Biomarkers for Infectious Disease Diagnostics in the Developing World:

Status of Technologies and Approaches for Biomarker Discovery

Andrew Blasband, Paul Neuwald, Laura Penny, Katherine Tynan,

Mickey Urdea and Richard Thayer

May 2006

Halteres Associates, LLC

5858 Horton Street, Suite 550

Emeryville, CA 94608

510-420-6733

www.halteresassociates.com

Table of Contents

List of Tables and Figures

I. Introduction

Despite the fact that many of the infectious diseases that are prevalent in the developing world are treatable, they continue to pose enormous health burdens in many populations. A wide variety of factors contribute to this situation, with perhaps the most obvious being limited access to effective vaccines and drug treatments for many of the world's poorer individuals. Another factor that is less well known, but still vitally important to the success of any disease management program, is the lack of a practical way to identify the patients who require treatment. Diagnosing patients in need of a particular treatment, and efficiently distinguishing them from individuals who have non-specific symptoms, is a major barrier to disease management programs in the developing world. The burden of infectious diseases could be significantly reduced if diagnostic tests with the appropriate performance characteristics were accessible to the large populations that need them.

Laboratory-based tests with useful sensitivities and specificities already exist for most of the diseases that affect the developing world, although these tests are generally not available in the extremely resource-limited testing sites, such as the peripheral health outposts, or the poorly-funded urban public health centers, which serve the majority of the population.

Leaving aside economic and other socio-political factors, there are a number of factors that limit the dissemination of many advanced tests to resource-limited testing sites, and other factors that adversely affect their performance characteristics in such sites. These factors are often related to 1) the biological limitations of the biomarker(s) that has been selected to predict disease, 2) the concentration of the biomarker(s) in the specimen types that are practical to collect in resource-limited settings, and 3) the technology platform (or the complexity of its supply chain) that is currently available to detect the biomarker. In some instances, there are no known biomarkers that would provide adequate performance in the specimen types that are practical to use in resource-limited settings, especially in high-prevalence populations, with life-long exposures to many infectious organisms, and/or who are infected with more than one organism.

A test's performance is measured by its sensitivity and specificity, which will vary significantly depending on the circumstances under which the test is performed (technical issues) and the prevalence of the disease in the population being tested. The resource-limited settings considered in this report usually do not have access to power, clean water, sterile conditions, temperature-controlled supply chain or storage, or even basic equipment, nor do they have trained personnel to perform venipuncture or even what are considered low complexity assays in resource-rich sites. For these reasons, many currently available tests for biomarkers of infectious diseases that were designed for a laboratory setting do not perform adequately, or cannot be attempted at all, in these resource-limited settings.

Other biomarkers might be simpler to detect, but the assays to detect them require more than a day to complete. Because of the challenges that patients face in reaching some of these sites, and the chance that they will be lost to follow-up, tests that take more than several hours to perform become impractical because they do not allow a clinical decision to be made in a meaningful time frame.

There are other biomarkers for infectious diseases that were selected based on their predictive power in low prevalence, first world populations. When used in high-prevalence populations in the developing world, these biomarkers are often not clinically helpful because their predictive power is dramatically different in the populations being served by resource-limited testing sites.

This report surveys the current status of scientific knowledge regarding known biomarkers for five major disease categories that currently pose significant disease burdens in the developing world: Acute Lower Respiratory Infections (ALRI), HIV and other STDs (Syphilis, Chlamydia and Gonorrhea), TB, Malaria and Diarrheal Diseases. The issues and current diagnostic options for specific infectious disease intervention points will be discussed. For each intervention point (clinical decision) that is relevant to disease management in resource-limited settings, the availability and performance characteristics of known biomarkers are reviewed. The areas of need, barriers to entry, and the tools required for the validation of biomarkers in the relevant populations are discussed. In addition, scientific strategies for the discovery of new, potentially better biomarkers are outlined. Finally, recommendations are provided regarding the most promising approaches, including specific biomarkers (or biomarker types), specimen types, and technology platforms for delivering a test with useful sensitivities and specificities in resource-limited settings for each clinical decision point within the next 1 to 3 years.

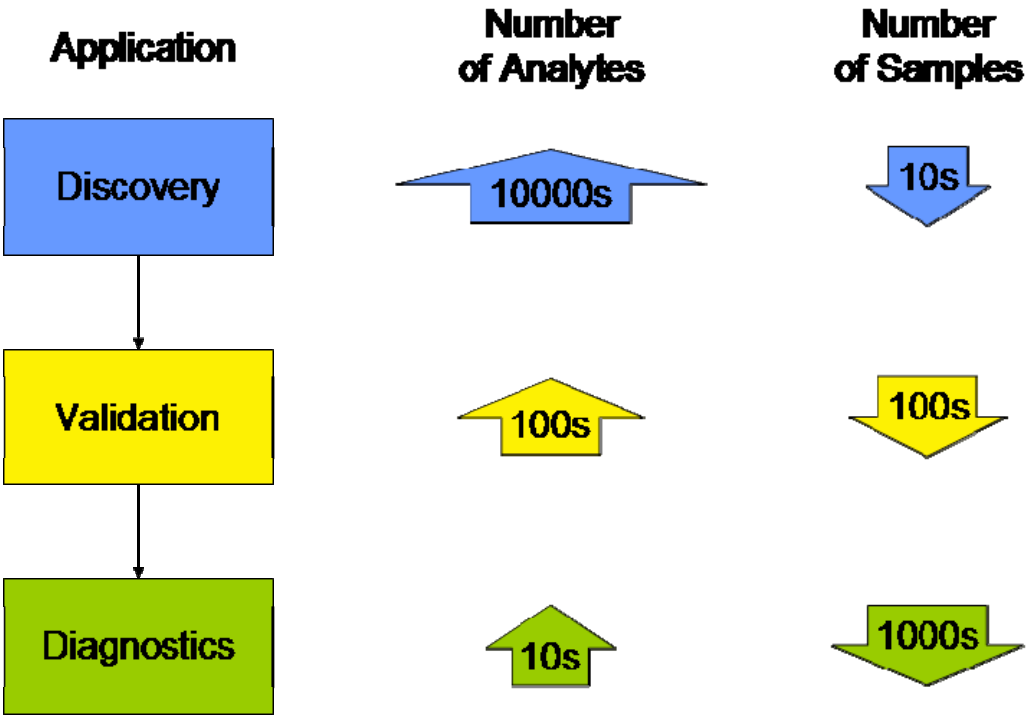
II. Evaluation of Technologies for the Discovery of Biomarkers

A biomarker, or biological marker, is defined as a characteristic that is measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacological responses to a therapeutic intervention.[1, 2] The application of biomarkers in medicine is not new and has been used effectively in the diagnosis and management of cardiovascular diseases, infections, cancer, immunological and genetic disorders.[3] The key issue is identifying an appropriate biomarker for a disease. A good biomarker has a clearly defined relationship to the disease and can be used to predict clinical benefit. However, single biomarkers are sometimes unreliable and often lack sensitivity, such as in the case of prostate specific antigen (PSA), which is used as a biomarker for prostate cancer. Over the past several decades, developments in genomic, proteomic, transcriptomic, and metabolomic technologies such as DNA sequencing, mass spectrometry, nuclear magnetic resonance, protein arrays, and DNA microarrays have enabled researchers to profile potential biomarker molecules in normal and diseased tissues to discover multiple biomarkers or biomarker signatures that are indicative of disease, drug response and safety. These multiple biomarkers are often more sensitive and selective than the use of single biomarkers in diagnostic tests. While it is one thing to identify new biomarkers for a disease, developing a diagnostic assay that can measure a biomarker in a clinical sample presents another set of challenges, particularly if the preferred sample type is an easy-to-access body fluid.

While conducting research for this evaluation of practical and empirical technologies used in biomarker discovery for proteins, RNA, and metabolites, several themes arose that are common in all three fields. The most startling revelation is that biomarkers are being discovered utilizing technology that is for the most part between 20 and 50 years old. The mature, established, reliable, and trustworthy technologies like mass spectrometry, two dimensional gel electrophoresis, liquid chromatography, and nuclear magnetic resonance are the stalwarts of protein and metabolite biomarker discovery efforts. When it

comes to RNA, it is the microarray that is almost exclusively used for discovering RNA biomarkers. Another common theme is that the technologies used to discover biomarkers are not generally being used as platforms for commercially-marketed diagnostic tests, though some might also be used for validating biomarkers. For example, in experimental programs aimed at discovering biomarkers, the number of samples is typically small, but the number of molecules being analyzed in each sample is large. In validation, the objective is to narrow the range of candidate biomarkers, yet enlarge the number of samples that can be processed in a meaningful time frame. When a diagnostic assay is used in commercial clinical testing, the goal is to have a small number of biomarkers that confer both sensitivity and specificity for a disease, that are measured using a platform that can easily process thousands of samples (see Figure 1).

Figure 1. Objectives of Biomarker Analysis by Application Mode



This figure illustrates the relationship between the number of analytes that need to be evaluated and the number of samples that need to be analyzed for three applications of biomarker analysis (discovery, validation, and commercialization) In the discovery process, thousands of analytes are interrogated in a small number of samples. The candidate biomarkers identified in discovery are then validated by screening hundreds of samples to narrow the candidates down to tens of analytes. In platforms used for commercialized diagnostic products, this small number of validated biomarkers will be assayed in thousands of samples.

One cautionary note needs to be mentioned regarding the issue of increasing the chances of false positives when surveying a large number of analytes in parallel as is commonly done during the discovery process. A common problem arises in –omics experiments where a large number of probes are surveyed simultaneously in the hopes of finding something of "significance". Typically, multiple "significance

tests" are applied to the same data set in the quest to uncover a "hidden message" in the massive amount of data generated during the analyses. Unfortunately, this can lead one to applying "significance" to a result that normally should be labeled insignificant. In order to avoid such an occurrence, statistical methods called "multiple testing procedures" are performed on the data to help cull out false positives. The Bonferroni Method of adjustment and the Benjamini-Hochberg step-up procedure[4] are just two examples of procedures that help minimize false positives and, for that matter, false positives due to application of statistical significance test applied to large data sets.[5]

It also appears that the day of the single biomarker diagnostic test is moving by the wayside as the ability to display profiles of biomarkers becomes more plausible with today's technologies. Multiple biomarkers are being discovered for many diseases that will lead to the development of diagnostic test panels that show more specificity and sensitivity than assays based on single biomarkers.

In the discovery process, enormous volumes of data are generated that must be compared with existing data sets from other studies. To date, there are few if any standards for collecting, analyzing, storing, and exchanging data within, or between, laboratories. The principal bottleneck identified as impeding the discovery of protein and metabolic biomarkers by most thought leaders in these fields is the lack of standards for both bioinformatics and analysis. Methods that are used for handling large proteomic datasets are even less standardized. This leads to problems when attempting to share information between researchers. Although RNA data standardization remains a popular topic, to date, the methods are debated and difficult to compare. Most laboratories that are engaged in biomarker analytical research use multiple bioinformatic and statistical approaches, hoping that at least one pans out.

Finally, efforts to discover biomarkers for diseases of particular importance to developing nations is fragmented, decentralized, uncoordinated, and unfocused. This problem is only exacerbated by the lack of data analysis and bioinformatic standards that make it difficult to assess if the discovery of a biomarker from one laboratory is truly novel or whether it has merely been rediscovered. This type of redundancy in efforts between disconnected groups with common goals seems to be the norm and not the exception. Reducing duplicate efforts and coordinating groups with similar interests focused on a single goal will lead to more efficient use of time and resources in the discovery of novel biomarkers for infectious diseases. We will take these issues into consideration in the recommendations section.

1. Technologies for Discovering Protein Biomarkers

Technologies used in protein biomarker discovery efforts fall into three general categories, separation methods, detection instrumentation, and bioinformatics tools. The repertoire of technology employed in today's biomarker discovery laboratories is surprisingly small. The technologies are relatively mature, reliable, and established corner stones in protein research laboratories around the world. Most protein biomarkers today have been discovered via technologies that utilize a combination of the following methods: hardware and software tools; separation of proteins by gel electrophoresis or chromatography; identification or detection by mass spectrometry (MS), nuclear magnetic resonance (NMR), or protein arrays; and extensive bioinformatics algorithms for protein/peptide pattern recognition and identification. The overall trends today in protein discovery tools are to increase the throughput, resolution and

sensitivity of established technologies currently in use, rather than on the development of novel biomarker discovery technologies.

One key point to note in protein biomarker discovery to date is that while there have been many potentially useful biomarkers proposed in the literature, none have been validated to meet Food and Drug Administration (FDA) requirements.[6] Is this to say we should abandon the current technological approaches being used today and invent new discovery tools for biomarker discovery? Many thought leaders think not. The problem more likely lies in the fact that biomarker discovery at this point in time is a needle-in-the-haystack expedition where no one is exactly sure what the haystack or the needle looks like. However, the analytical advances in current technologies made thus far in areas such as greater acquisition times, sensitivity, dynamic range and resolution, coupled with better bioinformatics tools that help identify protein biomarkers from massive amounts of data, have accelerated a greater and more comprehensive understanding of the proteome content of many biofluids today. It is only now that we are beginning to understand the complexities of the proteomes that are common to many disease states while at the same time refining what the "haystack" looks like.

Protein Separation Technologies: Two-Dimensional Gel Electrophoresis

Two-dimensional gel electrophoresis (2DGE) is the only technique currently available with sufficient resolving power and range of application. Consequently, it is the most widely used protein separation technique in the discovery of protein biomarkers. 2DGE is over 30 years old and is routinely performed manually in many laboratories. There have been few significant improvements in the history of 2DGE, and most researchers agree that innovation in the method is necessary and well overdue. The instrumentation can range from a single electrophoresis gel apparatus to more sophisticated units that can run up to 12 gels at one time. Despite the popularity of 2DGE, the technique is renowned for being complex and time-consuming. The techniques involved can be difficult to learn and require the user to gain a high degree of skill. Furthermore, the hands-on nature of traditional 2DGE methods often results in irreproducible data. 2DGE will always be complemented rather than replaced by any new enabling technologies. Drug discovery companies have found 2DGE to be an unparalleled tool for identifying therapeutic targets and thus it plays a crucial role in the drug target discovery process.[7]

In 2DGE, the sample is supported by a polyacrylamide gel through which proteins are separated first in one direction, and then in a perpendicular direction. For the first separation, known as isoelectric focusing, a pH gradient is applied to the gel with the top being more acidic than the bottom. Proteins are loaded at the point where the pH is neutral, and a voltage is applied. The proteins migrate through the gel to the pH at which they have no net charge (their isoelectric point). For the second separation step, the gel is soaked in a denaturing solution (to unfold protein structures) containing a negatively charged detergent such as sodium-dodecyl-sulphate (SDS). The proteins become coated in SDS so they are all equally negatively charged. A voltage is applied across the gel, and the proteins migrate at a rate dependent primarily on their molecular mass, with smaller proteins moving faster than larger ones.[7] There are two electrophoresis approaches in 2DGE, namely non equilibrium pH gradient electrophoresis (NEPHGE) which offers 10 times higher resolution than the second form, immobilized pH gradient (IPG). NEPHGE can separate up to 10,000 proteins whereas only 1000 to 2000 proteins can be separated by IPG. IPG however, has resulted in increased reproducibility and allows very accurate quantitative

comparative protein mapping.[3] Only about one percent of today's users run their samples on NEPHGE while the majority of researchers prefer IPG because it has been commercialized in a way that makes it easier for new users to implement it by companies such as GE Healthcare and Bio-Rad.

