

Comprehensive phenotyping and biological marker discovery

Aaron B. Kantor

SurroMed, Inc., 2375 Garcia Ave, Mountain View, CA 94043, USA

Tel.: +1 650 230 1561; Fax: +1 650 230 1960; E-mail: akantor@surromed.com

Abstract. There is an enormous unmet need for biological markers to characterize disease type, status, progression, and response to therapy. We are developing and applying an integrated bioanalytical platform and clinical research program to facilitate comprehensive differential phenotyping of patient samples and enable the discovery of biomarkers. The platform employs high-throughput, quantitative analysis for the characterization of thousands of parameters including cell populations, cell-surface antigen density, soluble proteins and soluble low molecular weight biomolecules, from small-volume biological samples in a clinical research laboratory-like setting.

Keywords: Biomarker, cytometry, proteomics, mass spectrometry, clinical studies

1. Comprehensive phenotyping and biomarkers

We are applying a broad-based integrated technology platform for comprehensive phenotyping to biological marker discovery. Phenotype is the constellation of observable traits and measurable constituents of an organism. It includes all biochemical and physiological characteristics and is the result of genotype plus development in a specific environment. Phenotype is much more than characteristics observable with the naked eye. X-rays and magnetic imaging provide information on the phenotype of the skeleton and internal organs. Cytometry provides information on the composition and nature of cells in blood and other tissues. Immunoassays and mass spectrometry can quantify the amount of proteins and low molecular weight biomolecules in blood, tissues and various fluids.

The phenotype of an individual changes constantly throughout life as a result of complex interactions between an individual's genotype and environmental factors. Environmental factors such as exposure to infectious agents, diet, stress, exposure to chemicals and the physiological changes associated with aging affect phenotype at the molecular and macroscopic levels. Different environments can influence the development of inherited traits. Height and weight, for example, are

affected by the available food supply and nature of the diet. The average height of young adults in Japan today is half a foot taller than their grandparents because of diet, not heredity. Environmental differences can alter expression by similar genotypes such as twins maturing in dissimilar families. Disease often results in a profound change in the phenotype of the individual.

Biological markers are measurable phenotypic parameters that are useful for characterizing the state of health or disease of an individual. They can be an indicator of normal biological processes, disease processes or pharmacologic responses to therapeutic intervention. A subset of biomarkers that can accurately substitute for a clinical endpoint becomes accepted as a surrogate endpoint over time and is used as a diagnostic and prognostic measure in the clinic. Biomarkers vary in molecular complexity and molecular mass. They range from simple low molecular weight molecules, such as sugars, fatty acids peptides and steroids, to soluble and cell surface proteins, and to complex clinical phenotypes (Fig. 1). Some biological markers are well established as clinical diagnostics. In diabetes fasting blood glucose is a short-term indicator and glycated hemoglobin (HbA1c) an intermediate-term indicator of glycemic control. Similarly, CD4 T cell counts are a

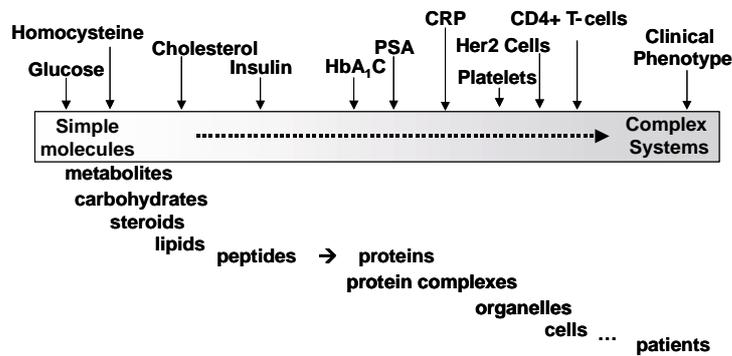


Fig. 1. Biomarkers enable objective evaluation of disease and drug response. Biomarkers can range from simple molecules to complex systems.

