powerful for evaluating disease phenotypes.

Heterogeneity of samples

In many cases, the biological samples analyzed by technologies such as microarrays are heterogeneous; that is, they are composed of multiple different cell types, each with its own gene and protein expression signatures. The frequency of different cell types might vary markedly between specimens, as it does, for example, in peripheral blood samples (2–10-fold differences in frequency among various cell types).[87] In the case of gene expression microarrays, for example, the tissue sample is lysed to isolate the mRNA, which is then analysed by microarray. Traditional microarray analysis methods

do not take into account any information on cell-type heterogeneity in the sample and so cannot distinguish between variations in gene expression attributable to an actual physiological change in a cell type and those attributable to differences in actual cell-type frequency. Moreover, the contributions of the different cell types to the total measured gene expression cannot be identified.[88–90] Therefore, the ability of these methods to detect differentially expressed genes is strongly affected by variation in the frequencies of different cell types in the sample;[88,89,91] moreover, the interpretation of results is made difficult as transcripts are described as part of a single system, without cellular boundaries or context (Figure 5A, right). Techniques to circumvent this issue by isolating specific cell types and profiling each type separately affect the underlying biology to a varying extent and make a strong underlying assumption on the cell type of interest. As a result, the perspective of the overall system is missing; that is, any information about non-profiled cell types is unknown and the effects of cell-to-cell interaction are lost (Figure 5A, left).

A methodological innovation is to use statistical deconvolution techniques to achieve a middle ground between cell-type specific and system-wide information levels (Figure 5A, middle; Figure 5B). This approach exploits the fact that the majority of genes are expressed to a varying degree in multiple cell types. By tracking how gene expression fluctuates between samples in relation to cell-frequency changes, the average gene expression of each cell type within an analyzed group of samples, as well as the cell-type specific expression differences between groups, can be accurately estimated *in silico*.[92–94] The sensitivity of cell-type specific expression analysis performed in this manner is often orders of magnitude higher than that obtained by analyzing heterogeneous tissue samples, yet is likely to be lower than that achieved by isolating the individual cell types. Moreover, as the deconvolution methodology does not require any cell separation, the cell type responsible for any detected differences in expression can be identified whilst avoiding the requirement to isolate the cell type of interest. In contrast to traditional techniques, increased variation in cell frequencies between samples actually improves the performance of statistical deconvolution in accurately estimating cell-type specific expression and group differences.[94] Groups of specific cell types have been shown to be reliably detected for cells whose frequency in the sample is as low as 5–10%, though the minimal cell-type frequency for which detection of group differences is possible can only be determined empirically owing to the large number of factors involved.[94] Notably, statistical deconvolution-based techniques are not restricted to microarray gene expression but may be easily adapted to a large number of other assays (including deep sequencing, intracellular flow cytometry, mass cytometry, and protein arrays, as well as bead-based profiling) in which the biological samples analyzed are heterogeneous with respect to cell type.

As in all analyses performed in humans, a large amount of variability exists between samples, which is attributable to genetic differences, environmental factors, medical conditions and medication taken. A balanced experimental design between study groups to control for major factors (such as gender, age, BMI and so on)is recommended, yet accounting for all factors within the study is nearly impossible. We therefore recommend a combined solution comprising: a careful and detailed documentation of as many confounding variables as possible; rigorous statistical testing to

