Using modified nucleotides and selecting for slow off-rates in the SELEX procedure, we have evolved a special class of aptamers, called SOMAmers (slow off-rate modified aptamers), which bind tightly and specifically to proteins in body fluids. We use these in a novel assay that yields 1:1 complexes of the SOMAmers with their cognate proteins in body fluids. Measuring the SOMAmer concentrations of the resultant complexes reflects the concentration of the proteins in the fluids. This is simply done by hybridization to complementary sequences on solid supports, but it can also be done by any other DNA quantification technology (including NexGen sequencing). We use measurements of over 1000 proteins in under 100 μL of serum or plasma to answer important medical questions, two of which are reviewed here. A number of bioinformatics methods have guided our discoveries, including principal component analysis. We use various methods to evaluate sample handling procedures in our clinical samples and can identify many parameters that corrupt proteomics analysis.

© 2012 Elsevier Ltd. All rights reserved.
into a robust science that will answer important medical questions.

We do this by using SOMAmers (slow off-rate modified aptamers) as detection agents.1 Before describing this unique class of aptamers, let us review the problem of sensitive and specific protein detection in blood. The most abundant protein in blood is albumin; it is present at a concentration of about 800 μM to 1 mM. Other very abundant proteins (about 250 μM IgG and 50 μM fibrinogen) exist. In contrast, many growth factors and signaling proteins are found in blood in the range of 100 fM to about 800 μM. In other words, an agent that detects a particular most antibodies, are capable of doing such discriminatory binding. This is why the standard antibody test for blood proteins is the ELISA or some variant of it. Here, the problem is solved by using two antibodies to two different epitopes on a protein to gain adequate discriminatory power. As long as one is examining few proteins, the ELISA method is capable of seeing blood proteins in the picomolar range (although finding two antibodies of different and powerful specificity can be challenging). The problem with this approach arises when one tries to apply it to what we consider the modern challenge in proteomics, namely, analyzing thousands of proteins simultaneously, and doing this in a rapid and inexpensive manner. Multiplexed ELISAs are limited to 30 or 40 analytes,1 mainly because the secondary antibodies tend to not have enough specificity in these reactions (especially when they are polyclonals, as most of them are in the commercially available tools).

We have been able to solve the multiplexing problems by using SOMAmers, which are aptamers with two special characteristics.1 They employ at least one modified base in their makeup, and they are specifically doubly selected, once for low $K_d$ values, and (pseudo-)independently for very low $k_{off}$ rates (dissociation half-life $>30$ min). SOMAmers are single-stranded nucleic acids (in the data presented here, they are all DNA) subject to repeated rounds of selection–amplification (SELEX)2 until tight-fitting protein single-stranded DNA complexes are formed.2−4 Because the selection process in each round involves a kinetic challenge with a nonspecific anionic compound, the winning SOMAmers for each selection tend to stay bound to their cognate proteins on kinetic challenge, whereas those “wrong” proteins that were bound during the equilibrium binding steps mostly dissociate after 30 s or less in the presence of the anionic competitor.5

As long as the assay conditions mimic the conditions used in the selection procedure, we are able to get specific binding using one SOMAmer per protein, thus solving the problem introduced by secondary antibodies during multiplexed ELISAs, namely, nonspecific cross reactivity of the secondary antibody.

In Table 1, we show how this use of SOMAmers has dramatically increased our ability to routinely evolve such reagents. We show a list of 12 proteins that repeatedly failed (defined as a $K_d$ greater than $3 \times 10^{-5}$) the SELEX procedure using the four canonical deoxynucleotides in the original library. When 5-benzyl deoxyUTP-containing libraries (as a replacement for the natural thymidyl TP—5-methyl deoxyUTP), 5-isobutyl deoxyUTP-containing libraries, or 5-trytophanyl deoxyUTP-containing libraries are used, these recalcitrant proteins yielded SOMAmers with $K_d$ values low enough to be used as capture reagents.6 We use deoxyuridine-5-carboxamides as starting materials to synthesize libraries containing over 20 different modifications, including hydrophobic residues, amino acids, and known pharmaceutical templates to enrich the binding.