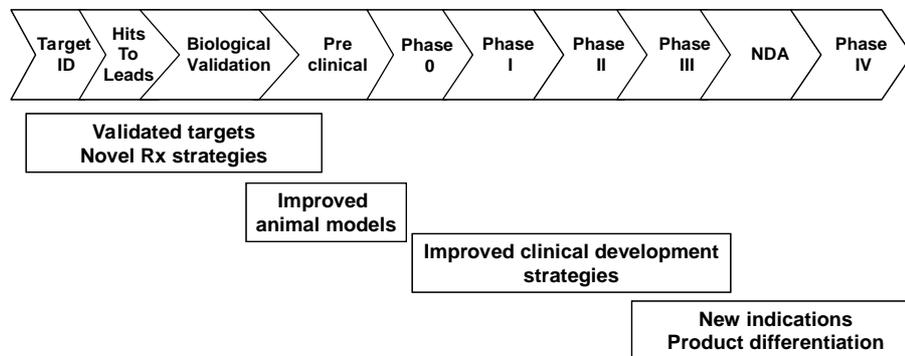


Fig. 2. Biomarker-enabled drug discovery, development and commercialization. Biomarkers can have utility at all stages of the drug development process.

commonly used indicator of immune system status in acquired immunodeficiency disease.

Biomarkers can change in response to a particular therapeutic intervention and have large utility in drug development, ranging from discovery to pre-clinical evaluation, to all aspects of clinical trials and commercialization (Fig. 2). They can yield clinical drug development strategies that are more effective, less costly and executed more rapidly because they require fewer patients. Clinical endpoints can be quite subjective and long trials may be required before efficacy of such endpoints can be assessed. The success with currently available biomarkers presages the enormous value and need for biomarker discovery. Cholesterol and lipoprotein ratios are well-established biomarkers of coronary artery disease and substitute for clinical endpoints, major coronary events and cardiac mortality, in the clinical trials of HMG-CoA reductase inhibitors (statins), providing effective readouts in months rather than years. Biomarkers can also be used to select patients who are likely responders before enrollment in a clinical trial. A good example is the use of HER2 screening prior to enrollment in the Herceptin[®] (trastuzumab) Phase

III clinical trial. HER2 is over expressed in about 25 percent of breast cancers and the level of expression increases with the aggressiveness of the tumor. By limiting the clinical trial to patients who over express HER2 in an immunohistochemistry assay, Slamon et al were able to demonstrate efficacy using less than 500 patients with metastatic disease [5]. This and other tests for HER2 continue to be used in additional clinical trials and as a commercial diagnostic/prognostic test prior to treatment.

There is a strong demand for additional biomarkers. We are developing and applying proprietary strategies and technologies to comprehensive phenotyping in order discover new biological markers of disease progression and response to therapy. The approach is highly integrated and combines capture of patient clinical signs and symptoms from physical examination and systems testing with broad bioanalytical profiling including measurements of blood, urine and other biological samples. An informatics infrastructure, consisting of a proprietary database and software tools for data acquisition, analysis and mining has been developed to handle the diverse types of measurements.

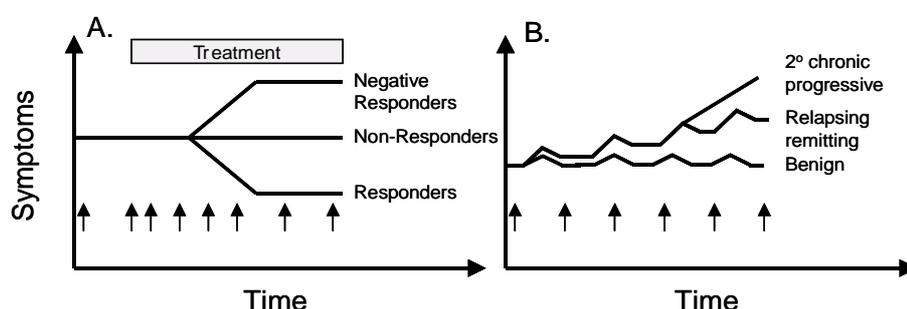


Fig. 3. Biomarker studies connected to clinical outcome. A. Response to treatment. B. General disease progression, independent of treatment. Arrows represent sample collection.