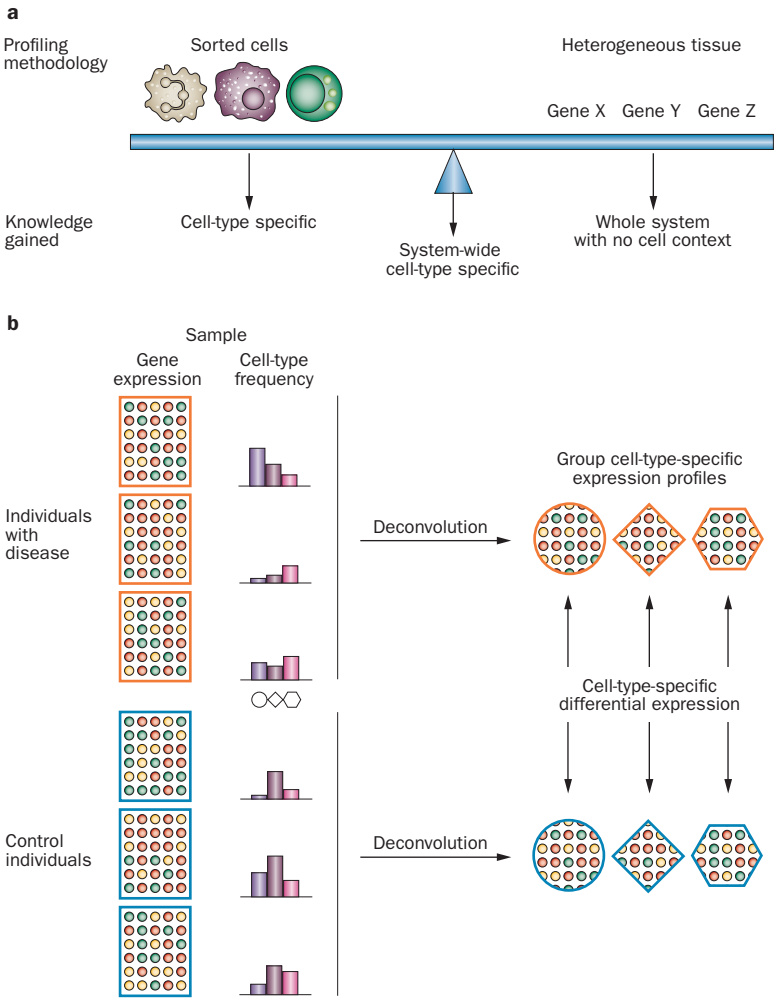

Figure 5 | Statistical deconvolution enables detection of system-wide cell-type specific differences between groups without cell-type isolation. **a** | The majority of biological samples comprise multiple cell-types that can vary dramatically in frequency from one sample to another. Traditional sample profiling, either by isolating specific cell-types of interest or by profiling heterogeneous tissues, provide a system-level understanding or cellular context respectively. Statistical deconvolution-based techniques offer a middle ground by providing system-wide cell-type specific differences between groups. **b** | The csSAM methodology provides a high-resolution and sensitive differential expression analysis that is localized to a specific cellular context. Quantifying the frequency of the different cell-type subsets in each sample enables the average gene expression profile of each cell type in each group to be estimated by statistical deconvolution. These estimated expression profiles can then be utilized to detect cell-type specific differences without sorting of the heterogeneous tissue, and reconstitute whole tissue as individual samples that are independent of frequency variations associated with cell type. Abbreviation: csSAM, cell type–specific significance analysis of microarrays. Permission obtained for part b from Nature Publishing Group © Shen-Orr, S. *et al*. *Nat. Methods* **7**, 287–289 (2010).

measure the effects of the confounder variables at the start of the analyses, and the introduction of the major variables into the statistical model, sample size allowing, as per the classical statistical literature; post-discovery retesting of the relationship between findings and confounder variables; and follow-up experiments aimed at testing detected relationships between main findings and confounder variables.