### Table 1. SELEX library affinities ($K_d$, in molar) with unmodified and modified nucleotides

<table>
<thead>
<tr>
<th>Target protein</th>
<th>dT</th>
<th>Benzyl-dU</th>
<th>Isobutyl-dU</th>
<th>Tryptami-no-dU</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-1BBa</td>
<td>Failedb</td>
<td>$6 \times 10^{-9}$</td>
<td>Failed</td>
<td>$4 \times 10^{-9}$</td>
</tr>
<tr>
<td>B7a</td>
<td>Failed</td>
<td>$1 \times 10^{-8}$</td>
<td>Failed</td>
<td>$7 \times 10^{-9}$</td>
</tr>
<tr>
<td>B7-2a</td>
<td>Failed</td>
<td>Failed</td>
<td>Failed</td>
<td>$6 \times 10^{-9}$</td>
</tr>
<tr>
<td>CTLa-4a</td>
<td>Failed</td>
<td>Failed</td>
<td>Failed</td>
<td>$1 \times 10^{-9}$</td>
</tr>
<tr>
<td>sE-Selectina</td>
<td>Failed</td>
<td>Failed</td>
<td>Failed</td>
<td>$2 \times 10^{-9}$</td>
</tr>
<tr>
<td>Fractalkine/CXC3L-1</td>
<td>Failed</td>
<td>Failed</td>
<td>Failed</td>
<td>$5 \times 10^{-11}$</td>
</tr>
<tr>
<td>GA733-1 proteina</td>
<td>$9 \times 10^{-9}$</td>
<td>$3 \times 10^{-9}$</td>
<td>$5 \times 10^{-9}$</td>
<td>$5 \times 10^{-10}$</td>
</tr>
<tr>
<td>gp130, solublea</td>
<td>Failed</td>
<td>$6 \times 10^{-9}$</td>
<td>$2 \times 10^{-8}$</td>
<td>$1 \times 10^{-9}$</td>
</tr>
<tr>
<td>HMG-1</td>
<td>Failed</td>
<td>Failed</td>
<td>$2 \times 10^{-8}$</td>
<td>$5 \times 10^{-9}$</td>
</tr>
<tr>
<td>IR</td>
<td>Failed</td>
<td>$2 \times 10^{-9}$</td>
<td>$1 \times 10^{-8}$</td>
<td>$2 \times 10^{-10}$</td>
</tr>
<tr>
<td>Osteoprotegerina</td>
<td>Failed</td>
<td>$5 \times 10^{-9}$</td>
<td>$9 \times 10^{-9}$</td>
<td>$2 \times 10^{-10}$</td>
</tr>
<tr>
<td>PAI-1</td>
<td>Failed</td>
<td>$4 \times 10^{-10}$</td>
<td>$9 \times 10^{-10}$</td>
<td>$2 \times 10^{-10}$</td>
</tr>
<tr>
<td>P-Cadherina</td>
<td>Failed</td>
<td>$4 \times 10^{-9}$</td>
<td>$5 \times 10^{-9}$</td>
<td>$3 \times 10^{-9}$</td>
</tr>
<tr>
<td>sLeptina</td>
<td>Failed</td>
<td>$2 \times 10^{-9}$</td>
<td>Failed</td>
<td>$5 \times 10^{-10}$</td>
</tr>
</tbody>
</table>

* The protein used was expressed as a fusion to the Fc of human IgG1. No detectable binding of the active library to an alternate Fc fusion protein was observed.

b Pool $K_d$ $>30$ nM.
capacities of these nucleic acid libraries. Both nucleoside triphosphate synthesis for SELEX and phosphoramidite synthesis for chemical synthesis of “winning” sequences are done for each modification. A sampling of these structures is shown in Fig. 1. By using multiple libraries with each protein, we almost always find a SOMAmer that binds tightly and specifically to individual proteins.

Sequence analysis of SOMAmers selected to 850 human protein targets suggests that, at least for the benzyl dUTP library, the modified nucleotides are positively selected during the rounds of SELEX. Using the base composition of our starting (random) libraries and the base composition of our winning sequences, we find a 1.33-fold enrichment for the modified nucleotide. Moreover, many trinucleotide motifs are selected for and against during SELEX with the benzyl dU-containing library. Co-crystal structures of three of these SOMAmer–protein complexes demonstrate why this may be so. Novel benzyl clustering stabilizes the SOMAmer and novel benzyl–nucleotide base and benzyl–aromatic amino acid stacking interactions figure prominently in these structures. The SELEX procedure has been automated (with off-the-shelf robotic components) so that hundreds of proteins can be run each time a SELEX procedure is started.