The major advantages of 2DGE are that it has very high resolution (separation based on pI and size/mass) while also being relatively quantitative. Consequently, 2DGE is routinely used in protein/peptide profiling and for expression level comparisons. Some of the major drawbacks of the technology are that it has limited capability to analyze very small (< 2 kilo Daltons [kDa]), very large (>130 kDa), hydrophobic or low abundance proteins, requires large sample volumes, has low throughput with long turn-around times of approximately two and a half to three days, and remains a manual and complex technique requiring technical expertise and a lot of 'art'. Furthermore, it is often very difficult to share separation (protein spot) analysis data obtain from the same sample between instruments, let alone between laboratories. The potential for errors is great and is only compounded by the variability of users, reagents, and instrumentation.

Trends in Electrophoretic Protein Separation: Automation of Electrophoresis

The trend in 2D gel electrophoresis can be summarized in one word: automation. Obviously, with such a long tenure as the premier protein/peptide separation technique there must be compelling reasons to put up with all the inadequacies of 2D gel electrophoresis. Several commercial organizations have attempted to address the most pressing issues related to the manual components of the process.

Next Gen Sciences (UK) has developed an automated 2D electrophoresis (a2DE) platform.[8] The a2DE has been designed to save time, reduce costs, and improve results. The a2DE will deliver gels within 24 hrs, with only about one hour of hands-on time during the system setup.

Shimadzu Biotech (Japan), has developed an automated 2D gel isolation technology "Xcise", that runs a gel, images a gel, excises bands, digests peptides, extracts and then deposits the sample on a matrix assisted laser desorption ionization (MALDI) mass spectrometry plate or on an liquid chromatography LC -mass spectrometry plate. Shimadzu Biotech has also launched the Chemical Inkjet Printer (ChIP).[9] Using a combination of image analysis and inkjet microdispensing, ChIP can efficiently process membrane bound protein arrays. Protein arrays are electro-blotted to an ideal storage medium that is typically a polyvinylidene fluoride (PVDF) membrane. ChIP technology can deliver picoliter volumes of reagent to specific locations on the protein spot, which can then be subjected to protein identification by MS. The membrane can then be archived for further analysis at a later stage.[10] The major advantage of the ChIP technology is that the protein is stored on the membrane and only a small portion of it is consumed in an analysis and consequently, many analyses can be performed on one sample.

Genomic Solutions, a subsidiary of Harvard Biosciences, has leveraged its proprietary technology, Investigator™ Proteomic System, into almost every aspect of the proteomic workflow processes, such as protein separation using 2D electrophoresis and sample preparation for protein identification. Genomic Solutions has been able to provide a one-stop solution for researchers working on array-based experiments as well as spectrometry-based proteomics.[7, 11]

Trends in Electrophoretic Protein Separation: 4D Fractionation

As mentioned previously, a major drawback to 2DGE is that it cannot be used for low abundance proteins. Another general trend in protein biomarker discovery is to increase sensitivity by separating the sample into as many fractions as possible based on biochemical properties of the molecule (e.g., size, pH, pI, hydrophobicity, hydrophilicity). At the Wistar Institute, David Speicher has developed a novel four dimensional (4D) fractionation for the low abundance proteins in plasma & serum.[12] The 4D method fractionates proteins in four dimensions prior to analysis by MS. The procedure is able to isolate low abundance proteins by first depleting the top six most abundant proteins by immuno-affinity columns, followed by protein separation based on pH using Micro-Sol Isoelectrofocusing, next, running fractions on one dimensional PAGE and cutting out slices, and finally, digesting the slice with enzyme and running it on a nano-capillary reverse phase HPLC. The good news is that routinely ng/mL protein identifications can be made with occasional pg/mL protein identification. The bad news is that each sample generates 150 fractions and the entire process takes 11-15 days. However, if the biomarker of interest is in low abundance and can't be identified by traditional separation methods, 4D fraction appears to be a legitimate method to discover biomarkers. In infectious disease diagnostic testing, it is likely that many protein biomarkers will be in low abundance, and a method such as 4D fractionation could provide advantages for detecting these low-abundance biomarkers.

Trends in Electrophoretic Protein Separation: Lab-On Chips (LOC) Microfluidics

Caliper Life Sciences (Mountain View, CA) has virtually eliminated slab gel technology by automating the entire process on-chip. The LabChip 90 System streamlines the multiple manual steps of slab gel electrophoresis onto a small microfluidic chip.[13] The primary advantage of the LabChip 90 System is that it automates sample loading, electrophoresis, staining, destaining, detection, and quantitative data reporting, all in the channels of the microfluidic chip. These assays provide higher resolution, broader dynamic ranges and better reproducibility than agarose gel and SDS- polyacrylamide gel electrophoresis (PAGE) methods, while also generating quantitative sizing and concentration results in various digital data formats, including a simulated gel image.[7, 14] Biomarker discovery is facilitated by tools that allow researchers to analyze a number of proteins in a sample and compare the protein profile in one sample with the protein profile in a large number of other samples. The LabChip 90 automates the direct analysis of protein lysates providing a gel free platform for protein expression profiling.

Another microfluidic technology for protein separations is based on a compact disc (CD) developed by Gyros, AB (Uppsala, Sweden).[15, 10] The CD based microfluidic process is used to prepare samples for analysis by MALDI-MS (see below). Capillary and centrifugal forces are employed in tandem to load the sample into the microstructures. Preparation steps such as desalting, derivatization for chemically assisted fragmentation, enzymatic digestion, capture on immobilized metal-chelate affinity chromatography (IMAC) pads, and any combination of these steps can be carried out on the columns of each micro-structure in the CD. The major advantages of the CD approach is the low sample volume requirement of only 500 nL, in addition to the fact that 96 samples can be processed in less than one hour. Furthermore, separated proteins are captured on streptavidin-coated beads that can be recovered and used at another time. As the trend in biomarker discovery moves toward minimal sample requirements and the

ability to reuse samples for subsequent analyses, technologies utilizing microfluidics will start to gain greater acceptance in the proteomics biomarker discovery laboratories.

Protein Separation Technologies: Liquid Chromatography

High throughput liquid chromatography (LC) may yield higher sensitivity, as well as better throughput, than 2DGE but it does not address the issue of resolution. LC is a method by which large quantities of protein in their native state can be fractionated and then subjected to further analyses like mass spectrometry or functional protein arrays. The basic procedure entails the use of porous beads packed into a column, through which a liquid containing the sample is drawn. Proteins and other biomolecules are selectively retained within the column based on many biochemical and biophysical characteristics, such as size (gel filtration chromatography), charge (ion [anion or cation] exchange), specific binding affinity (affinity chromatography; affinity groups can be non-specific hydrophobic), IMAC (immobilized metal chelate ions), substrate, antibody, or differential partitioning (partition chromatography where the separation is due to differences in portioning in polar stationary phase [beads] or non-polar mobile phase [organic solvent]). This comes in reverse phase (RP) LC where proteins are dissolved in a non-polar mobile phase first, and then a polar phase is applied. Unwanted material passes through the column, while the protein of interest is eluted from the column and typically fractionated and automatically prepared for further analysis, usually by mass spectrometry. The key advantages of LC are that it is more automatable than 2DGE, and it can be used as a preparatory method for native proteins that can be subsequently used in functional protein assays.[3]

High pressure liquid chromatography (HPLC) principles are virtually the same as for any LC method, with the exception that a pump is used to push the aqueous sample through a small column under pressure. The primary benefits of HPLC over conventional LC methods are the ability to reduce elution volumes, to obtain higher resolution, and to reduce analysis times. In practical terms, this means a protein or peptide fraction eluted from an HPLC instrument can be applied directly to an electrospray ionizing mass spectrometer (ESI-MS) (see below).

Protein Separation Technologies: Protein Microarrays

Protein microarrays (sometimes called protein “chips”) provide a platform that can perform some aspects of both the separation and the characterization of protein functions and interactions. They often exploit the high affinity interaction between a protein and a ligand (e.g., an antibody, substrate, DNA or RNA, aptamer, carbohydrate, or small molecule). Either the ligand or the protein(s) of interest are immobilized as a series of very small spots (sometimes called features) in a regular, pre-defined pattern on a glass slide, a membrane, or in the bottom of microtiter plate wells. There are numerous variations on the general theme of protein microarrays, but they are generally classified into two types, ‘expression’ (or ‘analytical’) arrays and ‘functional’ arrays. In this report, the focus will be primarily on ‘expression’ (analytical) protein microarrays utilized in protein separation, enrichment, or fractionation for subsequent analyses. However, one promising “functional” protein microarray type termed the MHC Peptide Array (see below) shows great potential for being utilized in discovery of host response biomarkers to various diseases. For a recent, concise review on protein microarrays, please refer to Chen and Zhu.[16]

Analytical protein microarrays can be used to monitor the quantity of known proteins (alternatively referred to as expression). One of the most commonly used types of analytical microarray is the antibody array, where antibodies are arrayed on a solid surface and are used to determine the quantity of particular proteins in the sample. One of the major drawbacks to this approach for biomarker discovery is the requirement that an antibody (typically a monoclonal antibody) must already exist for a protein of interest, and be available to the researcher. Additionally, the monoclonal antibodies that are used to recognize the proteins of interest should be of high specificity and affinity, which is not always easy to achieve. All of these requirements tend to make the protein antibody microarray laborious, time-consuming, and expensive to manufacture. To this end, several alternative approaches have been developed to accelerate the production of alternatives to antibodies that still bind proteins with high specificity, such as, aptamers,[17] phage display-derived protein binders,[18] affibodies,[19] mRNA display,[20] and ribosome display.[21] While these ligands are very specific and bind with high affinity to their target proteins, they are all binding ligands that have been molecularly evolved to bind to known, defined protein targets. Though protein microarrays are not generally very useful in biomarker discovery applications, they are powerful tools to validate identified biomarkers, and may ultimately be developed for commercial diagnostic tests.

There is one caveat to the statement that antibody protein arrays are not practical for discovering unknown biomarkers due to the expense of making antibodies to all known and unknown proteins. The caveat lies in the scenario where if there were indeed available antibodies against all human proteins without the necessity of knowing the protein ligand of an antibody *a priori* to discovering it. If this were true, it would be possible to array the antibodies against all the human proteins and look for differentially expressed biomarkers in normal and diseased samples. If an unknown biomarker bound to an antibody, it could later be identified by a variety of analytical means. There is actually one company Milagen, (Emeryville, CA)[22] which claims to have made mouse polyclonal antibodies to all human proteins by a proprietary cDNA expression vector technology.[23] Milagen has been screening biofluids from patients in 15 major disease categories for the presence of differentially expressed protein biomarkers. Milagen is in the process of developing immuno-based diagnostic assays based on this unique approach to protein biomarker discovery. Even more interesting, Milagen claims it has made antibodies against all of the *M. tuberculosis* proteins and that it has antibodies that can distinguish between active and latent TB infections.[24] If indeed Milagen has the TB antibodies it claims, it would provide a significant jump-start to the possibility of developing a novel TB diagnostic assay.

The company Invitrogen has embarked upon an effort to produce as many human proteins as possible. They claim to have the manufacturing capacity to add 1000 to 3000 human proteins per year to their list of proteins that they can produce. They have commercialized a set of protein microarrays, and their microarray of human proteins currently has about 1000 proteins. These microarrays have a number of uses, though they could be used to evaluate the specificity of antibodies, which might be useful in the evaluation of antibodies that could be incorporated into a diagnostic assay platform. Perhaps autoimmunity diseases could also be investigated with the arrays.

Protein Separation Technologies: The Ciphergen Protein Chip®

Ciphergen has exploited the use of their ProteinChip® product in the discovery of patterns of protein expression (sometimes called signatures) that can be used as biomarkers. (Note: the Ciphergen Lifescience Research business was purchased by Bio-Rad.) The ProteinChip® is a platform by which proteins can be separated or fractionated based on biochemical or biophysical properties.[10, 25] The ProteinChip® is not a microarray, but rather a grid of different solid surfaces that can be used to rapidly prepare and enrich classes of protein. The chemical ProteinChip® has four separation surfaces: 1) an immobilized metal-chelate affinity surface, a hydrophobic surface, a cationic exchange surface, and an anionic exchange surface. After a sample containing proteins is incubated with the four surfaces on a ProteinChip®, a different spectrum of proteins associates with each surface. The proteins that associate with each surface are then subjected to surface enhanced laser desorption ionization (SELDI-MS (see below), and protein signatures (or profiles) of the samples are generated.

The Ciphergen technology has many advantages over traditional protein separation technologies described earlier. The use of four different surfaces results in a large discrimination of proteins during the pre-separation process. Greater than ten orders of magnitude in dynamic range can be achieved in the analysis, allowing low abundance proteins to be observed. Profiles of expressed proteins obtained via SELDI may represent over 3000 individual proteins or peptides, yet the analysis only requires approximately 200 uL of sample.[26, 27] Finally, the Ciphergen platform can be used in the validation of biomarkers once they have been discovered.

The major disadvantage of the Ciphergen SELDI-MS technology is that it is destructive in nature (e.g.,. the entire sample is consumed in the analysis) and consequently specific proteins cannot be recovered from the sample if it is found to contain a protein of interest. Additionally, the SELDI analysis results in a profile of expressed proteins (MS spectra) and does not necessarily identify any single biomarker unless it has been previously identified in a protein database.

Recently, Ciphergen announced the release of a new product called Equalizer Beads®,[25] which are composed of 64 million beads attached to different hexameric peptide ligands that bind to proteins in a biological sample. High abundant protein beads wash away preferentially over low abundant protein beads. The major advantage of this technology is there is no pre-fractionation of proteins, consequently no loss of sample complexity in the preparation process. As a result, low abundance proteins, i.e. those present at less than one ng/mL, can be detected. Currently, it is not known whether the equalization process will make it impossible to quantify the proteins in a sample.[26] Ciphergen intends to commercialize their technologies as a complete high-throughput platform for protein biomarker discovery, including sample preparation, detection, identification, and validation.

Major histocompatibility complex (MHC) peptide microarrays are used to measure the host responses to antigen presentation.[28] [29] MHC peptide arrays can both detect and characterize CD4+ and CD8+ T cells using peptide-loaded MHC molecules which have been immobilized in the same feature with multiplexed cytokine capture antibodies in a spatially addressable manner. MHC peptide arrays have been successfully used to detect and characterize T cell clones that specifically react to HIV, Vaccinia, and Influenza.[29] [30] The technology has also been used to correlate long-term patient survival to cytokine response patterns in a melanoma cancer vaccine trial[31] and to monitor changes in the recognition of epitope binding sites during a malaria vaccine trial.[32] Most recently, MHC peptide arrays have been used to detect low frequency autoimmune CD4+ T cells in a Type 1 diabetic patient.[33] [34] MHC peptide arrays have the potential to detect host-derived biomarkers upon exposure to pathogenic microorganisms. For example, MHC arrays could be designed to identify host immune response biomarkers in patients infected with *M. tuberculosis*, and possibly even to distinguish between active versus latent TB infections, HIV+/-, and PPD+ but TB negative.[35]

The MHC peptide array approach has many advantages over current technologies. MHC peptide arrays can detect low affinity T cell interactions that cannot be detected by flow cytometry. In addition, the sample size required for MHC peptide array analysis is significantly smaller than required by ELISPOT or flow cytometry, an important consideration when the samples available for analysis are small. Finally, MHC peptide arrays can detect physiologically-relevant secretions of activated T cells, and these T cells can be viably recaptured for future study, something not possible using intracellular cytokine staining. MHC peptide arrays can provide optimal screening and monitoring of immune responses. MHC peptide arrays do require that disease-associated antigens be identified in order to make MHC:peptide pairs for a disease-specific array. However, with the availability of complete genomic sequences for important infectious pathogens, such as HIV, *MTb*, *Treponema pallidum*, and *Neisseria gonorrheae*, it is fairly straightforward to derive the protein and peptide sequences from a complete genome sequence, in order to construct the proper MHC:peptide pairs to monitor the host response to a specific disease.