Conclusions

In this Review, we have discussed new technologies that will be used in future immune phenotyping analyses: mass cytometry, peptide and protein arrays, and BCR and TCR sequencing. These novel assays offer the promise of new information to improve the management of autoimmune disease and represent the latest methodology for analyzing cells, soluble proteins, and genes, respectively. New technologies for the analysis of gene expression in whole blood samples and for deconvolution of the resultant datasets enable the expression of specific genes to be assigned to cell subsets, without isolation and manipulation of the blood cells; in this way they offer a much improved method of looking for actionable biomarkers. From such highly multiplexed analytical approaches, panels of actionable biomarkers will undoubtedly be extracted that will be useful for diagnosis, prognosis, clinical subtyping, and selection and monitoring of therapy. Given the complexity of the immune system and the high degree of crosstalk between cells, biomarkers would be expected to be not only of a single measure, but also of relationships between measures. It may be too early to tell which of these new methods will prove most practical and useful, but we strongly believe that future clinical decisions may be guided, in part, by biomarkers that can only be defined at as high dimensional. Hence, we advocate for increased training in quantitative methods.

Review criteria

Articles were identified by performing a search of the PubMed database between December 2011 and January 2012 using search terms that included: "FACS", "CyTOF", "gene expression", "high throughput TCR and immunoglobulin sequencing", "protein arrays", "peptide arrays", "autoimmunity", "rheumatic disease", "rheumatology", "proteomics" and "biomarkers". This Review is limited to full-text articles published in English. Because many of these methodologies are evolving quickly, references were limited to those published between the years 2000 and 2012. Personal libraries and references from identified papers were also used to write this manuscript. Preference was given to techniques that represented a marked improvement over existing methodologies, and have been made readily accessible to other groups, and have been used in published research of considerable impact to the field of clinical immunology.