At present, SomaLogic has SOMAmers to about 1100 human proteins, which have been tested for specificity and are used in our assay. We have an additional 1000 SOMAmers being prepared for the menu, and we plan to have SOMAmers to 5000 proteins by 2014.

The key to using SOMAmers to measure all these proteins simultaneously in under 100 μL of serum or plasma is a resultant of their intrinsic specificity and sensitivity and of our ability to reduce biological background in the samples. Figure 2 shows a cartoon of the structure of each SOMAmer used in the mixture of reagents in solution at the start of our assay. The three-dimensional structure of the SOMAmer is necessary for correct binding to its cognate protein, and the additional functional groups do not disturb this structure. They are as follows (Fig. 2): a fluorophore (F), usually Cy 3, which stays on the SOMAmer (S) until the end of the assay and is what is finally measured in relative fluorescence units; a photocleavable group (PC), o-nitrobenzyl ether, which is necessary for separating two of the steps in the assay procedure; and biotin (B), which is a substrate for streptavidin beads.

Fig. 1. SOMAmer building blocks: dU-5-carboxamides. The chemical structure of CE-phosphoramidites and nucleoside triphosphates are shown on the left, where R represents the various derivatives of the carboxamide substituent on the 5-position of deoxyuridine. The structure of the R groups that we employ are shown on the right—Bn-dU is benzyldeoxyuridine, dT* is methyl-dU (thymidine), iB-dU is isobutyldeoxyuridine, Th-dU is 2-thieno-methyldeoxyuridine, FBn-dU is 4-fluoro-benzyl-deoxyuridine, Pe-dU is 2-phenyl-ethyl-deoxyuridine, Pp-dU is 3-phenyl-2-propyl-deoxyuridine, Tyr-dU is tyrosyl-deoxyuridine, MBn-dU is 3,4-methylene-dioxy-benzyl-deoxyuridine, Nap-dU is 1-naphthydeoxyuridine, 2Nap-dU is 2-naphthydeoxyuridine, Ne-dU is 2-(1-naphthyl)-ethyldeoxyuridine, 2Ne-dU is 2-(2-naphthyl)-ethyldeoxyuridine, Trp-dU is tryptophanyldeoxyuridine, and Bt-dU is 2-(3-benzo[β]thiophenyl)-ethyldeoxyuridine. See Ref. 6 for details.

Fig. 2. Decorated SOMAmers are used in the assay. The SOMAmer structure includes the single-stranded DNA SOMAmer shown in blue. These are usually 80 nucleotides long, but some have been truncated to 40–60 nucleotides. At the 5′ end of each SOMAmer are, covalently linked through an aliphatic linking chain, a fluorophore (F), usually the Cy 3 dye, then an o-nitrobenzyl ether moiety, which is photocleavable UV light (PC), and then a biotin (B), for capture by streptavidin beads.
Biochemical details of all these steps can be found in Ref. 5. SOMAmer (plus hybridization controls) is now carried out. DNA microarray containing probes complementary to each original sample. (6) Quantitative hybridization to a custom accurate measure of how many cognate proteins were in the back into solution. The SOMAmers in solution are now an Elute. The SOMAmers are eluted off their cognate proteins beads are only those bound to their cognate proteins. (5) SOMAmers are washed away. The SOMAmers now on beads by the biotin groups on the protein, and unbound protein complexes are captured onto new streptavidin magnetic plexes back into solution. (4) Catch 2. The SOMAmer out with UV light, releasing the SOMAmer bound proteins. (3) Cleave. The photocleavage step is carried No dissociates the SOMAmers from the protein in each complex, and instead of measuring the protein, one measures the SOMAmer by either hybridization to complementary probes on a chip or by quantitative PCR. The measurement of protein levels has been transformed into a measurement of single-stranded DNA, a much easier and more sensitive task. The results we present below have been derived from the hybridization of SOMAmers to custom Agilent chips with up to 15,000 spots per array, displaying complementary probes to our known SOMAmer sequences.