2. Bioanalytical platform and clinical studies

The bioanalytical platform enables high-throughput, quantitative analysis and characterization of thousands of parameters including cell populations, cell-surface antigen density, soluble proteins and soluble low molecular weight biomolecules, from small-volume biological samples in a clinical research laboratory. We emphasize readily available samples such as blood and urine that provide the best potential for future diagnostics. Our current bioanalytical platform can be divided based on the use of antibody reagents. Immunophenotyping includes cytometry and immunoassays to measure specific known analytes. Mass spectrometry-based analysis of proteins and low molecular weight molecules is used to measure both known and unknown analytes. It is well suited for discovering unexpected changes during the course of disease progression or therapeutic intervention. It is also ideal for identifying post translational modifications that cannot be identified by gene expression technologies. The informatics infrastructure accommodates all SurroMed data streams as well as those generated by additional technologies, including gene expression and genotypic data, which may be incorporated in a given study.

Ideally, biomarker studies are tied to clinical outcome. Two general examples are shown in Fig. 3. In the first example, biomarkers are needed to predict the outcome of a particular treatment either before treatment begins or shortly thereafter. From a medical viewpoint, the need is greatest when the time to identify a response with an established clinical endpoint is long and the treatment is expensive, such as for many new biologic drugs. An example would be the response to anti-TNF-alpha therapy in rheumatoid arthritis [2]. In rheumatoid arthritis therapeutic studies, clinical assessment typically includes tender and swollen joint counts, C-reactive protein, duration of morning stiffness, over-

all physician assessment and for longer times, X-ray evaluation of the joints as described by the American College of Rheumatology (ACR). About one-third of the patients are good responders, showing a 70% improvement based on ACR scores, and about one third are non-responders. In the second example disease progression is considered independent from treatment. The example would be appropriate for different forms of multiple sclerosis. Early markers are needed to distinguish which patients will have relapsing remitting disease from those that develop secondary progressive disease.

3. Sample size and initial biomarker discovery

How many samples should be used for an initial biomarker discovery study? Concern is often raised about the number of samples needed to detect statistically significant differences, without false positive or false negative conclusions, as the number of variables is increased. A calculation demonstrates that the number of samples needed for these types of studies can be quite practical. The broad-based technology permits the collection of many more bioanalytical variables than subjects. Consequently, many multivariate statistics cannot be used and the data must be analyzed using univariate test statistics. Figure 4 shows a power calculation as the number of variables is increased from 1 to 5000. The plot calculation uses paired measurements of one group (e.g. before and after drug) and a power of 90%. Here we take a very conservative approach to protect against false positive conclusions. To maintain the overall experiment-wise false positive rate at 0.05, the p-values from the univariate test statistics has been adjusted using the step-down Bonferroni method of Holm. The utility of any variable is determined by the effect size, the difference of the means

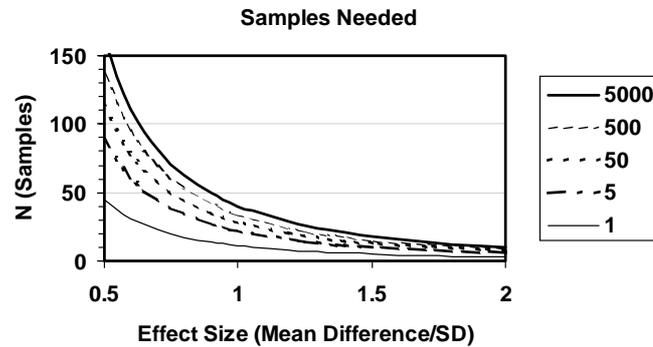


Fig. 4. Samples needed vs. effect size and number of variables. Calculation is based on paired comparisons, a power of 90% and an overall study-wise p-value of 0.05 using the step down Bonferroni correction of Holm.

of the measure, relative to the standard deviation of the measure (MD/SD). Often discovery studies are initiated in hypothesis generation mode, without an *a priori* estimate of the effect size for most measures. It is reasonable to power a first study to detect effect sizes of around 1. The number of samples required for a given effect size increases logarithmically with the number of variables. For 50 measures, this would require about 28 samples for each of the two time points. For 5000 measures the number of samples increases to about 40 samples. The need for additional samples per additional variable is far less than one-to-one. Note that the number would approximately double for unpaired comparisons. These calculations are encouraging because even using a conservative analysis technique adequate power for a typical study is maintained without requiring a prohibitive number of subjects.