1. Capdeville, R., Buchdunger, E., Zimmermann, J. & Matter, A. Glivec (STI571, imatinib), a rationally developed, targeted anticancer drug. *Nat. Rev. Drug Discov.* **1**, 493–502 (2002).
2. Nahta, R. & Esteva, F. J. HER-2-targeted therapy: lessons learned and future directions. *Clin. Cancer Res.* **9**, 5078–5084 (2003).
3. LaGasse, J. M. *et al.* Successful prospective prediction of type 1 diabetes in schoolchildren through multiple defined autoantibodies: an 8-year follow-up of the Washington State Diabetes Prediction Study. *Diabetes Care* **25**, 505–511 (2002).
4. van der Woude, D. *et al.* The ACPA isotype profile reflects long-term radiographic progression in rheumatoid arthritis. *Ann. Rheum. Dis.* **69**, 1110–1116 (2010).
5. Zethelius, B. *et al.* Use of multiple biomarkers to improve the prediction of death from cardiovascular causes. *N. Engl. J. Med.* **358**, 2107–2116 (2008).
6. Maclaren, N. *et al.* Only multiple autoantibodies to islet cells (ICA), insulin, GAD65, IA-2 and IA-2β predict immune-mediated (Type 1) diabetes in relatives. *J. Autoimmun.* **12**, 279–287 (1999).
7. Goekoop-Ruiterman, Y. P. *et al.* Clinical and radiographic outcomes of four different treatment strategies in patients with early rheumatoid arthritis (the BeSt study): a randomized, controlled trial. *Arthritis Rheum.* **52**, 3381–3390 (2005).
8. Moreland, L. W. *et al.* Treatment of rheumatoid arthritis with a recombinant human tumor necrosis factor receptor (p75)-Fc fusion protein. *N. Engl. J. Med.* **337**, 141–147 (1997).
9. Li, Q. Z. *et al.* Protein array autoantibody profiles for insights into systemic lupus erythematosus and incomplete lupus syndromes. *Clin. Exp. Immunol.* **147**, 60–70 (2007).
10. Sokolove, J. *et al.* Autoantibody Epitope Spreading in the Pre-Clinical Phase Predicts Progression to Rheumatoid Arthritis. *PLoS ONE* (in press).
11. Hueber, W. *et al.* Blood autoantibody and cytokine profiles predict response to anti-tumor necrosis factor therapy in rheumatoid arthritis. *Arthritis Res. Ther.* **11**, R76 (2009).
12. Querec, T. D. *et al.* Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. *Nat. Immunol.* **10**, 116–125 (2009).
13. Irish, J. M. *et al.* Single cell profiling of potentiated phospho-protein networks in cancer cells. *Cell* **118**, 217–228 (2004).
14. Bandura, D. R. *et al.* Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Anal. Chem.* **81**, 6813–6822 (2009).
15. Ornatsky, O. *et al.* Highly multiparametric analysis by mass cytometry. *J. Immunol. Methods* **361**, 1–20 (2010).
16. Bendall, S. C. *et al.* Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332**, 687–696 (2011).
17. Edwards, B. S., Oprea, T., Prossnitz, E. R. & Sklar, L. A. Flow cytometry for high-throughput, high-content screening. *Curr. Opin. Chem. Biol.* **8**, 392–398 (2004).
18. Pyne, S. *et al.* Automated high-dimensional flow cytometric data analysis. *Proc. Natl Acad. Sci. USA* **106**, 8519–8524 (2009).
19. Walther, G. *et al.* Automatic clustering of flow cytometry data with density-based merging. *Adv. Bioinformatics* 686759 (2009).
20. Qiu, P. *et al.* Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat. Biotechnol.* **29**, 886–891 (2011).
21. Stanford University Gary P. Nolan Laboratory. CytoSPADE: cytoscape-driven spanning tree progression of density-normalised events [online], http://cytospade.org/ (2011).
22. Robinson, W. H. *et al.* Autoantigen microarrays for multiplex characterization of autoantibody responses. *Nat. Med.* **8**, 295–301 (2002).
23. Bussow, K. *et al.* A method for global protein expression and antibody screening on high-density filters of an arrayed cDNA library. *Nucleic Acids Res.* **26**, 5007–5008 (1998).
24. Ekins, R. P. Multi-analyte immunoassay. *J. Pharm. Biomed. Anal.* **7**, 155–168 (1989).
25. Joos, T. O. *et al.* A microarray enzyme-linked immunosorbent assay for autoimmune diagnostics. *Electrophoresis* **21**, 2641–2650 (2000).
26. Kanter, J. L. *et al.* Lipid microarrays identify key mediators of autoimmune brain inflammation. *Nat. Med.* **12**, 138–143 (2006).
27. Wang, D., Liu, S., Trummer, B. J., Deng, C. & Wang, A. Carbohydrate microarrays for the recognition of cross-reactive molecular markers of microbes and host cells. *Nat. Biotechnol.* **20**, 275–281 (2002).
28. Quintana, F. J. *et al.* Antigen microarrays identify unique serum autoantibody signatures in clinical and pathologic subtypes of multiple sclerosis. *Proc. Natl Acad. Sci. USA* **105**, 18889–18894 (2008).
29. Michaud, G. A. *et al.* Analyzing antibody specificity with whole proteome microarrays. *Nat. Biotechnol.* **21**, 1509–1512 (2003).
30. Zhu, H. *et al.* Global analysis of protein activities using proteome chips. *Science* **293**, 2101–2105 (2001).