Protein Detection and Identification Technologies: Mass Spectrometry

The most ubiquitous technique used in the field of proteomics and biomarker discovery is mass spectrometry, which was developed over 50 years ago. Throughout the life span of the mass spectrometer there have been numerous improvements that included increased sensitivity, a reduction in sample quantity, higher resolution, the ability to quantify, increases in throughput, more powerful analysis algorithms, and lower instrument costs. The mass spectrometer is an instrument that measures the mass-to-charge ratio (m/z) of molecules such as proteins or peptides that have been electrically charged. A mass spectrometer typically couples an ionization device, a mass analyzer, and a detector. The most common ionization techniques are ESI (electron spray ionization), MALDI (Matrix Assisted Laser Desorption Ionization), SELDI (Surfaced Enhanced Laser Desorption Ionization) and FT (Fourier-transform). After the sample is ionized, mass analysis is determined through either time-of-flight (TOF), ion trap, or quadrupole time-of-flight (QTOF) analyzers.[7] For a current, comprehensive, and concise review of mass spectrometry and protein analysis, please refer to the article by Domon and Aebersold.[36]

In a typical biomarker discovery experiment, proteins of interest are excised from a 2DGE gel or fractionated by LC and digested with trypsin. The sample is then evaporated in a vacuum (desorption) and exposed to a high voltage, to convert the sample molecules into gas phase ions (ionization). The mass analyzer then separates the ions according to their charge to mass ratio (m/z) through the application of an electrical potential difference. Finally, a detector registers the number of ions at each m/z ratio and produces a mass spectrum, i.e. a mass frequency distribution.[3] A search algorithm is used to compare the experimentally determined peptide masses with theoretical masses of trypsin fragments for each protein from a genomic database. This process is known as peptide mapping, peptide-mass mapping, or peptide mass fingerprinting. The resulting "peptide mass fingerprint" can be used to search protein databases to identify unknown peptides or proteins in the sample.

Failure to identify a protein at this stage will lead to a more detailed analysis by using tandem mass spectrometry (MS/MS) (aka, orthogonal or 'coupled' MS). In this technique, selected peptide ions from the first round of MS are fragmented by energetic collision with gas, and subjected to another round of MS to provide amino acid sequence data in the form of collision induced spectra (CID). Although the CID does not directly provide sequence data, it provides additional information that can be used to further interrogate protein databases and identify an unknown peptide or protein.[3, 7, 36]

Ionization Techniques for Mass Spectrometry. Since proteins have low volatility, they must be ionized prior to mass spectrometry. Two ionization methods that are used most commonly are electrospray ionization and laser desorption ionization (LDI) methods such as MALDI and SELDI. When the goal is mass fingerprinting, the approach most often used is MALDI-MS or SELDI-MS (Table 1). Alternatively, for peptide identification patterns the preferred approach is tandem MS/MS using either electrospray ionization or MALDI. Mass fingerprinting relies on the accurate mass measurement of the separated proteins, whereas the peptide identification requires the fragmentation patterns.

Electrospray Ionization. Electrospray ionization (ESI) is best suited for determining the mass of very large proteins (>130 kDa) or complex mixtures of proteins. The typical sample is prepared for ESI from liquid chromatography or capillary electrophoresis (see Metabolomics section) and is directly injected (DI) into an ionizer tube (DIMS). In the ionizer, the molecules become highly charged in an electric field and are "sprayed" out of the tube into fine ion droplets. Mass determination either by Fourier Transform (FT), Time of Flight (TOF) or Quadrupole (Q) mass spectrometry is used to analyze the highly charged-ionized droplets. The output from ESI is a highly complex spectrum of peaks representing the charge of each molecule that is then compared against known protein databases in order to identify the peaks based on mass. Although electron spray ionization seems to be helpful for identifying post-translational modifications in particular, it does not seem to have a major impact in the proteomics workflow at present.[3]

Laser Desorption/Ionization. The two most common methods of laser desorption/ionization (LDI) are matrix assisted laser desorption (MALDI) and surface enhanced laser desorption ionization (SELDI). In these methods, the sample is presented on a solid phase "probe" and ionized using lasers. The main differences between the two methods are that MALDI requires pure protein samples and consequently is coupled with 2DGE purification or LC fractionation, whereas SELDI does not require purification of the

sample, and is best suited to analyze proteins from complex samples such as urine, blood, serum, and whole cell lysates. SELDI-MS has gained more traction in the discovery of biomarkers since it accepts "dirty" samples. The SELDI "probe" actually is used in the purification, extraction, and enrichment of the sample. As discussed above, SELDI-MS uses the ProteinChip® for purification and fractionation of samples based on general protein biochemical/biophysical properties.[7, 25]

A comparison of the ionization methods used for mass spectrometry is presented in Table 1.

Table 1. Mass Spectrometry Ionization Methods

Ionization	Description of Common Applications
ESI	Best used with proteins greater than 130 kDa or complex mixtures of proteins. Samples primarily derived from LC separation technology to yield 'clean' samples. Sensitivity at ng/mL. Used to analyze intact proteins aka Top-Down Approach Proteome coverage ~ 0.2% (200 proteins per analysis).
MALDI	Requires pure (and simple) samples, maximum of ~ 6 proteins. Best for proteins and peptides < 5 kDa. Used to generate protein or peptide fingerprint profiles. Typically coupled with 2DGE or LC protein separation to yield 'clean' samples. More tolerant than ESI to contaminants in sample (e.g. salts and detergents). Most commonly used MS platform to identify peptides < 5 kDa at ng/mL sensitivity. Degraded protein fragments typically labeled with isobaric tags for quantitative analysis. aka Bottom-Up Approach Coverage of proteome between ~0.01- 0.3% (6 - 300 proteins per analysis).[37]
SELDI	Accepts 'dirty' samples from complex mixtures such as blood, urine, serum, and cell lysates. Employs the ProteinChip® to purify, extract, and enrich the sample based on biochemical properties. Can detect low abundance proteins due to enrichment process at ng/mL sensitivity. Used to analyze whole proteins and naturally degraded proteins up to ~ 20 kDa. Proteome cover is ~ 1.0% (1000-2000 proteins per analysis).

Mass Analysis Technology. Mass spectrometry analyzers have seen extraordinary technical advancements throughout the years. Manufacturers are constantly 'out doing' their competitors on a yearly basis in terms of sensitivity, workflow, instrument costs and most importantly, major improvements in software and analysis algorithms to identify unknown proteins. MALDI technology has become an indispensable and state-of-art tool in the world of proteomics, and advancements in ionization methods have facilitated protein identification through techniques, such as digestion analysis and peptide sequencing, for post-translational modifications in particular. MALDI-TOF is typically used for studying relatively simple samples, whereas ESI-MS (aka Direct Injection MS-DIMS) systems are applicable for complex mixtures as they are typically coupled to liquid-based separation tools. A major trend in mass spectrometry is the coupling of two analyzers such as LC-Q-TOF MS-MS, in order to gain the maximum benefit from the respective combined strength of each analyzer. In Europe, proteomic laboratories are quickly adapting to SELDI-TOF, FT-MS, QTOF.[7] The following section briefly describes the current mass analyzers commonly used to identify protein biomarkers. A more basic explanation for some of the analytical techniques can be found in the Domon and Aebersold review.[36]

Time-of-Flight Mass Spectrometry (TOF-MS) separates ions of different masses based on the time needed for the ions to traverse a fixed distance. It is useful in identifying individual proteins, which can define the difference between a healthy and a disease state.[38]

IonTrap MS systems obtain peptide sequence information by fragmenting peptides followed by measurement of the fragment ion mass/charge ratios. While the mass accuracy is not as good as with quadrupole time-of-flight (Q-TOF) instruments, IonTrap MS is more sensitive. The sensitivity, reliability and lower cost of IonTrap instruments make it the best choice for most protein identification when *de novo* sequencing is not required.[3]

MALDI-TOF instruments are capable of analyzing sub-picomole amounts of peptides of mass less than six kDa with a mass accuracy of 50 parts per million (PPM), and approximately 100 PPM for proteins up to twenty kDa using internal calibration. Analysis of larger proteins requires more material, and the resolution and mass accuracy can decline rapidly at higher masses. One of the benefits is the ability to use mixtures of proteins and peptides and salts, and even some detergents are fairly well tolerated.[3]

Q-TOF MS is useful for peptides of up to approximately five kDa. It has similar sample requirements, with similar or better resolution, than a MALDI-TOF instrument, with the added advantage that MS/MS experiments can be performed to obtain sequence information about the peptides, or to obtain detailed information about posttranslational modifications. Peptides derived from proteolytic digestion of proteins by trypsin are particularly suitable for sequencing by Q-TOF MS, largely because of the size of the peptides and because they contain the basic amino acids lysine or arginine at their C termini. Sequences of up to 30 amino acids can be obtained in this way.[3]

Liquid Chromatography (LC) Q-TOF MS-MS spectrometers are hybrid Quadrupole TOF (LC-Q-TOF MS-MS) instruments that offer advantages for the identification of biotransformation pathways of drugs and biomolecules. This older technology has made a comeback due to advances in electronics and computer control. Its spectral resolution, coupled with high-speed data acquisition, makes it suitable for mass-spectrometric structural identification in the context of ultra-fast separation methods. The LC Q-TOF MS-MS can be used for parallel determinations on series of samples injected on several different columns. The Q-TOF technology is also suitable for multiple parallel kinetic studies.[3]

Analysis and Bioinformatics for Protein Biomarker Discovery

One of the most significant improvements in field of protein characterization over the past few years has been the development of protein analysis software and search algorithms that can help to identify proteins by comparing multi-parameter experimental data in protein/peptide databases. Nearly all manufacturers of mass spectrometers now tout their latest and greatest software packages to quantify and identify peptides from the mass spectrum. A good example of an integrated set of protein analysis tools that are used today is ExPASy (Expert Protein Analysis System), which is hosted by the Swiss Institute for Bioinformatics (SIB).[39], [40] By using the analysis packages at ExPASy, researchers can match their mass spec data with internet-based protein identification databases such as Pepmapper, PeptideSearch or ProteinProspector. Other data types, such as isoelectric point and molecular weight, can be aligned against protein sequence or mass spec data in MultiIdent and TagIdent search engines.[40] The irony in an

effort like SIB (and companies like Rosetta BioScience and Sage-N-Research, see below) is that while there are efforts to centralize the many varied proteomic analysis and bioinformatics applications, the proteomics field by and large still lacks the necessary standards by which data are generated and then can be shared across platforms, between laboratories, and within and across fields.

Taking note of the SIB example above, the trend now is for commercial organizations to deliver more value to their customer base by integrating several stand-alone protein analysis and identification software tools to improve the automation of analyses, and to therefore increase the analytical throughput and productivity in the discovery of protein biomarkers. As a representative example, very recently Rosetta BioSoftware and Sage-N-Research announced a collaboration to establish interoperability between Rosetta's Elucidator® system for protein expression analysis and Sage-N-Research's Sorcerer™ proteomic integrated data appliance to offer customers an advanced solution for protein expression research and protein identification.[41] The Elucidator system is a flexible, scalable system to manage and analyze large volumes of proteomic data that can automatically identify and list differentially expressed protein/peptides generated from mass spectrometry data. The Sorcerer is a data appliance that provides processing throughput for protein identification using proprietary software and hardware. In the future, there will be more announcements by companies regarding the integration of proprietary protein analysis and bioinformatics tools to provide better analysis solutions for the massive amount of proteomic data that is accumulating. However, what are truly needed are universal standards for protein analysis and informatics tools. With universal analysis and reporting standards the exchange of data, generated on different technology platforms, in different laboratories on different samples will be seamless, redundancy of experiments between laboratories will be reduced, and corroboration of data and discovery of biomarkers can be accelerated.[42]

Proteomics Summary and Discussion

The majority of thought leaders in the field of proteomics agree that all of the necessary analytical tools to discover protein biomarkers are in existence today. The primary analytical instrument is the mass spectrometer, in all of its various forms. While there will always be improvements in speed, sensitivity, reproducibility, and throughput, the mass spectrometer is meeting the needs of the proteomic research community today. Sample preparation on the other hand, seems to need major improvements in automation, ease of use, and overall speed. 2DGE is the dominant technology used for separating proteins to be analyzed by mass spectrometry. While efforts to automate the 2DGE process are being pursued by several commercial organizations, it is probably the microfluidics technologies, such as Lab-on-Chip (LOC), that show the most promise for making protein sample preparation routine, reliable, and less labor intensive. The attraction of LOC technology when analyzing clinical samples is in the significant reduction in the volume of sample that is required for analysis. For example, the CD approach developed by Gyros only requires 500 nL of sample for analysis by MALDI-MS.

The one area of proteomics that requires the most development is the field of data analysis software, protein/peptide search and identification algorithms. Enormous amounts of data are generated during each experiment and the data need to be compared against other protein/peptide fragment databases in order identify proteins of interest. Domon and Aebersold said it best "…incremental improvements in instrument performance will continue to translate into more-sensitive, faster and more reliable proteome

analyses. However, it is not clear whether such advances will be sufficient to eliminate the major bottlenecks encountered in the current proteomics approaches…proteomics needs to undergo a paradigm shift to reach the goal of robustly and globally analyzing proteomes. The essence of this shift is the transformation of proteomics from a mode where, in every experiment, the proteome is rediscovered, to a mode in which the information from the prior proteomic experiments is used to guide the present experiments. For mass spectrometry instrumentation and strategy, this shift of paradigm requires the development of instruments and data acquisition protocols that support the fast, sensitive and robust analyses of previously generated lists of target peptides."[36] In other words, protein biomarker discovery can be accelerated if the field stopped generating the same data over and over in separate laboratories, and instead could leverage the data that is already out there, but which cannot currently be accessed due to non-standardization in the protein analysis and identification algorithms and bioinformatics tools.