4. Cytometry

The SurroScanTM microvolume laser scanning cytometry (MLSC) system is used to classify and quantify hundreds of immune and inflammatory cell populations and cell surface antigen expression in small volumes of whole blood, processed blood or other fluids [1,6]. It is designed for robust daily use in a clinical research laboratory. The system combines instrumentation, reagents, disposable plastic capillary arrays and software to enable the rapid, sensitive, automated and comprehensive cellular phenotyping. Advantages over conventional flow cytometry systems include improved ease-of-use, higher degree-of-multiplexing, decreased sample volumes, absolute quantitation of cell counts, automated data capture and analysis and consistent results among instruments.

In MLSC, as with flow cytometry, fluorophore-labeled antibodies specific for cell surface antigens are used to identify, characterize, and enumerate specific leukocyte populations. The staining reaction can be done in whole blood or other single cell suspensions. In general, there is no need to wash the reagent away; quantitative dilution of the blood-antibody mixture is usually sufficient sample preparation. The cell-antibody mixture is loaded into an optical-quality capillary of known volume and analyzed with a laser-based fluorescence-imaging instrument. In order to operate with whole blood, we use fluorophores that can be excited in the red region (> 600 nm) of the spectrum, such as Cy5, Cy5.5 and Cy7-APC. White blood cells isolated following ficoll or red cell lysis, can also be analyzed.

In contrast to flow cytometry, the laser scans over stationary cells rather than cells flowing past the laser. A small cylindrical laser spot is scanned across the capillary in one direction while the capillary is translated relative to the optical system in a second direction. An example of the resulting raster-scan profile for a single fluorescence channel is shown in Fig. 5(A). Each peak represents a cell and can be converted to a list mode data file and displayed as a dot plot (Fig. 5(B)). Typically three antibody reagents, each with a different fluorescent tag and each detected in a different channel are used per assay. Three cell populations are indicated in the figure: CD4 T cells, CD8 T cells and monocytes.

Our clinical studies typically utilize panels of 64 three-color cellular assays, arranged in two disposable capillary arrays, that allow the identification and enumeration of hundreds of different cell types and cell-associated molecules that are relevant to immune, inflammatory and metabolic processes. Each reagent cocktail typically contains one or two antibodies to the major cell populations – neutrophils, eosinophils,

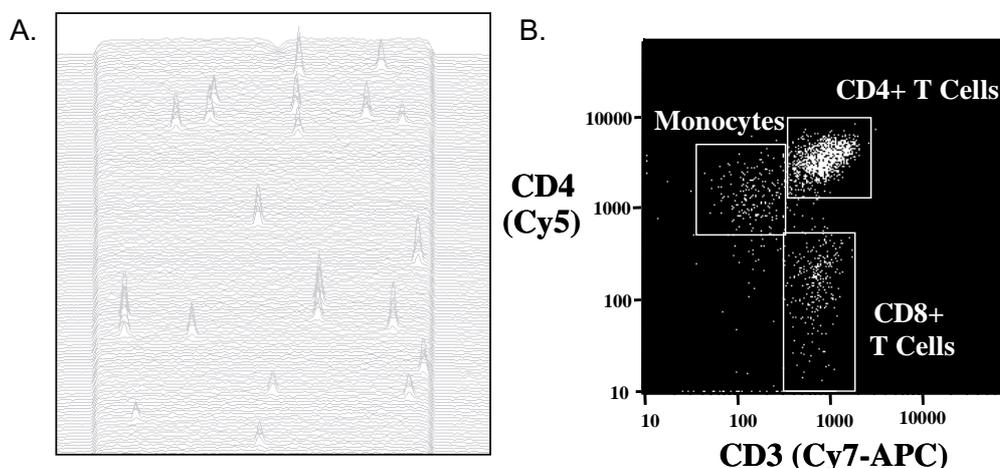


Fig. 5. Cytometry data. Whole blood was labeled with Cy5-anti-CD4, Cy5.5-anti-CD8 and Cy7-APC anti-CD3 antibodies, diluted and scanned. A. Raster scan generated for one fluorescence channel. Peaks correspond to antibody-labeled cells identified with image processing software that produces a list-mode data file with parameters for every detected cell event [4]. Unlabeled cells (erythrocytes and other leukocytes) are not identified. The baseline pedestal is unreacted reagent that was not washed away. B. Dot plot. Resultant data can be graphed with each dot representing one cell. Intensity data is compensated for spectral overlap, so the shown parameters are proportional to the amount of dye-antibody reagent on each cell.