31. Tabakman, S. M. *et al.* Plasmonic substrates for multiplexed protein microarrays with femtomolar sensitivity and broad dynamic range. *Nat. Commun.* **2**, 466 (2011).
32. Chen, R. J. *et al.* Noncovalent functionalization of carbon nanotubes for highly specific electronic biosensors. *Proc. Natl Acad. Sci. USA* **100**, 4984–4989 (2003).
33. Chen, Z. *et al.* Protein microarrays with carbon nanotubes as multicolor Raman labels. *Nat. Biotechnol.* **26**, 1285–1292 (2008).
34. Gaster, R. S. *et al.* Quantification of protein interactions and solution transport using high-density GMR sensor arrays. *Nat. Nanotechnol.* **6**, 314–320 (2011).
35. Arbuckle, M. R. *et al.* Development of autoantibodies before the clinical onset of systemic lupus erythematosus. *N. Engl. J. Med.* **349**, 1526–1533 (2003).
36. Garren, H. *et al.* Phase 2 trial of a DNA vaccine encoding myelin basic protein for multiple sclerosis. *Ann. Neurol.* **63**, 611–620 (2008).
37. Robinson, W. H. *et al.* Protein microarrays guide tolerizing DNA vaccine treatment of autoimmune encephalomyelitis. *Nat. Biotechnol.* **21**, 1033–1039 (2003).
38. Tan, E. M. *et al.* The 1982 revised criteria for the classification of systemic lupus erythematosus. *Arthritis Rheum.* **25**, 1271–1277 (1982).
39. von Muhlen, C. A. & Tan, E. M. Autoantibodies in the diagnosis of systemic rheumatic diseases. *Semin. Arthritis Rheum.* **24**, 323–358 (1995).
40. Campbell, P. J. *et al.* Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc. Natl Acad. Sci. USA* **105**, 13081–13086 (2008).
41. Liu, F., Whitton, J. L. & Slifka, M. K. The rapidity with which virus-specific CD8+ T cells initiate IFN-gamma synthesis increases markedly over the course of infection and correlates with immunodominance. *J. Immunol.* **173**, 456–462 (2004).
42. Meffre, E. *et al.* Immunoglobulin heavy chain expression shapes the B cell receptor repertoire in human B cell development. *J. Clin. Invest.* **108**, 879–886 (2001).
43. Weinstein, J. A., Jiang, N., White, R. A. 3rd, Fisher, D. S. & Quake, S. R. High-throughput sequencing of the zebrafish antibody repertoire. *Science* **324**, 807–810 (2009).
44. Baechler, E. C. *et al.* Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. *Proc. Natl Acad. Sci. USA* **100**, 2610–2615 (2003).
45. Crow, M. K., Kirou, K. A. & Wohlgemuth, J. Microarray analysis of interferon-regulated genes in SLE. *Autoimmunity* **36**, 481–490 (2003).
46. Lau, C. M. *et al.* RNA-associated autoantigens activate B cells by combined B cell antigen receptor/Toll-like receptor 7 engagement. *J. Exp. Med.* **202**, 1171–1177 (2005).
47. Rifkin, I. R., Leadbetter, E. A., Busconi, L., Viglianti, G. & Marshak-Rothstein, A. Toll-like receptors, endogenous ligands, and systemic autoimmune disease. *Immunol. Rev.* **204**, 27–42 (2005).
48. Bauer, J. W. *et al.* Elevated serum levels of interferon-regulated chemokines are biomarkers for active human systemic lupus erythematosus. *PLoS Med.* **3**, e491 (2006).
49. Bauer, J. W. *et al.* Interferon-regulated chemokines as biomarkers of systemic lupus erythematosus disease activity: a validation study. *Arthritis Rheum.* **60**, 3098–3107 (2009).
50. Hueber, W. *et al.* Proteomic analysis of secreted proteins in early rheumatoid arthritis: anti-citrulline autoreactivity is associated with up regulation of proinflammatory cytokines. *Ann. Rheum. Dis.* **66**, 712–719 (2007).
51. Axtell, R. C. *et al.* T helper type 1 and 17 cells determine efficacy of interferon-beta in multiple sclerosis and experimental encephalomyelitis. *Nat. Med.* **16**, 406–412 (2010).
52. Green, N. M. & Marshak-Rothstein, A. Toll-like receptor driven B cell activation in the induction of systemic autoimmunity. *Semin. Immunol.* **23**, 106–112 (2011).
53. Sharp, V. & Utz, P. J. Technology insight: can autoantibody profiling improve clinical practice? *Nat. Clin. Pract. Rheumatol.* **3**, 96–103 (2007).
54. Holst, J. *et al.* Scalable signaling mediated by T cell antigen receptor-CD3 ITAMs ensures effective negative selection and prevents autoimmunity. *Nat. Immunol.* **9**, 658–666 (2008)
55. Richez, C. *et al.* IFN regulatory factor 5 is required for disease development in the FcγRIIB–/–Yaa and FcγRIIB–/– mouse models of systemic lupus erythematosus. *J. Immunol.* **184**, 796–806 (2010).
56. Thibault, D. L. *et al.* IRF9 and STAT1 are required for IgG autoantibody production and B cell expression of TLR7 in mice. *J. Clin. Invest.* **118**, 1417–1426 (2008).