How much does it cost to analyze a sample for the presence of protein biomarkers? It turns out that question can be very difficult to answer if a detailed breakdown is desired or the answer is very easy if you solicit the services of an organization in the business of biomarker discovery. There is little to no publicly available information on the cost of performing a mass spectrometry analysis on a sample or even preparing a sample for MS by 2DGE for that matter. Steve Martin, Senior Vice President and Chief Technology Officer, of B-G Medicine (B-GM, Waltham, MA) shared invaluable insight on the topic of biomarker discovery costs. At B-G Medicine, the cost of analyzing a single primary sample for a variety of biomarkers is $10,000. Included for $10,000 is one protein panel, three metabolite panels (lipids, organic and polar compounds), one mRNA transcript profile, and collection of all metadata associated with the sample, (e.g. tissue, bacterial load, drug compounds, source, etc). B-G Medicine believes in a comprehensive approach to biomarker discovery, rather than focusing on a single class of biomarkers. B-GM uses LC-MS, GC-MS, MALDI (iTRAQ)-MS and NMR to analyze protein and metabolite biomarkers. When asked about the cost of a single MS run or an LC run, Martin stated, "nobody breaks it down to a single run, since a primary sample such as urine or plasma will generated 100s-1000s fractions that all need to be analyzed-it's not like a single MS or NMR analysis is done on the sample and you get a panel of biomarkers." He did say B-GM is transparent about their costs and a 30% margin is included in the $10,000 pricetag.[37] While Martin didn't breakdown the costs exactly by category, he did mention that it costs ~3.5 times more to generate a protein biomarker panel than it did to generate the three metabolite panels. By extrapolation one can estimate the B-GM price of approximately $7,800 per sample, with about $800 for an mRNA transcript profile, $1,500 for three metabolite profile panels, and $5,000 for one protein biomarker panel. Martin also mentioned that when he was employed at Applied Biosystems as Director of the Proteome Research Center, the academic institutions he interacted with did not calculate a per sample cost. In academic organizations the only concern was the initial capital instrument costs and not the running costs per sample since that cost is paid for via grants and departmental overhead charges.

A summary of common "profiling" methods used for the discovery of protein biomarkers and their characteristics is presented in Table 2.[40]

Table 2. Comparison of Technologies Used in Protein Biomarker Discovery

Method	Quantitation Capability*	Sample volume	LOD	Max # Proteins /Sample	Sample complexity to be analyzed	Protein-Peptide size (Da)	Sample thru-put
Separation Methods							
2DGE	+	Sample requirement high < 50 uL loaded but sample is concentrated. Total ug protein loaded vs vol.	Poor for low adundant + hydrophobic	< 1000	High	2000 -> 250000	3.0 days
IPG	+			< 2000	High		
NEPHGE	+			< 10000	High		
LC- Rev Phase Immuno-Affinity Ion Exchange	+++ + +		10^4 dynamic range	< 10	High - Samples first depleted for high abundance proteins In plasma-200+ proteins	Method mostly used with digested proteins Peptides < 5000	8/day
Ion Exchange/Rev Phase	+++			< 2000			
Ion Exchange -Chromato-focusing	+			< 2000			
4D Fractionation	+++		ng/mL-pg/mL	~ 150	High	< 20000	11-15 days
Mass Spectrometry							
ESI	+ No	< 15 uL	ng/mL	200 Shot-gun	Med – 100s proteins	130,000	200 per day
MALDI	+++ [iTRAQ]	< 50 uL	ng/mL	< 6	Low-6 proteins	<5000 -> 20000	1000s per day
SELDI-Protein Chips	+ Relative	200 uL	ng/mL	3000	High – 1000s of proteins	< 20,000	1000s per day
Hyphenated/Coupled Mass Spectrometry							
MS-MS	++	< 50 uL	50 fg/mL	< 2000 Shot-gun	High –100s of proteins	< 5000	1000s per day
LC-MS	+++	< 50 uL	fg/ml	Varies	Med-High-	Peptides	100s per day

* + = may be possible / moderate quantitation capability
++ = good / high quantitation capability
+++ = very high / excellent quantitation capability

2. Technologies for Discovering Metabolite Biomarkers

The metabolome, as defined by the Metabolomics Society, should consist only of those native small molecules (definable and non-polymeric compounds) that are participants in general metabolic reactions and that are required for the maintenance, growth, and normal function of a cell. The human metabolome is a list of all compounds capable of being produced by human cells that conform to the above definition. Examples of metabolic molecules are some vitamins, amino acids, nucleotides, lipids, sugars, ketones, alcohols, -OH, -NH, -SH functional groups, amides, amines, thiols, sulfo-acids, and organic acids. A sampling of molecules excluded from the human metabolome include enzymes, and other proteins and peptides, RNA, DNA, structural molecules (e.g., glycosaminoglycans, polymeric units), polymeric compounds such as glycogen, metabolites of xenobiotics (e.g., foreign chemicals like pesticides or man-made toxins), and essential or nutritionally required compounds not synthesized *de novo*.[43-46] By extension, the metabolome of microorganisms is defined similarly to that of humans and consequently many classes of metabolites are shared between the two. However, the specific molecules synthesized in each group can differ. Hence, in order to be able to assign the origin of a metabolite found in a patient infected with a foreign pathogen, it is imperative to have a fundamental knowledge or foundation of all metabolites found in man under normal conditions and under conditions of infection. An even more desirable scenario would be to also define the metabolomes of pathogenic microorganisms, independent of its host and under stages of infection. Knowledge of this kind is indispensable for the discovery of metabolic biomarkers (whether pathogen or host derived) associated with disease. Clearly, the growing field of metabolomics and specifically microbial metabolomics is an untapped source of information that has the potential of revealing disease specific biomarkers that could lead to specific and sensitive diagnostic tests.

In contrast to genomics and transcriptomics, which have dominant analytical technologies to interrogate biomolecules, in metabolomics there is no single technology that is suitable for the analysis of all of metabolic molecules. Chris Beecher, Vice President of Biochemistry and Technology at Metabolon Corporation states "while mass spectrometry is a tremendous improvement over other methods [e.g., NMR]… we have also found that no single technology will provide the entire answer."[47 45] This is due to the diverse structural and chemical nature of metabolites. Consequently, a mixture of technologies and protocols are employed to analyze metabolites, depending on the goals and objectives of the experiment as well as the specific chemical moieties under investigation. These strategies along with the optimal technical approaches are described in Table 3.[48]

While there are many strategies used in metabolite research depending on the goal, in biomarker discovery 'metabolite profiling/metabolic profiling' is the most commonly used strategy to identify and quantify biomarkers. In contrast, in efforts such as the Canadian Human Metabolite Project, a global metabolomics strategy is used since the goal of the project is to catalogue all metabolites in the human body.[44] In early 2007, University of Alberta, home of the Human Metabolome Project, in conjunction with Genome Alberta and Genome Canada, announced completion of the first draft of the human

metabolome. A total of 2,500 metabolites, 1,200 drugs, and 3,500 food components have been assembled in a first of its kind Human Metabolome Data Base (HMDB).[49]

The efforts of a group at the University of Alberta to generate the first annotation of the human metabolome are certainly a good first step, providing a basis from which to explore and expand this important component of human physiology. As mentioned in a recent Science article however, success in this case will depend on the ability to expand upon their "first draft" by detecting and identifying sub-micromolar concentrations of analytes.[50] It will also be important to be able to distinguish the normal-state human metabolome from disease-state conditions, and other naturally occurring variations in the metabolome within and between individuals. There is no doubt that, when completed, the annotation of the metabolome will become a very interesting source of biomarkers, but we are some years away from being able to claim that our understanding of the metabolome is complete.

Figure 2 illustrates the relationship between the desired outcome of the metabolic discovery experiment and the technological method(s) that can be employed to best meet the experimental objectives.

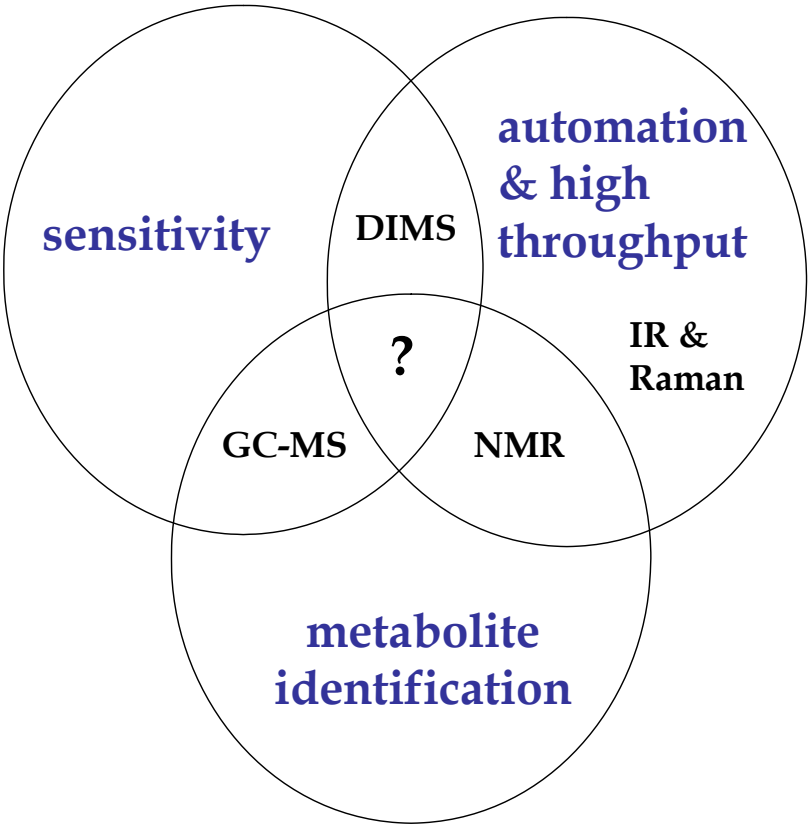
Table 3. Strategies for Metabolomics Research

Strategy	Description	Technical Approach
Metabolomics	Non-biased identification and quantification of all metabolites in a biological system. The analytical technique(s) must be highly selective and sensitive. No one analytical technique, or combination of techniques, can currently determine all metabolites present in microbial, plant, or mammalian metabolomes.	GC-MS LC-MS (MS)
Metabonomics	The quantitative measurement of the dynamic multi-parametric metabolic response of living systems to patho-physiological stimuli or genetic modification. This term refers to NMR data when used in metabolomics – typically in pharmacological analyses.	NMR
Metabolite or Metabolic profiling	Analysis to identify and quantify metabolites related through similar chemistries or metabolic pathways. Normally employs chromatographic separation before detection with minimal metabolite isolation after sampling.	GC-MS LC-MS (MS) NMR
Metabolite target analysis	Quantitative determination of one or a few metabolites related to a specific metabolic pathway after extensive sample preparation and separation from the sample matrix and employing chromatographic separation and sensitive detection. Metabolic pathways describe the processes involved in producing an end compound, or the processes involved in a disease state or disorder. A full metabolic pathway includes all intermediates. Of particular importance in defining metabolic pathways is the state of flux of the system.	DIMS CE-MS NMR
Metabolic fingerprinting	Global, rapid, and high-throughput analysis of crude samples or sample extracts for sample classification or screening of samples. Identification and quantification is not performed. Minimal sample preparation.	NMR DIMS IR and Raman Spectroscopy

Strategy	Description	Technical Approach
Metabolic footprinting	Global measurement of metabolites secreted from the intra-cellular volume in to the extra-cellular spent growth medium. High throughput method not requiring rapid quenching and time consuming extraction of intra-cellular metabolites for microbial metabolomics.	NMR DI-MS IR and Raman Spectroscopy

As in the case of protein biomarker discovery, the mass spectrometer is the workhorse analytical instrument platform for metabolomics applications. It is sensitive and allows the identification of metabolites by providing accurate mass spectrum interpretations and comparisons against libraries of mass measurements. Most samples introduced for analyses are prepared by chromatography (gas chromatography, [GC-MS]); high performance liquid chromatography [HPLC-MS], and capillary electrophoresis; [CE-MS]) prior to direct injection-MS (DIMS). As mentioned previously, one of the drawbacks to mass spectrometry is that it is a destructive method, especially when compared to other mainstay technologies used in metabolomics analysis such as Nuclear Magnetic Resonance (NMR) and vibrational spectroscopy (Raman, Infrared). Comprehensive reviews of technologies used in metabolomics research can be found in the articles by Weckwerth and Morgenthal, Goodacre, et. al., and Dunn, et. al.[43],[51],[43, 48, 51]

Figure 2. Analytical Methods versus Objectives for Metabolite Analysis.



The drawing depicts a matrix by which to select an analytical method based on the desired goal(s) of the metabolite analysis (e.g. sensitivity, identification, automation and or high-throughput). The three circles are interconnecting and include an analytical method(s) that can be utilized to achieve the desired goal. For example, if the goal of the experiment is to maximize sensitivity and to analyze hundreds of samples, Direct Injection-MS (DIMS) would be the best analytical method by which to achieve the experimental goal. The question mark represents the desired experimental goal. (from Dunn et. al.)[48]

One of the key points about metabolomics is the enormous volume of data generated by the various analytical methods. In single experiments such as a global metabolic fingerprint or a more extensive metabolic profile, hundreds of individual chemical structures can be detected while analyzing a sample. Consequently, a large number of potential disease-associated chemical structures remain unidentified. Moving forward, in order for metabolic biomarkers (either host or exogenously derived) to be linked to disease, it will be imperative to establish databases of all biomarkers derived from man and infectious agents. One example is the previously mentioned Human Metabolome Data Base.[52] These databases must be comprehensive in nature and contain annotated information about the metabolite such as location and concentration in normal and disease states, and the expression profile of the biomarker and structural information.

Mass Spectrometry in Metabolomics

Direct injection MS (aka ESI-MS) is the preferred method for analyzing metabolites since it is rapid, accepts crude samples or extracts, and is high-throughput in nature (1-3 min/sample, 1000 analyses /day).[48] ESI-MS is typically used for metabolite analysis rather than MALDI- or SELDI-MS due to increased sensitivity. While SELDI and MALDI are not typically used in metabolomic analysis today, their popularity will increase as the need to deliver global, rapid and high throughput analyses on a large sampling is necessitated. DIMS-MS has been successfully used to characterize and identify bacteria including *Escherichia coli* strains, *Bacillus spp* and *Brevibacillus laterosporus*.[48] With today's bench top mass spectrometers, the typical microbial metabolomics analytes are detected at micromolar concentrations. However, in clinical diagnostics one needs micromolar to nanomolar limits of detection in order to deploy a viable assay. Metabolon (Marlboro, MA) uses mass spectrometry and finds hundreds of metabolites per sample in the femptomolar range.[53] One of the major trends in mass spectrometry today is the use of orthogonal, multi-dimensional separations to increase the number of metabolites and sensitivity of the assay. Biocrates Life Science uses tandem MS/MS to measure low concentration metabolites in the nanomolar range without chromatographic instrumentation.[54]

HPLC-MS, CE-MS in Metabolomics

Due to the high resolving power of HPLC, samples containing complex mixtures of metabolites can be separated into well-defined spectra that provide greater signal to noise (S/N) ratio, which results in increased sensitivity during the analysis by QTOF-MS. According to a recent review of metabolomics technologies, applications of HPLC-MS in microbial metabolomics are small in number. In general, the lack of HPLC-MS applications to microbes is most likely due to the lack of a prioritized, coordinated, focused effort to investigate the metabolites of infectious bacteria by the academic research arena.[48] Capillary electrophoresis (CE)-MS is very similar to HPLC-MS in its ability to deliver highly resolved molecules for analysis by mass spectrometry. It is used primarily for analysis of polar and thermolabile compounds. However in the field of metabolomics, GC-MS and HPLC-MS are by far the favorite MS analysis tools in laboratories today. While these methods provide unambiguous identification and quantification of compounds in complex samples, their throughput levels are much lower than other technologies such as NMR, IR-Spectroscopy and direct infusion MS.[43] Obviously, increasing throughput

on HPLC- and CE-MS platforms will be a key to making these analysis platforms more useful for large-scale metabolic studies.