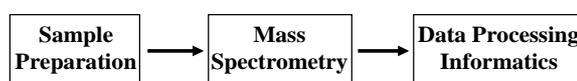
monocytes T cells, B cells, NK cells, and platelets – and one or two antibodies to subsetting antigens which may indicate the functional state, activation state or adhesion characteristics of the population.

The SurroScan cytometer is designed for robust use in a clinical research laboratory. Proprietary informatics tools enable automated collection, processing and analysis of the data. This data pipelining approach speeds data flow and reduces user error. For each clinical study, information on the individual assays, including reagents, cell populations, and gating strategy are stored in a common Oracle database. The system is designed such that every cytometer yields the same results for a given sample. Consequently, common gates, the boxes used to identify specific cell populations, can generally be used throughout the study. Resultant information on cell counts and relative antigen levels on individual cell types is immediately available for review and use in study wide analyses.

5. Profiling protein and low molecular weight molecules in serum samples

Mass spectrometry is a mature bioanalytical platform for the measurement of specific analytes with a high degree of mass accuracy. The challenge is to adapt this technology to robust screening of large numbers of biological samples. There are three main technical components for differential profiling of proteins and low

molecular weight molecules from biological samples:



mass spectrometry instrumentation, sample preparation, and data processing informatics. We acquire high-end instruments and integrate these with proprietary sample preparation methods and data processing informatics. Our approach has been to use high resolution, high mass accuracy time-of-flight mass spectrometers for screening purposes. Screening with the time-of-flight instruments provides an advantage in the ability to track low abundance species in complex mixtures like serum. With a 0.002 Dalton mass accuracy, common compounds are readily identified from complex mixtures by comparison to pre-established proprietary data libraries. Additionally, follow-up tandem mass spectrometry, with the ion trap and triple quadrupole instruments, is used for the identification and focuses on candidate biomarkers that are significantly different between study groups.

Biological samples, like unfractionated serum, are very complex and often dominated by a few components present at very high concentration that can interfere with the detection of all other molecules by MS. For example albumin (35–45 mg/ml) and immunoglobulin (IgG, IgA and IgM, 25–35 mg/ml) make up about 90% of total serum protein. Simplification of the com-

plex mixture enables quantitation of a broader set of analytes.

Our proteomics and metabolomics sample preparation process for serum focuses on simplifying the material: First, serum is fractionated into low and high molecular weight fractions because they require different analytical methods at the preparation, instrumentation and identification steps for subsequent resolution into hundreds of distinct molecular species. The low molecular weight fraction is analyzed by two techniques: gas chromatography-mass spectrometry and on-line HPLC-mass spectrometry. Chemical modification before the gas chromatography enhances the volatility of the compounds and the ion impact ionization mass spectrometry enables quantitation and a high degree of reproducibility. In the second technique, electrospray ionization mass spectrometry is used because it can be readily coupled to on-line liquid chromatography. We are among the first to establish reproducibility and quantification for the electrospray ionization technique. Average CV's for analytes in complex mixtures are about 25%, which establishes a reasonable level for looking at biological variation. The types of compounds in the low molecular weight fraction are diverse. Identification is made with the assistance of the AMDIS computer program and spectra library of ~ 100,000 compounds from National Institute of Standards and Technology. Tandem mass spectrometry is also used for identification.