57. Thibault, D. L. *et al.* Type I interferon receptor controls B-cell expression of nucleic acid-sensing Toll-like receptors and autoantibody production in a murine model of lupus. *Arthritis Res. Ther.* **11**, R112 (2009).
58. Kattah, M. G., Coller, J., Cheung, R. K., Oshidary, N. & Utz, P. J. HIT: a versatile proteomics platform for multianalyte phenotyping of cytokines, intracellular proteins and surface molecules. *Nat. Med.* **14**, 1284–1289 (2008).
59. Price, J. "On silico" peptide microarrays for high resolution mapping of antibody epitopes and diverse protein-protein interactions. *Nat. Med.* (in press).
60. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
61. Boyd, S. D. *et al.* Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Sci. Transl. Med.* **1**, 12ra23 (2009).
62. Freeman, J. D., Warren, R. L., Webb, J. R., Nelson, B. H. & Holt, R. A. Profiling the T-cell receptor β-chain repertoire by massively parallel sequencing. *Genome Res.* **19**, 1817–1824 (2009).
63. Glanville, J. *et al.* Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc. Natl Acad. Sci. USA* **106**, 20216–20221 (2009).
64. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
65. Robins, H. S. *et al.* Comprehensive assessment of T-cell receptor β-chain diversity in αβ T cells. *Blood* **114**, 4099–4107 (2009).
66. Venturi, V. *et al.* A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing. *J. Immunol.* **186**, 4285–4294 (2011).
67. Wang, C. *et al.* High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. *Proc. Natl Acad. Sci. USA* **107**, 1518–1523 (2010).
68. Wu, Y. C. *et al.* High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood* **116**, 1070–1078 (2010).
69. Liao, H. X. *et al.* Initial antibodies binding to HIV-1 gp41 in acutely infected subjects are polyreactive and highly mutated. *J. Exp. Med.* **208**, 2237–2249 (2011).
70. Wu, X. *et al.* Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science* **333**, 1593–1602 (2011).
71. Robins, H. *et al.* Ultra-sensitive detection of rare T cell clones. *J. Immunol. Methods* **375**, 14–19 (2011).
72. Warren, R. L. *et al.* Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Res.* **21**, 790–797 (2011).
73. Lebecque, S. G. & Gearhart, P. J. Boundaries of somatic mutation in rearranged immunoglobulin genes: 5' boundary is near the promoter, and 3' boundary is approximately 1 kb from V(D)J gene. *J. Exp. Med.* **172**, 1717–1727 (1990).
74. Boyd, S. D. *et al.* Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *J. Immunol.* **184**, 6986–6992 (2010).
75. Wang, Y. *et al.* Genomic screening by 454 pyrosequencing identifies a new human *IGHV* gene and sixteen other new *IGHV* allelic variants. *Immunogenetics* **63**, 259–265 (2011).
76. Wardemann, H. *et al.* Predominant autoantibody production by early human B cell precursors. *Science* **301**, 1374–1377 (2003).
77. Yurasov, S. *et al.* Persistent expression of autoantibodies in SLE patients in remission. *J. Exp. Med.* **203**, 2255–2261 (2006).
78. Yurasov, S. *et al.* Defective B cell tolerance checkpoints in systemic lupus erythematosus. *J. Exp. Med.* **201**, 703–711 (2005).
79. Edwards, B. M. *et al.* The remarkable flexibility of the human antibody repertoire; isolation of over one thousand different antibodies to a single protein, BLyS. *J. Mol. Biol.* **334**, 103–118 (2003).
80. Glanville, J. *et al.* Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. *Proc. Natl Acad. Sci. USA* **108**, 20066–20071 (2011).
81. Jackson, K. J. *et al.* Divergent human populations show extensive shared IGK rearrangements in peripheral blood B cells. *Immunogenetics* **64**, 3–14 (2011).
82. Massengill, S. F., Goodenow, M. M. & Sleasman, J. W. SLE nephritis is associated with an oligoclonal expansion of intrarenal T cells. *Am. J. Kidney Dis.* **31**, 418–426 (1998).
83. Murata, H. *et al.* T cell receptor repertoire of T cells in the kidneys of patients with lupus nephritis. *Arthritis Rheum.* **46**, 2141–2147 (2002).
84. Winchester, R. *et al.* Immunologic characteristics of intrarenal T cells: Trafficking of expanded CD8 T cell β-chain clonotypes in progressive lupus nephritis. *Arthritis Rheum.* http://dx.doi.org/10.1002/art.33488.
85. Gillespie, G. M. *et al.* Functional heterogeneity and high frequencies of cytomegalovirus-specific CD8(+) T lymphocytes in healthy seropositive donors. *J. Virol.* **74**, 8140–8150 (2000).
86. Hadrup, S. R. *et al.* Longitudinal studies of clonally expanded CD8 T cells reveal a repertoire shrinkage predicting mortality