In June 2005, Human Metabolome Technologies (HMT, Japan) and Agilent Technologies announced a collaboration that is aimed at integrating HMT biochemical assays with Agilent's CE-MS capabilities to profile, identify and quantify metabolic biomarkers. The aim of the collaboration is to develop a new and unique mass spectrometry-based workflow for high throughput analysis.[55] Currently, HMT can detect more than 1600 metabolites simultaneously in a high-speed and high-resolution manner.

GC-MS for Volatile Metabolites

Metabolites are classified into two classes, volatile metabolites that do not require chemical manipulation prior to analyses, and non-volatile metabolites that must be chemically derivatized prior to analyses. GC-MS has become a valuable tool for metabolite discovery due to its sensitivity and comprehensiveness.[43] Volatile metabolites do not require chemical derivatization prior to being analyzed by MS. Consequently, sample collection is rather simple when looking for volatiles. Typical volatile metabolites are found in mammalian breath, gases absorbed onto absorbent solids (e.g., filters) and extraction of liquid or solids with solvents. Analyses of volatile compounds are routinely performed via gas chromatography (GC). For increased sensitivity and specificity, GC is coupled with mass spectrometry. GC-MS has been utilized for over 20 years to diagnose disease and in recent years utilized for volatile metabolite discovery and diagnosis. In fact, there is extensive data on and a non-invasive clinical diagnostic test for the diagnosis of *Helicobacter pylori* infections based on the detection of urea from infected patients.[48]

In contrast, sample preparation for GC-MS analysis of non-volatile metabolites is rather extensive and includes a drying step. This step is followed by a chemical derivatization that is used to increase volatility and thermal stability to the metabolite. Sensitivities for metabolites analyzed by GC-MS fall in the micromolar to nanomolar LOD range. GC-TOF-MS is used to provide complete mass spectra for all metabolites present in ultra-complex samples. Unknown or novel metabolites can be identified post analysis by comparisons with catalogued spectral data in annotated databases.[48]

NMR in Metabolomics

Nuclear magnetic resonance has been a favored analytical tool for chemists for over 50 years. NMR's main attraction is its specificity, while at the same time it is non-selective. This means each compound in a sample will have its own, unique resonance spectrum that provides rich and definitive structural information in the context of the molecule's chemical environment. NMR data are multi-parametric in nature (chemical shift, spin-spin coupling, relaxation) and allows quick identification of all chemical components contained in a sample, if the sample is not too complex. An NMR spectrum provides atom-to-atom connectivity information and is very informative for studying protein-structure function relationships. Unlike MS, NMR is non-destructive and does not require an upfront decision on how to prepare the sample (e.g., LC, GC, ion exchange) prior to analysis. NMR experiments are usually run in the liquid state so preparation of biological fluids (urine, blood, saliva, etc.) for NMR analysis is straightforward with a simple addition buffer to maintain pH.[48, 51] Hardware improvements to the NMR probe have lead to higher sensitivity with low volumes samples in the range of 2 to 20 uL. NMR is becoming more appealing for metabolomics since obtaining metabolite profiles in complex mixtures is

starting to gain popularity among researchers. As NMR databases become populated with more metabolite resonance information, it will be easier to identify unknown molecules in complex mixtures. Global metabolomics cataloguing projects such as the Canadian Human Metabolome Project will hopefully improve the ability to rapidly identify and characterize newly discovered metabolites in a variety of samples.

While NMR is typically used with solution state samples, it also has the ability to analyze intact tissue by a technique called magic angle spinning (MAS) NMR. In this technique, the sample is spun very fast which results in very sharp spectrum data in tissue sections that would be poorly resolved in conventional NMR analyses. The power of this technique lies in the fact that, unlike MS, it is non-destructive and does not require sample extraction, is more sensitive than FT-IR (see below) and yields the ability to actually visualize the compartmentalization of metabolites.[48] Time studies can also be conducted. MAS-NMR could possibly be used to directly visualize and discover metabolites in to infectious diseases. For example, the development of metabolite biomarker analysis in whole tissues by Magic-Angle-Spinning NMR (MAS-NMR) could be a method by which TB specific biomarkers are discovered. For instance, during latent TB infections, *M. tuberculosis* resides within macrophages. MAS-NMR might be used to analyze individual macrophages for metabolite profiles specific for the infected cell and compare it to the profile of an uninfected macrophage. This type of analysis could lead to a latent TB specific metabolite that differentiates it from an active TB infection.

Trends in NMR

Over the last several years, there have been many new advances in technology to improve NMR sensitivity and throughput. One of the major advances has been in the availability of higher-field magnets. In NMR, the signal quality effectively increases as the square of the magnet field. This means a 4X increase in field strength will yield a 16X increase in signal.[56] Additionally, coupled with an increase in higher magnetic fields, is technology to shield the magnet field with smaller footprints, thus freeing up precious laboratory space. Another innovation that has also lead to improved sensitivity in NMR is the cryogenically cooled probe. The probe is one of the main contributors to signal sensitivity but can also be responsible for introducing background or noise in the analysis. By cooling the probe, the noise in the system is reduced and the signal to noise ratio improves, usually greater than or equal to 4X overall. As mentioned previously, sensitivity is particularly important when analyzing proteins or metabolites for low concentration biomarkers. The two market leaders in NMR, VARIAN Inc. and Bruker BioSpin Corporation are constantly introducing NMR sensitivity improvements to the field. Finally, automation is the latest innovation in NMR. Typically, NMR is not considered to be a high throughput analysis platform. Recently, Protasis Corporation (Marlboro, MA)[57] has developed the One-Minute NMR automation platform. The system automates sample analysis from microtiter plates (384-well), uses less than 10 uL of sample and enables researchers to acquire data on at least 130 samples per day.[56] While this may not seem like high throughput when compared to mass spectrometry or LC, it certainly is an improvement over typical NMR throughputs of only about 10 samples per day (Table 3).

Raman and IR Spectroscopy in Metabolomics

Another analytical tool that is gaining popularity in the metabolomics research community is spectroscopy, specifically infrared (IR) and Raman spectroscopy. Optical spectroscopy primarily

measures the rotations and vibrations of molecular functional groups when energized with a radiation source (e.g., infrared or visible light) and the subsequent transitions to the atoms in the molecule such as electronic excitation, vibrational, and/or rotational changes. The significance of these transitions is each molecule or functional group has its own unique "fingerprint" spectrum that can be used to identify it in a sample.[48]

Over the past fifteen years, Raman spectroscopy has gained more acceptance in the field of biological sciences. There are many published reports of using Raman for the identification and characterization of microbial organisms such are *Escherichia*, *Enterococcus*, and *Staphylococcus* in clinical isolates.[48] Furthermore, resonance Raman and surface enhanced Raman spectroscopy (SERS) have been used to discriminate bacteria in urinary tract infections (UTI).[48] Interestingly, until recently Raman spectroscopy has primarily been used only in the area of microbial identification, not for metabolite discovery.

In contrast, IR Spectroscopy has been used in microbial characterization, biomarker discovery, medical diagnostics, and quality assurance. Originally IR Spectroscopy was used to classify microbial organisms, but with the development of Fourier transform (FT) -IR, applications moved into biological and medical applications. An IR spectrum consists of many bands arising from the vibrational motion within a molecule and is characteristically unique for a sample in regards to number of bands, frequency, and intensity. Consequently, as in Raman spectroscopy, a unique fingerprint of a sample can be established. In fact, in IR spectroscopy, a portion of the spectra is termed "The Bacterial Fingerprint" region since it is often used to classify bacterial organisms.[48] One of the primary applications of FT-IR is for the study of complex microbiological systems in medical and industrial settings. The technique is truly high throughput, rapid in nature, requires no reagents, and is relatively inexpensive. FT-IR has been used for the rapid identification of clinical bacterial in isolates to differentiate Candida and to characterize *Streptococcus* and *Enterococcus* species in addition to UTI samples.[48] The advent of FT-IR technology in clinical settings may lend itself as a possible diagnostic tool for resource limited environments particularly for the identification of the causative agent(s) in ALRI and diarrheal diseases.

Analysis and Informatics in Metabolomics

"*Data does not equal information; information does not equal knowledge; and, most importantly of all, knowledge does not equal wisdom. We have oceans of data, rivers of information, small puddles of knowledge, and the odd drop of wisdom.*" Henry Nix, 1990 "A National Geographic Information System – An Achievable Objective?"

Similar to the information challenges in protein biomarker discovery, one of the key challenges in elucidating the human metabolome, let alone the metabolome of infectious agents of man, is the massive volume of non-standardized data that is being generated daily and that already exists. When the data are not standardized in a universal open format, the ability to search databases for information within and across analysis platforms between laboratories around the world is tedious, time consuming, and inefficient. In addition to the elucidation of unknown metabolite chemical nature and structures, metabolomics data must also contain meta-information such as sample origin, tissue and experimental conditions.

The U.S. National Institute of Standards and Technology (NIST) provides one of the most comprehensive databases for chemical and structural properties of compounds.[58, 43] There are ongoing efforts to standardize mass spectrometry data from both commercial and academic organizations.[43] Unfortunately, many different research communities follow different analysis standards, which limits access to data derived from disparate laboratories. Matej Oresic, PhD, of VTT Technical research Centre of Finland [www.vtt.fi] confirms this sentiment. He states the two most pressing challenges in metabolomics today are 1) "…making sense of all the mounds of data generated" and 2) "the lack of an appropriate standardization of metabolomics data so that we can better compare and validate each other's findings".[47] This conundrum is perhaps made even more obvious in the case of Lipomics Technologies, the leading commercial service provider of qualitative and quantitative lipid metabolite profiles and lipomic information. While Steven Watkins, President and Chief Scientific Officer, Lipomics Technologies states, "A major challenge of metabolomics today is the need to integrate data from different metabolomics platforms," the company is actually developing proprietary analytical techniques and bioinformatics tools for the accurate and quantitative profiling of lipid metabolites.[47] Unfortunately, this type of conflict between doing what might be best for the field of metabolomics versus what might be best for the financial gain of a for-profit organization (which might in fact undermine the goals of the field of metabolomics overall) is the norm, not the exception, in the field today. During the course of the U.S. government-funded project to sequence the human genomes (e.g., the Human Genome Project), a similar lack of coordination between wet-lab and bioinformatics approaches plagued the effort for years, until a common vision prevailed that overcame the shortcomings of the individual contributors.

Amid all of the data being generated, and the data that currently resides in any number of databases, lays the necessity to convert these disparate kernels of information into "puddles of knowledge." Currently, two multivariate statistical analysis methods are being used to better visualize the interaction of many components in a biological system, Principal Components Analysis (PCA) and Independent Component Analysis (ICA). Both of these methods enable pattern recognition and biomarker identification using an unsupervised statistical analysis approach that eliminates bias in the analysis (that typically comes unknowingly from the researcher). Metabolite correlation analysis allows comparisons between networks and can be extended to integrate matrices of data about proteins, transcripts and environmental conditions.[43] This integrative approach places the field of metabolomics in a complementary position to proteomics and transcriptomics. Furthermore, it is the inclusion of data from all aspects and components of biological systems that empowers the growing field of Systems Biology or the attempt to understand the total biochemical interactions of an organism. There is excitement in the metabolomics field regarding the possibility that within the foreseeable future, a higher level of biological understanding will be achieved, which will lead to the identification of physiological and clinically relevant biomarkers that typically would not be possible.[43]

Similar to protein analysis, the collection and analysis of data from metabolomics experiments is a daunting problem. The data can be organized and evaluated in a number of ways, and the six types of databases that are necessary for metabolite biomarker discovery are presented in Table 4.

Table 4. Database Types Required for Metabolite Discovery Using Profiling Technologies[51]

TYPE OF DATABASE	DESCRIPTION
Metabolite Profiles	Databases storing detailed metabolite profiles, including raw data and detailed metadata (e.g., data about the data)
Single Species - Metabolite Profiles	Single species-based databases that will store 'relatively' simple metabolite profiles
Multi-Species Metabolite Profiles	Databases storing complex metabolite profile data from many species in many different physiological states
List of metabolites	Databases listing all known metabolites for each biological species, including suitable metadata. These databases could be extended to contain temporal and spatial information
Biochemical	Databases such as KEGG- a compilation of established biochemical facts
Integrated	Databases that integrate genome and metabolome data with an ability to model metabolic fluxes

Most importantly, these databases will have to be organized and configured in order to be useful to the wider community. In order not to repeat the same errors of the Human Genome Project's DNA sequencing effort, it is imperative to ensure that the data entered into these databases are accurate, validated, and annotated at the present time, so others can corroborate the data and compare it with their own findings.[43]

Metabolomics: Summary and Discussion

Metabolomics is one of the new '-omics' that is expanding at a rapid pace. As the scientific community realizes the value of understanding the interactions between small molecule metabolites and other biomolecules, the need to enlarge the knowledge base of metabolic compounds increases as well. The most important trend in the discovery of metabolic biomarkers today is the shift from simply studying metabolic pathways to analyzing the interconnection between networks of metabolic pathways. Goodacre et al., present the argument quite clearly regarding the absolute necessity to elucidate and visualize metabolic "neighborhoods" rather than pathways to understand the structural properties of the network.[51] This can only be accomplished at the level of the metabolome since the fluxes of metabolites between neighborhoods and the relationship of metabolites between networks cannot be calculated accurately from either expression levels of transcripts or from protein levels.

To this end, NMR and mass spectrometry (LC-MS, GC-MS) are the two main technologies used to study metabolite biomarkers while IR spectroscopy is starting to gain in its popularity for use in analyzing bacterial metabolic fingerprints and profiles. However, there is no single technology currently in use that is suitable for analysis of all metabolite molecules. Biofluids such as urine, blood and its fractions, and saliva are typical specimen types for studying metabolites. Frequently, when studying microbial metabolites, i*n vitro* cultures are employed as a starting point for analyses. Obviously, the objective of the metabolite experiment can drive the analytical method of choice but more frequently it is driven by the available sample size. For example, NMR routinely requires ~250 uL of starting material, compared

to the mere 15 uL that is required for an LC-MS analysis.[59] However, there is very little if any sample preparation required for an NMR experiment, while there is considerable sample preparation required for analysis by LC- or GC-MS analysis.