The high molecular weight fraction is processed further. High abundance proteins such as albumin and immunoglobulin that dominate the fraction are removed by affinity methods to increase dynamic range. The high molecular weight fraction is analyzed in several ways. First tryptic digestion is done to generate peptide fragments and the sample is applied to an on-line reverse phase liquid chromatography column coupled to mass spectrometry with electrospray ionization and time-of-flight detection. This method has a duty cycle of a few hours and can be automated for routine use in a clinical research laboratory. Second, the albumin and Ig fractions can be analyzed independently for associated molecules. Third, two dimensional chromatography can be used to extend the range of molecules that can be profiled. This approach requires a longer duty cycle and may be more appropriate for a subset of samples. The high molecular weight fraction is largely proteins and tandem mass spectrometry, protein databases and commercial software are used to identify peaks of interest and build a compound library. Post translational modifications such as phosphorylation, N and

O-linked glycosylation and ubiquitination can also be identified.

An example of the richness of the chromatogram and mass spectra associated with such a human serum sample is shown in Fig. 6. Approximately 1000 molecular species can be readily resolved from the high molecular weight fraction after removal of albumin and immunoglobulin even with a very fast chromatography run. Only two hours is needed for the process. Each protein gives rise to several peptide species that can be identified with tandem mass spectrometry. We have identified hundreds of different proteins in serum and a partial list is shown in the figure.

SurroMed has developed proprietary informatics methods for comparison of subject samples and quantification of proteins, peptides or small molecules, in which these components are quantified relative to the same, corresponding molecules in a different samples, usually a control or normal sample. The approach relies on the assumption that biological samples consist of complex mixtures of multiple biological components, of which only a minority is relevant to the comparison. The majority of components are relatively constant for the same individual over time or across subject populations. The majority of components whose concentrations do not vary across samples are used as an intrinsic internal standard to normalize the concentrations of components that do vary. Spectra from individual samples first undergo nonlinear filtering to remove noise, dynamic thresholding to separate peaks from noise and vectorized two-dimensional peak peaking to take advantage of information in both the chromatography and mass-to-charge dimensions [3]. Second, common components in the samples are compared to enable normalization and time warping to correct for small differences in the runs. Third, peak lists from all of the samples are compared and a merged peak list is developed and applied for monitoring a common integrated set of peaks across all samples. This list of peaks and intensities resides in the study-wide database for statistics and data mining.

6. Conclusions

The large amount of patient data across a broad range of bioanalytical and clinical measurements necessitates an array of informatics tools. We have developed tools for the primary analysis of raw data on each bioanalytical platform, pipelining that data into an Oracle database and visualizing and analyzing the results on