and an increased number of dysfunctional cytomegalovirus-specific T cells in the very elderly. *J. Immunol.* **176**, 2645–2653 (2006).
87. Janeway, C. et *al.* Immunobiology: The Immune System in Health and Disease (5th edition) (Garland Science, New York, 2001).
88. Cobb, J. P. *et al.* Application of genome-wide expression analysis to human health and disease. *Proc. Natl Acad. Sci. USA* **102**, 4801–4806 (2005).
89. Whitney, A. R. *et al.* Individuality and variation in gene expression patterns in human blood. *Proc. Natl Acad. Sci. USA* **100**, 1896–1901 (2003).
90. Xu, Q. *et al.* Investigation of variation in gene expression profiling of human blood by extended principle component analysis. *PLoS ONE* **6**, e26905 (2011).
91. Palmer, C., Diehn, M., Alizadeh, A. A. & Brown, P. O. Cell-type specific gene expression profiles of leukocytes in human peripheral blood. *BMC Genomics* **7**, 115 (2006).
92. Clarke, J., Seo, P. & Clarke, B. Statistical expression deconvolution from mixed tissue samples. *Bioinformatics* **26**, 1043–1049 (2010).
93. Kuroda, M. J. *et al.* Human immunodeficiency virus type 1 envelope epitope-specific CD4(+) T lymphocytes in simian/human immunodeficiency virus-infected and vaccinated rhesus monkeys detected using a peptide-major histocompatibility complex class II tetramer. *J. Virol.* **74**, 8751–8756 (2000).
94. Shen-Orr, S. S. *et al.* Cell type-specific gene expression differences in complex tissues. *Nat. Methods* **7**, 287–289 (2010).
95. van Lochem, E. G. *et al*. Immunophenotypic differentiation patterns of normal hematopoiesis in human bone marrow: reference patterns for age-related changes and disease-induced shifts. *Cytometry B. Clin. Cytom*. **60**, 1–13 (2004).

Acknowledgements
The authors wish to thank Dr Hongwu Du for his advice on the systems immunology section of the article and for assistance with editing the manuscript.

Author contributions
All authors researched data for the article, substantially contributed to the discussion of content and selection of references and reviewed/edited the manuscript before submission. H. T. Maecker, T. M. Lindstrom, W. H. Robinson, P. J. Utz, M. Hale, S. D. Boyd and S. S. Shen-Orr wrote the article.