Oddly enough, as in the protein biomarker field, there is little mention of cost per sample for discovering metabolite biomarkers. Steve Martin (B-G Medicine) stated that his main criterion for choosing one technology over another (MS versus NMR) is based on the size of the sample, not the cost of the analysis (Table 5). For instance, if Martin only has 50 uL of urine (e.g. from a mouse or rat) to analyze, the choice of analytical methods is limited to LC-MS (see Table 5). But if the sample was 500 uL of plasma, analyses can be performed on both LC and GC-MS and NMR instruments. Even though costs per sample are difficult to determine, by extrapolation of data supplied by B-GM, a single metabolite panel (e.g. organic via GC-MS, lipid via LC-MS or polar molecules via LC-MS) will cost approximately $ 600 to perform (see page 19).[37]

The characteristics of the selected technologies used to profile metabolite biomarkers are summarized in Table 5. This table was adapted from Weckwerth et al.[43] with specific numbers in the Sensitivity column contributed by Steve Martin at B-G Medicine and Elwin Verheij at the TNO.[60] Please note that sensitivity in Table 5 refers to instrumentation sensitivity and not the concentration of the metabolite in the sample. Many factors and assumptions can influence the metabolite concentration determination including the original sample volume, metabolite concentration in the original sample, sample volume loaded onto instrument, pre-concentration or extraction steps of the sample prior to analyses. Consequently, when data are reported in an actual study the more realistic manner to view the data are according to the source of sample (e.g., urine, plasma, CSF, tissue, etc.), the type of analytes found and the concentration range in which the analyte was found.[59] The primary purpose of Table 5 is to describe a relative comparison between analytical techniques deployed in metabolite biomarker discovery. In terms of resolving and separation power versus analytes per sample, LC-MS, GC-MS and CE-MS are superior analytical techniques to all of the others, followed by LC-NMR and then by the single spectrum methods such as DI-MS and FT-IR spectroscopy.

The human metabolome is predicted to contain anywhere from 2,500 to 10,000 molecules. The range in number is based upon the variation within the different tissues and fluids. Also, the concentration can play a significant role in the estimated total. For instance, estimates of 2,500 entities in the human metabolome might be derived from experiments in which only metabolites present at concentrations of 1 uM or greater were detected. On the other hand, estimates of 10,000 entities in the human metabolome might be based on experiments in which metabolites occurring at concentrations of < 1 uM could be detected. Unlike protein biomarker discovery analysis, where at best 1% of the proteome can be evaluated in a single experiment (e.g., SELDI-MS), in metabolomics, about 5 to 10% of the metabolome can be analyzed in a single experiment depending on the concentration of the biomarker in question (Table 5). The identification of all metabolites within the human body is the one of the most important goals in metabolomics today. This goal becomes even more relevant when considering how to distinguish biomarkers related to infectious diseases such as TB, HIV, and ALRI from those normally associated with the host organism. Since there is overlap in classes of metabolites between the two, it is paramount to

identify all biomarkers in each organism separately under normal circumstances in order that relevance can be made upon the presence of a biomarker(s) during an infection.

Table 5. Comparison of Methods Used in Metabolite Biomarker Discovery

	Sensitivity	Analytes / Sample	Sample Size	Samples / day	Sample complexity
NMR	>10 umol/L	10's	250 uL	10-130	High-(urine) Low-(plasma)
LC-MS	>1 umol/L	200-300	10 uL urine 50 uL plasma	30	High-sample fractionated
GC-MS	>1 umol/L	1000	100 uL	8 (1hr per run)	Med-samples derivatized
CE-MS	High	1000	10 uL urine 50 uL plasma	30	High-samples fractionated
DIMS	High	100's	10 uL	< 1000	High-dirty samples
LC-NMR	Real Time: 200 umol/L Stop Flow: >10 umol/L	100's	100 uL urine	< 300	High
FT-IR Spectroscopy	20,000/1 S/N	10's		1000s 1min/ sample	Med

To this end, the Canadian government under the direction of Genome Canada launched the 'The Human Metabolome Project' (HMP) in January 2005. The project mandate was to identify, quantify, and catalogue all metabolites that can be potentially found in human tissues and biofluids at concentrations greater than one micromolar (1 uM).[44] The HMP project utilized mass spectrometry, chromatography, and NMR spectroscopy. In early 2007, University of Alberta announced completion of the first draft of the human metabolome. A total of 2,500 metabolites, 1,200 drugs, and 3,500 food components have been catalogued to date and included in the Human Metabolome Data Base (HMDB).[49] The data are freely accessible to all researchers and all compounds are publicly available through the Human Metabolome Library.

The HMP brought together several Canadian universities, hospitals, research institutes and industry.[44] However, the HMP is only mandated to provide chemical data and compounds to the scientific community. It does not have funding or the resources to use the 'raw' metabolites for disease identification and characterization.

Similar to the situation in protein biomarker discovery, the lack of standardization among data management procedures, analytical methods, and identification algorithms remains one of the most important bottlenecks to discovering new metabolite biomarkers. The problem is exacerbated further by

non-standardized methods for generating data. Steve Martin (B-GM) and others argue one of the most important problems to solve is the standardization of experimental design and data generation for biomarker discovery.[37] [36] For example, with the experimental design, a standard should be developed for such variables as the number of people in the group, the number of samples to be collected from each person, at what time points, how the samples should be obtained and prepared, and the wet-lab analysis methods to be used. Once these types of standards are established, information sharing and database searches between organizations and agencies will allow researchers to cross-reference information between different laboratories. This should lead to more rapid discovery of biomarkers.

3. Technologies for Discovering Nucleic Acid Biomarkers

The transcriptome is defined as the full complement of activated (or expressed) genes, as evidenced by the presence of mRNAs, (also called transcripts), in a particular cell or tissue at a particular time.[38] In transcriptome analysis, regions of coding and non-coding DNA are interrogated by a variety of technologies including DNA sequencing, DNA microarrays, and real-time PCR (RT-PCR, sometimes called qPCR for quantitative-PCR). However, the single most relevant technology in use today to profile gene expression patterns in order to identify novel RNA biomarkers is the microarray (also referred to as biochips, DNA chips, DNA arrays, gene arrays, and by various product names, such as the GeneChips®).[38] As a note, the role of DNA sequencing in biomarker discovery today is limited, with preferred applications being in identification, validation, and confirmation of putative biomarkers originating from other discovery tools such as microarrays. Though not covered further in this manuscript, emerging DNA sequencing technologies have been reviewed elsewhere. Sequencing multiple genomes looking for sequence differences is costly and time consuming, best employed when looking for identified sequence variations related to common diseases or drug resistance through SNP analyses, resequencing and similar strategies. As new technologies under development help to bring the cost of whole genome sequencing down below $1000, the role of direct detection of human sequence variations may play a larger role in DNA biomarker discovery.

Microarrays allow the analysis of thousands of genes in a single experiment. More importantly, whole transcriptome microarrays (often called whole genome microarrays) permit theoretically complete coverage of a transcriptome to be analyzed in a single experiment. Consequently, it is possible to determine the abundance of a single mRNA (or several mRNAs) that may be differentially expressed in clinical samples and provides discriminatory value as a biomarker(s) of the disease. Gene expression profiles from disease states and normal states can be used to identify a smaller set of candidate mRNAs that might provide sufficient discriminatory power in a commercial diagnostic assay. This smaller set of candidate mRNAs can then be validated using a number of technology platforms including single-mRNA amplification methods (such as reverse-transcriptase PCR) or an inexpensive and rapid technology called MLPA (Multiplex Ligation-dependent Probe Amplification)[61] or even using microarrays that contain a smaller number of features than whole transcriptome microarrays ("focused" microarrays). The technological approaches that allow the detection or quantification of mRNAs on a whole-transcriptome basis are presented in Table 6, as these are the most useful for biomarker discovery. While these ultra high density microarray formats can be used for diagnostics, it is the low-density microarrays that ultimately are used commercially for diagnostic products. Virtually all of the mRNA profiling methods

can be used successfully for validation since at this stage, the number of candidate biomarkers has been reduced to a manageable number (100s to low 1000s). Because bacterial genomes are significantly less complex than the human, and usually express less than 5000 transcripts, almost all of the methods can be employed for the discovery of mRNA biomarkers for pathogens. However, the higher density microarray formats are more desirable since the total input of sample would be less than the amount required to conduct several lower density array experiments to cover the entire bacterial transcriptome. Because the majority of RNA biomarker discovery is exclusively conducted on microarrays, this section will focus primarily on the use of microarrays for biomarker discovery.

Table 6. Technologies Used to Profile or Quantify RNA

Method	Discovery	Validation	Commercial
Microarray	YES Whole Genome Analysis > 50,000	YES-but most validation experiments need better quantitation	YES –AmpliChip p450 microarray FDA approved
Bead array	YES Whole Genome Analysis > 50,000	YES - but most validation experiments need better quantitation	Possible but array size is targeted
RT-PCR	NO Targeted Low number analysis < 400	YES – most widely used validation method due to accurate quantitation	YES-the most widely used method for RNA analysis
MLPA	NO Targeted Low number analysis 100s	YES – new technique that is quantitative	Possible but not yet developed
DASL	NO Targeted Low number analysis < 2000	YES – mostly used with FFPE tissue sections	Possible Used with degraded RNA

Microarray Technology

There are two major applications for microarray technology, mutation detection in genes (presumably, mutations can lead to a disease state) and mRNA expression level analysis (altered levels of mRNA may correlate with altered protein levels that lead to disease). Generic microarrays are typically constructed on a solid support (e.g., glass slide, membrane, silicon wafer, or micro-beads to which oligonucleotides (20-100mers) or cDNAs (> 500 bases) are arrayed by robotics or photolithography at element-to-element distances that range from > 100 uM to less than 2 uM, respectively). By convention, the arrayed DNAs are called probes, and the samples of mRNA (converted to cDNA and typically labeled enzymatically with fluorescent molecules by PCR or reverse transcriptase) are called targets. Fluorescently labeled mRNA (cDNA) targets are hybridized to the microarray, the microarray is then washed to remove non-hybridized/non-specific target molecules with the resulting hybridized targets visualized and analyzed by a fluorescence microarray scanner and concomitant software algorithms. By comparing samples from "before and after" (e.g., normal versus disease states) conditions, a difference in gene expression pattern may be detected or elucidated. It is this differential pattern of expressed genes that must under go further more quantitative analysis (such as via RT-PCR or MLPA) to validate the results of the microarray data and consequently, the correlation between expression levels and disease status. As just described, this series of events from discovery of an RNA biomarker on a microarray to validation with a secondary technique (RT-PCR, MLPA) is a routine process for developing diagnostic assays based on RNA biomarkers. The RNA biomarker discovery and validation process is ideally described by GeneNews

(Toronto, ON) which has developed the Sentinel approach based on the premise that circulating blood reflects the health or disease status by virtue of the fact that interactions between white blood cells and disease tissues induce a differential gene expression in the cells. It is the differential gene expression patterns in the white blood cells that serve as biomarkers for diseases such as cancer, heart disease and CNS disorders to name a few. The Sentinel approach first utilizes microarrays to identify hundreds of candidate genes that deviate from normal expression patterns. GeneNews then applies RT-PCR to reduce the number to between 2 and 10 marker genes based on sensitivity and specificity. This approach has helped GeneNews to identify marker panels for diseases that have stumped other diagnostic developers. For example, GeneNews has a panel of RNA biomarkers that distinguishes schizophrenia from bipolar disorder.[62] Most recently, GeneNews announced a prospective clinical study with Kaiser Permanente to evaluate its blood-based molecular diagnostic test for colon cancer.[63] This test is the first diagnostic application of GeneNews' Sentinel Principle molecular diagnostic technology.

The Sentinel principle appears to be a rational approach toward the discovery of host-related RNA biomarkers in response to infectious diseases. White blood cells mount the first line of defense against pathogens such as *M. tuberculosis*, HIV, *Treponema pallidum* and *Neisseria gonorrheae*, and microbes responsible for ALRI and diarrheal diseases. Once a panel of host RNA biomarkers was validated, a diagnostic test could be developed. The key distinction in developing an assay using white blood cells as the biomarker source, is that unlike detecting pathogen specific RNA biomarkers which will often be present in low concentration and difficult to detect, there are plenty of white blood cells in blood especially during a microbial infection, and therefore it may be possible to detect the biomarker(s) with little (linear amplification rather than exponential) or no nucleic acid amplification.

Advantages of Microarrays

The true power of the microarray in biomarker discovery is found in its ability to interrogate and profile the entire human transcriptome in a single experiment in a highly parallelized manner. Affymetrix, a pioneer in the microarray field, was the first company to introduce a commercial microarray in 1994. Affymetrix utilizes a proprietary photolithographic process to produce their GeneChips®. Over the years, Affymetrix has led the field by increasing feature density by shrinking the feature size. For example, the Affymetrix Human Exon Array with 5 uM feature sizes, contains over 5.5 million oligonucleotides and allows the entire human transcriptome to be analyzed at one time.[64] While Affymetrix has mastered the high density, two-dimensional microarray, this attribute may be unnecessary when it comes to conducting a transcript expression profile on microbial pathogens, because most microbial genomes contain fewer than 5000 genes. Affymetrix does not have a microbial GeneChip® commercially available at this writing but is experimenting with a 150 organism chip in the field and plans to introduce it in the future.[65] On the other hand, NimbleGen (whose microarrays are very similar to Affymetrix GeneChips®) advertises it has hundreds of microbial expression arrays designed and ready to use. The high density array platform is very powerful for identifying the human gene response to infectious agents, yet when it comes to RNA biomarkers for pathogens, many of the lower density platforms (or focused microarrays) remain a more realistic option to analyze the mRNA expression profiles of infectious organisms. In this case, low density or focused arrays that contain hundreds or thousands of features may be all that is necessary to profile the gene expression of a pathogen (see CombiMatrix below). Recently, Rachman et

al describe a unique transcriptome signature of *M. tuberculosis* in patients with pulmonary tuberculosis using DNA microarrays.[66] In this study, the authors performed a genome-wide expression analysis of *M. tuberculosis* from clinical lung samples. They used microarrays to identify differentially expressed genes and analyzed these data in the context of computationally-inferred protein networks to reveal a number of genes involved in the active fortification and evasion from host defense systems.

Limitations of Microarrays

One of the main disadvantages of DNA microarrays has always been the inability to reliably and accurately quantify expression levels of low abundance transcripts. Consistently and accurately measuring changes in expression that are less than two fold has also been difficult.[38, 67, 68] In fact, it is this inability to achieve absolute quantification using a microarray that has lead to the emergence of reverse transcription quantitative PCR (RT-PCR) as the mainstream technology for quantification of gene expression. RT-PCR is fairly simple to perform, is sensitive, specific, and relatively quantitative. RT-PCR is currently the method that is routinely used to validate gene expression data generated on microarrays (Table 6).[68, 67] Unfortunately, the number of genes that can be interrogated by PCR in one reaction is limited, typically to 5 - 10, due to the difficulties in multiplexing the assays into one tube. Parallelized individual assays are often used, but the amount of sample then becomes limiting.