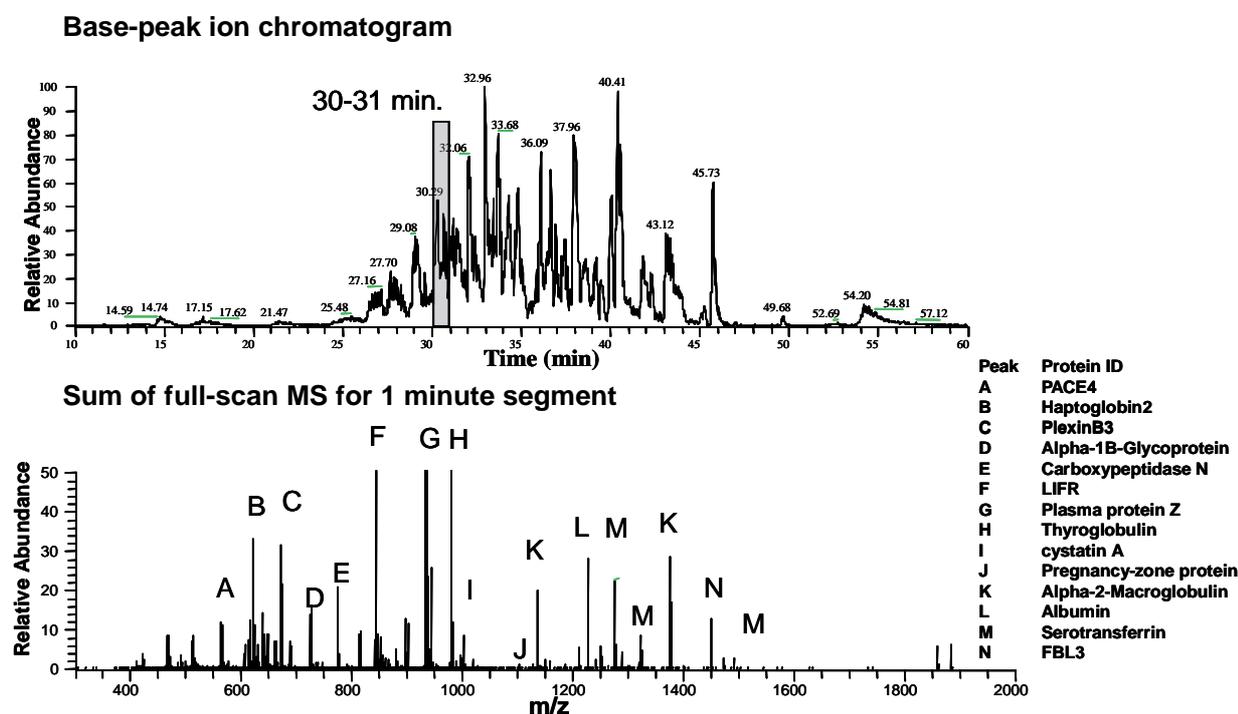


Fig. 6. HPLC-MS from Human Serum. The HMW fraction was digested with trypsin after removal of the majority of albumin and immunoglobulin. Twenty micrograms of the digest was run in a high throughput mode. Top: Base peak chromatogram shows the elution profile from the HPLC column. Bottom: Sum of the full scan MS for a 1 minute segment. Tandem MS and database searches can be used to identify hundreds of peptides. A representative set is listed on the right.

a study-wide basis. Web-based tools allow biologists, biochemists and clinicians to query the database and view statistical results on a study-wide basis. This aids in the identification of putative biomarkers and follow-up. Taken together, the concepts and technologies described here, as well as those being developed by others, will enable a revolution in comprehensive phenotyping and lead to improved treatment of patients in multiple disease areas.

Acknowledgements

This paper is an overview of the SurroMed biomarker discovery effort. A diverse group of biologists, chemists, engineers, informatics specialists and clinicians have contributed to the development of the platform. Michael Natan made an oral presentation of some of this material at the Biomarkers and Surrogate Endpoints Conference. Special thanks go to Karen Cheal for the power calculation for Fig. 4 and Tom Shaler for Fig. 6. Thanks to Howard Schulman, Susan Alters, John Kollins and Chris Becker for thoughtful comments on the manuscript.

References

- [1] L.J. Dietz, R.S. Dubrow, B.S. Manian and N.L. Sizto, Volumetric capillary cytometry: a new method for absolute cell enumeration, *Cytometry* **23** (1996), 177–186.
- [2] M. Feldmann and R.N. Maini, Anti-TNF alpha therapy of rheumatoid arthritis: what have we learned? *Annu. Rev Immunol* **19** (2001), 163–196.
- [3] C.A. Hastings, S.M. Norton and S. Roy, New algorithms for processing and peak detection in liquid chromatography/mass spectrometry data, *Rapid Commun. Mass Spectrom.* **16** (2002), 462–467.
- [4] S.M. Norton, J. Winkler and L.J. Dietz, Cell enumeration and characterization in microvolume laser scanning cytometry: a multicolor image processing package, *Optical Diagnosis of Living Cells III, Proceedings of SPIE* **3921** (2000), 20–30.
- [5] D.J. Slamon, B. Leyland-Jones, S. Shak, H. Fuchs, V. Paton, A. Bajamonde, T. Fleming, W. Eiermann, J. Wolter, M. Pegram, J. Baselga and L. Norton, Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2, *N. Eng. J Med.* **1344** (2001), 783–792.
- [6] I.D. Walton, L.J. Dietz, G. Frenzel, J. Chen, J. Winkler, S.M. Norton and A.B. Kantor, Microvolume laser scanning cytometry platform for biological marker discovery, *Proc. SPIE-Int. Soc. Opt. Eng.* **3926** (2000), 192–201.



Hindawi
Submit your manuscripts at
<http://www.hindawi.com>