Another issue that plagues microarrays (and for that matter, any RNA dependent protocol) is the fact that RNA in biological samples is an extremely labile and fragile molecule. RNA hydrolyzes easily and is susceptible to rapid degradation by RNAses. Consequently, samples must be collected and processed rapidly in order to inactivate RNAses. For example, in the practice of fixing tissues for pathology (a standard practice in the clinic), it has been estimated that over 400 million formalin-fixed, paraffin-embedded (FFPE) tissue samples have been archived in North America for cancer alone. Many of these samples have associated clinical outcomes — a potential gold mine of information when linked with underlying mRNA expression profiles. The problem lies in the fact that RNA becomes cross-linked and degraded by the fixation process. This results in rather discrepant data between matched tissue samples that have been either fixed or flashed frozen (freezing inactivates RNAses). These data indicated that RNA quantity and integrity obtained from fixed tissues were low and of poor quality, respectively.[69]

To address this problem and potentially tap into a gold mine of information, Illumina has recently introduced cDNA-mediated Annealing, Selection, extension and Ligation - the DASL Assay. The Illumina technology is based on the BeadArray™ that uses 3-micron beads that are situated at the end of a bundle of fiber optics (Array of Arrays™). The DASL assay provides a powerful gene expression solution designed to generate reproducible RNA profiles from degraded tissue samples such as formalin fixed, paraffin-embedded tissues. This is an exciting prospect for the discovery, validation and testing of biomarkers associated with complex diseases such as cancer or even for infectious disease pathogens.[70] While the DASL assay may alleviate some of the issues related to RNA degradation during sample collection, for practical matters, the fact that RNA is labile puts major constraints on sample collection from infectious disease patients in resource-limited environments, particularly if these samples are being acquired for biomarker discovery research.[14]

Finally, the cost of conducting whole transcriptome microarray experiments has typically been fairly expensive on a per sample basis. The routine cost for performing an experiment on an Affymetrix GeneChip® ranges from $300 to 700 depending on the array content (Table 7). This cost is similar to ABI's whole transcriptome microarrays but about 4 times greater than the Illumina whole transcriptome bead array ($160/sample). CombiMatrix will be able to fit two whole transcriptome arrays on their 90K array at a final cost of about $100/sample (see below).[71] While Affymetrix may argue that the cost per feature or data point has been reduced by orders of magnitudes, the simple fact remains that it costs several hundreds of dollars to analyze one sample on a GeneChip®. In direct response to Affymetrix, an entire industry has emerged with apparently one goal in mind, to provide microarrays less expensively than Affymetrix. While it is possible for academic laboratories to produce their own microarrays rather inexpensively (on the order of $5-10 per array), this is not a recommended way to proceed if one's goal is to achieve high quality, consistent, reproducible, and reliable data. Quality control and quality assurance issues often plague these "home-made" microarrays. For consistency, reliability, and data quality, purchasing microarrays, analyzers and software from a commercial supplier is highly recommended.

The characteristics of a number of technology platforms used to profile large numbers of transcripts are summarized in Table 7.

Table 7. Comparison of Selected RNA Biomarker Analysis Technologies Currently in Use

	Affymetrix GeneChip U133 2.0 +	**Illumina BeadArray Human 6**	**ABI Human Genome Survey v 2.0**	**ABI-TaqMan Low Density Array**	**MRC-Holland MLPA Tubes/wells**
Array type	silicon substrate microarray	virtual bead array	nylon/glass substrate microarray	microtiter plate-based array of individual assays	microtiter plate-based array of individual assays
# Transcripts Evaluated	47,000 Whole Genome	48,000 Whole Genome	33,000 Whole Genome	380/array Targeted	45/tube Targeted
% transcriptome Coverage	100	100	100	10^-7	10^-8
Most appropriate use	Discovery	Discovery	Discovery	Validation and Diagnostics	Validation and Diagnostics
Sensitivity	0.7-0.15 pmol	0.15 pmol	fmol	10 copies/cell	3-4 copies/cell
Assay Sample Amount RNA (amount blood)	10 ng (small) 1 ug (std) (500 uL blood)	50-100 ng (500 uL blood)	500 ng (500 uL blood)	100 ng (50-100uL)	20-500 ng (10-500uL)
Samples per Array Vessel	1	6	1	1/array	1/tube

	Affymetrix GeneChip U133 2.0 +	Illumina BeadArray Human 6	ABI Human Genome Survey v 2.0	ABI-TaqMan Low Density Array	MRC-Holland MLPA Tubes/wells
Samples per Day	25 2 fluidics station 1 scanner	192 1 autoloader	72 Hybridization off-line 20 min reads	24 (arrays)	1000's depending on # of tubes (wells)
Cost of Sample Prep	$100	$13	$13	$13	$13
Cost per Sample	$675 per array	$160 ($960/array)	$625 per array	~ $ 350/array ($33K/genome)	10/sample (~$7K/genome)

Summary and Discussion

Paolo Fortina, MD PhD, Professor of Medicine at Thomas Jefferson University, has recently conducted a comparative analysis of commercially available microarray formats (Affymetrix, NimbleGen, and CombiMatrix) to detect the presence of very small deletions in chromosome 22q in the DiGeorge syndrome.[72] He found that all three sources, in general, provided excellent results wherein small deletions (2- 3 bases) and low copy numbers (3-4 copies) were detected on the arrays.[73] The primary differentiating factors between the 3 platforms were background, ease of use and cost. While all three microarrays had low background, the platinum surface on which CombiMatrix produces its arrays had the lowest background noise level. In regards to NimbleGen, while their microarrays resulted in the "best of the best" data, the fact that a researcher must send their samples to Iceland to be analyzed by NimbleGen remains a major detraction in using the NimbleGen platform.[73]

CombiMatrix was by far the easiest and least expensive to use in the Fortina laboratory. CombiMatrix is now selling their CustomArray™ Synthesizer in addition to providing a custom array manufacturing service. While the capital cost for the instrument is ~ $250,000, the cost to run samples on a custom array, manufactured in your laboratory with overnight availability is about an order of magnitude less than an Affymetrix microarray analysis. The CombiMatrix microarray (12K and soon to be 90K chips) can be sectioned into 32 array regions and can be reused up to 4 times. The end result is about $20 to conduct an assay on the CombiMatrix system.[71] In terms of lower costs and greater flexibility, the CombiMatrix technology may be one of the most viable platforms to conduct microbial RNA biomarker discovery on since the assay and array configuration would accommodate microorganisms with < 5000 genes easily, while at the same time maintaining lower assay costs yet delivering high quality and sensitive results.

One of the most appealing features of a high-density microarray is the ability to analyze the entire transcriptome of any organism in single experiment. Currently at this time, there is no other biomarker discovery technology in use today that has the capability to cover the entire –'ome' in a single assay other than the microarray (Table 8). While this ability is appropriately applied for the study of most eukaryotic

organism transcriptomes, the high-density microarray is probably 'over-kill' when applied to an infectious disease organism transcriptome since it is not as complex as in humans.

Probably one of the most interesting approaches to RNA biomarker discovery today and the one with the most relevance to infectious diseases is the Sentinel Principle employed by GeneNews.[74] The idea that white blood cells (WBC) circulating within the body, will display differences in RNA expression profiles based on the status of disease within the body is quite elegant. The fact that microarrays are used to interrogate the host WBC expression patterns rather than the infectious organism's RNA profiles is very important for issues coupled with sample collection and assay specificity. For example, take the case of active versus latent tuberculosis infections, where the presence of acid-fast-bacilli (AFB) is indicative of disease TB. It is most likely that the body's WBCs will have a different yet distinguishable RNA expression pattern during an active infection as compared to a latent infection. During a latent infection, the presence of circulating AFB is not detectable. Furthermore, since it is the host WBC RNA profile and not the *M. tuberculosis* profile being analyzed, the number of WBCs in circulation in comparison to *M. tuberculosis* bacilli will be orders of magnitudes greater and that will enhance the sensitivity of the RNA analysis. Overall, the Sentinel method may be one of the most straightforward approaches to rapidly establishing, credible and reliable RNA biomarkers for the diagnosis of many different infectious pathogens.

4. Discussion and Recommendations

Observations

While evaluating the technologies used to discover biomarkers for identifying novel proteins, RNA and metabolites, several points of interest became apparent. The first point is that the technologies used to discover biomarkers are mature and established platforms in their fields. In proteomics and metabolomics, mass spectrometry and NMR, respectively are the prominent analytical tools. Both of these technologies are about 50 years old. In regards to sample preparation, the primary methods are two dimensional gel electrophoresis and liquid chromatography-both over 30 years old. In RNA discovery, the DNA microarray is mainstay and this technology is over 17 years old. The point is these are all established technologies that researchers can depend on to obtain results and are accepted warmly in the laboratory, warts and all. The consensus in the current literature and among many experts in the field is there is no need to develop new technologies to discover biomarkers. The manufacturers of these instruments will always be making incremental improvements that make them faster, more sensitive and less expensive. A summary of many of the commonly used analysis platforms and features associated with human biomarker discovery in transcriptomics, proteomics, and metabolomics is presented in Table 8.

The magnitude of the problem for the discovery of all possible biomarkers lies in the complexity of the "-ome" being analyzed. For example, the transcriptome (RNA) is estimated to contain between 30,000-50,000 molecules. It is currently, the only class of biomarker that can be evaluated in its entirety, in one experiment. One of the main reasons why the transcriptome can be analyzed easily is the fact that its analytes can be amplified *in vitro* from the amount found *in vivo* (e.g., RT-PCR or T7-amplification). On the contrary, analytes found in the proteome and metabolome cannot be amplified with existing

technologies. Endogenous concentrations of protein and metabolite analytes can differ by over fifteen orders of magnitude. Consequently, the technologies used to discover these biomolecules (e.g., MS and NMR) do not always possess the necessary sensitivity to detect the analytes at low concentrations. This is the case in metabolomic research where the metabolome is estimated to contain 2500 different molecules found at greater than 1 uM and 10,000 molecules found at concentrations less than 1 uM.

Table 8. Comparison of Discovery Technologies for Different Biomarker Types

	RNA	**Proteins**	**Metabolites**
Coverage of "-Ome" Based on Best Technology Available	99-100 % 30,000-50,000 transcripts	0.01-1% 1,000,000 proteins	0.2-10% 2,500-10,000 molecules
Analytical Method (% Genome Coverage)	DNA Microarrays (100%)	2DGE (2-10%) ESI-MS (0.2%) MALDI (0.01%) SELDI (1.0%) MS-MS (1.0%)	NMR (0.2%) Mass Spectrometry LC-MS (~10%) GC-MS (~10%)
Cost per Sample (or Study)	$160-675	$(5000)[36]	$(500)[36]
Throughput	25-192 samples/day	2DGE- 1per 3 days LC - 8/day MS – 200-1000s/day	MS 8-30/day NMR 10-130/day
Minimum Clinical Sample Volume	500 uL 10 ng-1 ug	15-200 uL	MS 10-100 uL NMR 250 uL
Sample Processing at or Near Collection Site	YES $10 (must inactivate RNAses)	YES (must inactivate proteases)	MS YES NMR NO (+/-)
Sample Complexity	HIGH 10,000s of different mRNAs	DIMS – MED (100s) MALDI - LOW (10s) SELDI - HIGH (1000s)	Spectroscopy HIGH (1000s) NMR – MED (100s)
Sensitivity of Analytical Method	HIGH (can detect about 10 copies of mRNA)	MS MED (nmol) MS-MS HIGH (amol)	MS - HIGH (amol) NMR – LOW (nmol)

Newer technologies, for example microfluidics (Lab-On-Chips) for protein separation are slowly gaining popularity but will be used to complement rather than replace the older technology in the lab. Protein arrays, while being touted as the best thing to hit proteomics, are still years away from being a useful biomarker discovery tool for the human proteome, but are perfectly suited for validation and commercialization applications. These arrays contain the proteins the analyte researchers are looking for, instead of the more useful tools, antibodies against the proteins, that are not available in sufficient numbers at this time. The simple fact is that while it is possible to create a whole genome DNA microarray for the 30,000-50,000 genes in the human that can be used to scan for all possible RNA biomarkers, the task is much more daunting to create a whole human proteome chip since it is much more

difficult to generate the required reagents (e.g. antibodies and proteins) and since the number of proteins is estimated to be as many as one million. So today, protein arrays represent only subsets of the whole human proteome and are designed in a selective, biased manner in contrast to a whole proteome chip design where all biomarkers could be analyzed in a global manner at one time. Microbial proteomes, however (and microbial genomes for that matter) are significantly less complex than the human proteome. Consequently, it may in fact be possible to utilize a whole microbe (e.g., *M. tuberculosis*) proteome array for discovering of biomarkers if antibodies can be prepared against all of the microbial proteins. As mentioned previously, Milagen (Emeryville, CA) claims to have generated antibodies against all *M. tuberculosis* proteins. Whole microbe proteomes could be a valuable biomarker discovery tool if antibodies are developed that recognize microbial proteomes (see Recommendations below)

Another observation extends the notion of a global or comprehensive approach to biomarker discovery. The other term used to describe this idea is "profiling", as in expression (RNA) profiling, metabolite profiling and protein profiling. The force behind profiling is based on the premise that a panel of biomarkers is much more reliable, specific, and diagnostic of a disease than a single biomarker. The technology choices being made today are those that favor the analysis of as many analytes as possible in a sample at one time. Consequently, technologies like MALDI-MS, NMR, and whole transcriptome expression arrays (Affymetrix and NimbleGen) are the popular platforms to observe hundreds to thousands of potential biomarkers in one experiment, respectively. However, these profiling types of experiments generate massive amounts of data that need to analyzed, interpreted, compared and appropriately stored in order to be used to identify a biomarker(s).

And so this leads into the third observation – bioinformatics and data analysis standards, or lack thereof. In all three biomarker discovery fields, proteomics, transcriptomics and metabolomics there is a lack of standards to analyze data, to interpret data, to store data and as a result, bioinformatics is identified as the major bottleneck to biomarker discovery.[3, 7, 38, 42, 48, 75] The lack of software standards prevents laboratories from exchanging data easily in order to corroborate findings and or to search within databases to help identified unknown analyte spectra or profiles. The problem is exacerbated even more in systems biology where a holistic approach is taken to analyze a system rather than just a single component of the system, for example a protein(s) or a metabolite(s). In systems biology, it is imperative to analyze all inclusive data sets made up of DNA, protein, RNA, metabolite, glycomics and other –omics data in a comprehensive manner in order to understand the connection between the networks of systems within an organism. For instance, at B-G Medicine (Waltham, MA) a biomarker discovery company, the biomarker discovery process takes every sample through a series of analytic instruments and methods to ultimately yield panels of protein, metabolite and RNA biomarkers.[37] A universal standard for data recording, analyzing, reporting, exchanging and storing would make more of an impact on biomarker discovery than just about any other advancement in technology.[48]

The antithesis of establishing software and data management standards is no better exemplified than by the state of DNA microarray software analysis market in the U.S. In the U.S. alone, there are 30 companies making DNA microarray analysis tools.[75] This large number of organizations is competing to gain market share and do so primarily by making proprietary software tools that ultimately make it difficult to share data across platforms and laboratories if so desired. Somehow, software and

bioinformatics standards have to be established for all aspects of protein, RNA, metabolites, and other –omics research as well. There needs to be an ability to exchange data and information easily among researchers. As Domon and Aebersold state in their recent review, there needs to be a paradigm shift away from rediscovering the proteome in every experiment to where the information from prior proteomic experiments is used to guide the present experiments.[36]

Very recently, the NIH announced a 5-year, $13.7M award for the Tuberculosis Structural Biology Project.[76] The award was made to a multi-disciplinary group of laboratories from four universities whose goal is to determine the three-dimensional structures of a 'large number' of *M. tuberculosis* proteins. While the concept is a good start, 5 years is a long time to wait for results. All of the biomarker discovery technologies are available today to identify new biomarkers and could do so as long as there is reason and a passion to do so.

Recommendations

At this juncture, there are several actionable recommendations regarding the establishment of programs directed at biomarker discovery for diseases such as ALRI, malaria, diarrhea, HIV, TB, syphilis, Chlamydia and gonorrhea that can be made based on the availability of technologies today. As identified previously, bioinformatics is a key bottleneck in biomarker discovery efforts today. In all three biomarker discovery fields, proteomics, transcriptomics and metabolomics there is a lack of standards to analyze data, to interpret data, to store data and to exchange data. Included in this assessment are the data collection requirements as well as the data base and statistical analysis tools. It is essential that agreement be reached on bioinformatics standards and that resulting requirements be made widely available to help minimize redundancies in discovery efforts and to facilitate application of experimental data across research efforts.

In each disease case, the existence of appropriate sample banks is also critical. This subject is addressed in detail in each specific disease report. However, the establishment of standardized sample collection requirements, including best available sample types, best practices for collecting, transporting and storing samples and quality control methodologies are needs that cross all disease areas. Related sample collection challenges critical to sustaining efficient and effective biomarker discovery efforts include recommended collection media and collection devices, validated for each sample type.

For diseases where multiple organisms could be the etiological agent (e.g., ALRI) or where several organisms can be infecting the host simultaneously (e.g., febrile children), fundamental studies concerning the overall microbial flora as a function of time, location and intervention are warranted and possible. In instances where multiple organisms are identified in the infected host, it is important to identify those that are the most critical causative agents, considering both health and economic perspectives. These studies can be conducted by sample collection at the clinical site and transport to laboratories with more sophisticated capabilities. These studies can be used to more completely understand the nature of the microbes being fought in the field and to monitor antimicrobial resistance. Although U.S. and European labs are now well equipped to conduct these studies, it is possible to disseminate the technologies to sophisticated labs in the countries under study. One laboratory that has been developed to conduct such studies is at Ibis division of Isis Pharmaceuticals in Carlsbad, California.

The protein and metabolite biomarker discovery fields both suffer from poor sensitivity of current discovery techniques and the lack of suitable profiling capabilities in complex biological samples. Analytes found in both the proteome and the metabolome can not be amplified and technologies used to discover biomarkers are not sensitive enough to detect low level endogenous concentrations of analytes. Unlike the genomics area where entire transcriptomes can be viewed in single experiments, protein biomarker discovery efforts visualize at best 1% of the proteome in a single experiment and metabolomics discovery efforts visualize from 5-10% of the metabolome present, depending on biomarker concentration and assay sensitivity. Additional focus on sample preparation techniques, such as microfluidics-based approaches like lab-on-chip, warrant further research. Ideally such techniques will be faster, utilize less sample, be non-destructive in nature and generate more consistent and reliable data.

For both proteomics and metabolomics, the ability to profile, or distinguish multiple biomarkers (proteomics) or metabolite networks (metabolomics) simultaneously in a biological sample is an area warranting further study. Moving from detection of individual proteins to panels of proteins will lead to the development of more sensitive assays, better able to distinguish disease state from normal, especially in the instance of multiply infected hosts. Likewise, the ability to detect metabolic networks, or interconnected metabolite pathways, may allow the development of tests currently not possible. An example of the latter may be an assay that distinguishes latent from active TB infection.

In the area of protein biomarker discovery, the use of analytical protein microarrays to determine a microorganism's protein expression profile is worth pursuing. Although an array composed of all the proteins of an organism could be useful in determining whether or not differential immune responses to protein antigens is indicative of active or latent disease, the most useful protein microarray would consist of antibodies raised against the entire proteome of a microorganism. As is the case of viruses such as HIV and HBV, where proteins can be produced independent of the production of the complete organism (e.g., p24 and surface antigen, respectively), perhaps this will be seen in other microbes, too. Antigen and antibody patterns against HBV have been particularly revealing. With the appropriate tools, it should be possible to generalize these types of investigations.

Contracting with a company that can produce antibodies against pathogen proteomes is a realistic path to pursue. In the case of tuberculosis, one company, Milagen claims to have generated polyclonal antibodies against four *M. tuberculosis* clinical isolates. To date, they have isolated ~3250 mouse polyclonal antibodies and screened about 800 of these antibodies and identified about 8 antibodies that may be able to differentiate between active and latent TB. Milagen is currently not interested in working on tuberculosis (due to business priorities and resource issues) and does not have any definite plans for moving the TB antibodies toward a diagnostic assay.[24]

Focusing on host response biomarkers to disease, MHC peptide arrays are probably the most promising technology to come along in a while that could be used to identify the specific host T-cell response to a specific infectious microorganism.[32], [31] Since the assay monitors the immune response of CD4+ and CD8+ T cells in the body in response to an infectious agent (or disease), very specific patterns or profiles of T cell responses (which are unique for every infectious agent) are identified that will be uniquely specific for any pathogen. Effectively, the unique patterns of CD4+ and CD8+ T cell types, in addition to

the cytokines secreted by them, could translate into a pathogen specific biomarker. Currently, ImmunoCyte, LLC (E. Hartford, CT) is in the early stages of developing the MHC Peptide array for a variety of commercial applications.[77, 78]

Turning to nucleic acid biomarkers, a focused RNA biomarker discovery effort could quickly be set up with a commercial organization such as NimbleGen or CombiMatrix to discover pathogen specific biomarkers. NimbleGen states that it already has the microbial microarrays while CombiMatrix has the ability to quickly and easily make microbe microarrays and to do so less expensively than NimbleGen. Contracting with a commercial organization would be the most efficient use of time and resources and would lead to an infectious disease RNA biomarker panel in the shortest period of time possible.

One potential benefit that might arise from a directed and focused RNA discovery program is that knowledge of differentially expressed microbial transcripts could lead directly to the investigation of the proteins they encode, which might be more easily measured in a diagnostic assay to be used in resource-limited settings than an RNA biomarker. However, it must be remembered that the overlap in proteome and transcriptome expression, where proteins and mRNA are quantified separately and directly, can be small (e.g. 10% in a recent oncology study; K. Ordonez, JPM Conference, San Francisco, 2005). A pilot study using both approaches in target organisms should be carried out.

GeneNews' Sentinel Principle warrants a serious evaluation as a rapid and directed method to identify host response biomarkers to infectious microorganisms.[79] The idea that white blood cells (WBC) circulating within the body, will display differences in RNA expression profiles based on the status of disease within the body is quite elegant. The fact that microarrays are used to interrogate the host WBC expression patterns rather than the infectious organism's RNA profiles is very important for issues of sensitivity coupled with sample collection and assay specificity, especially for infectious diseases in resource-limited-sites. In situations where the infectious pathogen is in low number and is difficult to isolate or to get enough of for analysis, the Sentinel approach obviates this limitation since the biomarker is host-derived, not pathogen-derived. White blood cells will not be in short supply in most cases and they provide a large source of biomaterial to analyze.

One other approach to characterize host response to infection generating significant interest today is the utilization of rapid sequencing technologies, including microarray based SNP detection, mismatch repair detection, massive parallel sequencing and others. Companies such as 454 LifeScience and Illumina (Solexa) are engaged in these technology development efforts.

Finally, though not directly applicable to the study of biomarker discovery, one must not loose site of how these biomarkers will be utilized in the field. Appropriate systems for field deployment are not available for each disease indication today in all target locals. Whereas some diseases, like HIV, can very likely be effectively diagnosed and monitored using existing rapid test methodologies, for others, such as ALRI, there is no suitable technology today. As explained further in the report on ALRI, it is postulated that volatile organic compounds may prove to be good biomarkers for the detection of disease. This is in part because readily available ALRI samples are generally infected with URI pathogens during the process of sample collection. To detect volatile organics, one can look to other non-medical industries, such as the

auto and cosmetics industries, to adapt electronic-nose technologies for remote field application. Finally, it might be argued that the holy grail in resource limited field deployable technologies would be the development of a rapid testing system for nucleic acids. The availability of such a system would certainly simplify biomarker discovery efforts in many disease areas. These matters are taken up in more detail in each of the disease reports.

References

1. FDA. (Food and Drug Administration, 2006).
2. Halsey, W. The Future of Systems Biology. Business Insights, 1-157 (2005).
3. Barton, C. L. The Future of Personalized Medicine: The impact of Proteomics on Drug Discovery and Clinical Trial Design. 1-156 (2005).
4. Benjamini & Hochberg. J. Roy. Statist. Soc. B 57, 289 (1995).
5. Personal communication, Eddie Moler, PhD, Novartis Institutes for Biomedical Research.
6. Conrads, T. P., Hood, B. L. & Veenstra, T. D. Sampling and analytical strategies for biomarker discovery using mass spectrometry. Biotechniques 40, 799-805 (2006).
7. Sullivan, F. Opportunities for Proteomics Technology Markets. Frost & Sullivan B638-55, 1-68 (2005).
8. Sciences, N. G. www.nextgensciences.com.
9. Shimadzu. www.ssi.shimadzu.com/products/index.cfm. (2006).
10. Sullivan, F. Biomarkers--Promising Research, Potential Applications and New Developments. D301, 1-132 (2004).
11. Solutions, G.
12. Lee, T. New 4D Fractionation Method Detects ng/mL Proteins, Relevant for Biomarkers. GenomeWeb - ProteoMonitor (2005).
13. Caliper.
14. Garone, L. Protein Electrophoresis: The Benchside View. G&P 6, 26-29 (2006).
15. Gryos. www.gyros.com.
16. Chen, C. S. & Zhu, H. Protein Microarrays. Biotechniques 40, 423 (2006).
17. Somalogic. www.somalogic.com.
18. Dyax. www.dyax.com.
19. Affibody. www.affibody.com.
20. He, M. & Taussig, M. Ribosome display: Cell free protein display technology. Briefings in Functional Genomics and Proteomics 1, 204-212 (2002).
21. Discerna. www.discerna.co.uk. (2006).
22. Milagen.
23. Jendoubi, M. (ed. USPTO) (2003).
24. Jendoubi, M. Milagen, CEO, Personal communication. (2005).
25. Ciphergen. www.ciphergen.com. (2006).
26. Lee, T.-S. in GenomeWeb - ProteoMonitor (2005).
27. Chan, D. Daniel Chan on Proteomics Biomarkers in Singapore and at Ciphergen. GenomeWeb - ProteoMonitor (2005).
28. Soen, Y., Chen, D. S., Kraft, D. L., Davis, M. M. & Brown, P. O. Detection and characterization of cellular immune responses using peptide-MHC microarrays. PLoS Biol 1, E65 (2003).
29. Cameron, T. O., Cohen, G. B., Islam, S. A. & Stern, L. J. Examination of the highly diverse CD4(+) T-cell repertoire directed against an influenza peptide: a step towards TCR proteomics. Immunogenetics 54, 611-20 (2002).
30. Danke, N. A. & Kwok, W. W. HLA class II-restricted CD4+ T cell responses directed against influenza viral antigens postinfluenza vaccination. J Immunol 171, 3163-9 (2003).

31. Chen, D. S. et al. Marked differences in human melanoma antigen-specific T cell responsiveness after vaccination using a functional microarray. PLoS Med 2, e265 (2005).
32. Parra-Lopez, C. et al. MHC and T cell interactions of a universal T-cell epitope from Plasmodium falciparum circumsporozoite protein. J Biol Chem (2006).
33. Danke, N. A., Yang, J., Greenbaum, C. & Kwok, W. W. Comparative study of GAD65-specific CD4+ T cells in healthy and type 1 diabetic subjects. J Autoimmun 25, 303-11 (2005).
34. Reijonen, H., Kwok, W. W. & Nepom, G. T. Detection of CD4+ autoreactive T cells in T1D using HLA class II tetramers. Ann N Y Acad Sci 1005, 82-7 (2003).
35. Schwander, S. K. et al. Pulmonary mononuclear cell responses to antigens of Mycobacterium tuberculosis in healthy household contacts of patients with active tuberculosis and healthy controls from the community. J Immunol 165, 1479-85 (2000).
36. Domon, B. & Aebersold, R. Mass spectrometry and protein analysis. Science 312, 212-7 (2006).
37. Martin, S. (2006).
38. Barton, C. L. The Future of Array Technologies: Impact on drug discovery and market growth in DNA, protein and tissue arrays. Business Insights (2005).
39. Bioinformatics, S. I. f. www.expasy.org.
40. Shaffer, C. in Drug Discovery & Development 22-27 (2005).
41. Rosetta, B. & Research, S.-N. (2006).
42. Chan, E. Integrating Transcriptomics and Proteomics. G&P 6, 20-26 (2006).
43. Weckwerth, W. & Morgenthal, K. Metabolomics: from pattern recognition to biological interpretation. Drug Discov Today 10, 1551-8 (2005).
44. HMP, T. H. M. P.-. (2005).
45. Beecher, C. in Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis (eds. George G. Harrigan, P. C., Chesterfield, MO, USA & Royston Goodacre, U. o. M. I. o. S. a. T. U., UK) (Copyright © 2003 Kluwer Academic Publishers. All rights reserved., 2003).
46. Metabolon. www.metabolon.com The Human Metabolome Defined. (2006).
47. Liszewski, K. Metabolomics Plays Crucial Discovery Role. Genetic Engineering News 26 (2006).
48. Dunn, W. B., Bailey, N. J. & Johnson, H. E. Measuring the metabolome: current analytical technologies. Analyst 130, 606-25 (2005).
49. Genome Alberta / Genome Canada. www.metabolomics.ca. (2007).
50. Canadian Groupo Claims "Unique" Database. Science 315, 583-584 (2007).
51. Goodacre, R., Vaidyanathan, S., Dunn, W. B., Harrigan, G. G. & Kell, D. B. Metabolomics by numbers: acquiring and understanding global metabolite data. Trends Biotechnol 22, 245-52 (2004).
52. Genome Alberta / Genome Canada. www.hmdb.com. (2007).
53. Inc., M. (2006).
54. Biocrates. www.biocrates.com.
55. HMT, H. M. T.-. (2005).
56. McGee, P. in Drug Discovery & Development 26-30 (2006).
57. Protasis. www.protasis.com.
58. NIST. www.nist.gov.
59. Personal communication, Steve Martin, B-G Medicine. (2006).
60. Personal Communication, Elwin Verheij, The Netherlands Organization for Applicd Scientific Research TNO. (2006).
61. MRC-Holland. www.mrc-holland.com.
62. BioScience, C.-. in BioScience Technology #52 14 (2005).
63. ChondroGene & GenomeWeb. in GenomeWeb Daily News Bulletin (2006).
64. Affymetrix. www.affymetrix.com. (206).
65. Smit, M. (ed. Blasband, A.) Discussion of technologies for RNA biomarker discovery other than microarrays (2006).
66. Rachman, H. et al. Unique transcriptome signature of Mycobacterium tuberculosis in pulmonary tuberculosis. Infect Immun 74, 1233-42 (2006).

67. Dinther, J. V. et al. in BioScience Technology 56-58 (2006).
68. Biosystems, A.-A. www.appliedbiosystems.com.
69. Seligmann, B. in G&P 32 (2006).
70. Illumina. www.illumina.com.
71. Amit Kumar, C. CEO CombiMatrix. (2006).
72. Fortina, P. Discussion on RNA biomarker discovery technologies (Philadelphia, 2006).
73. NimbleGen. www.nimblegen.com.
74. ChondroGene. www.chondrogene.com. (2006).
75. Sullivan, F. US DNA Microarray Markets. Frost & Sullivan F656-55 (2006).
76. GenomeWeb. NIH Awards $13.7M for Tuberculosis Structural Biology Project. (2006).
77. Enrico Picozza, C. Immunocyte, Inc. Personal Communication 508-561-6865. (2006).
78. ImmunoCyte, L. Enrico Picozza, CEO 508-561-6865.
79. DePalma, A. in BioScience Technology 14-16 (2006).