In Fig. 4, we show buffer dose–response curves for two of our lung cancer markers, EGFR and Endostatin. The limits of detection for these two SOMAmers by this assay are 100 fM and 70 fM, and the response is linear over about 4 logs of dynamic range. Figure 5 shows the lower limit of quantification (LLOQ) for 1033 SOMAmers. This is a cumulative distribution function of LLOQs and shows the median LLOQ to be about 500 fM. Because we use three different dilutions of serum or plasma to do these measurements, we actually can, for each blood sample, span a 7 log dynamic range. Our average coefficient of variation in repeated measurements is about 6%.

This assay protocol has undergone a number of improvements during the last 5 years or so, and we continue to search for ways to further lower backgrounds and increase specificity. Although the present version of the assay benefits from solution kinetics, fixing the SOMAmers onto a solid surface also allows a different version of the assay to work well. In fact, newer methods of allowing proteins and nucleic acids to rapidly interact (pressure cells, microfluidics, etc.) might allow us in the future to put SOMAmers onto a solid support and speed up the all the steps used in the assay. Note that the slow step in the assay is the 17-h hybridization of SOMAmers to their cognate probes on solid supports. Again, recently developed microfluidic methods could...
greatly shorten this time and open up a whole new set of medical applications for our assays.10,11

More recently, we have been able to adapt our assay protocol to analyze tissue homogenates.12 Fresh frozen (within 5–10 min of excision) tissue can be homogenized with a mortar and pestle, including a cocktail of protease inhibitors in the homogenization buffer, and diluted to equal protein concentrations, in order to normalize the results for tissue comparison. It was important to verify that our assay buffers did not allow any potential DNase activity from the tissue proteins to degrade the SOMAmers in the assay. Although our normalization was done for total protein concentration, it is clear that other types of normalization (cell type, housekeeping proteins, etc.) are possible. We shall come back to the tissue results when we discuss the medical applications of the SOMAmer multiplexed assay.

The motivation for developing SOMAmers was to be able to translate the sensitive measurements of thousands of proteins into answers for important medical questions. Biomarker discovery is based on the same intuition that led Willis in the 17th century to taste patients’ urine—namely, that body fluids are in equilibrium with all tissues and that dysfunction will be seen by changes in the composition of body fluids. Blood, which is a liquid organ, is clearly the most accessible body fluid, which is in contact with all other organs. What has become clear to us after more than a decade of examining blood for biomarkers is that they are there, at least for a high percentage of the medical questions for which we have sought answers, but that finding robust algorithms for disease detection is much more difficult than we had first imagined. Why is this? First of all, a detection algorithm is most useful when it is robust. Robustness means that it applies to a large percentage of the human population, across gender, age, and even genetic variation.13 Humans are extremely outbred, especially compared to the laboratory animals in which one usually starts studies in experimental medicine. Moreover, detection of human diseases implies that biomarkers will detect all stages of a disease, a requirement of utility not required in animal models of disease where genetically identical animals are all provoked into developing a disease (this is especially true of cancer studies) simultaneously. Another aspect of biomarker research in humans makes this a complicated affair. Relatively large numbers of samples (again, including variations in gender, age, and genetic makeup) are required in order to obtain statistically valid answers. This means that biomarker discovery ordinarily compares two populations, one with a disease, and a closely matched one, without the disease. However, the variability of protein concentrations is always much greater in a population than it is in one individual measured many times at different intervals.14 It is the latter, easier case that is the ideal situation for biomarker utilization, especially in widespread screening programs for disease—a person’s measurement is compared to previous measurements in that same person, rather than against the average or median of a population.
Nonetheless, these are the relatively easy problems to overcome. Much harder has been the nature of blood itself. Blood contains various kinds of white cells, red cells, platelets, lipids, and the proteins we seek to quantify. Moreover, some of these proteins are clotting substrates and factors, which are poised to initiate a proteolytic clotting cascade as soon as blood is drawn via venipuncture. Drawing samples for either serum or plasma preparation always involves a protocol that, even when followed rigorously (which is rare for practical considerations), leads to sample variability. For example, in serum preparation, blood is allowed to clot at room temperature before centrifuging out the clot and cells (arm to spin time) to give serum, which is then stored at $-80 \, ^\circ\text{C}$ (spin to freeze time). Lysis of white cells can take place during either of these time periods, in the first instance from uncentrifuged blood if one waits too long before centrifugation, and in the second instance if the centrifugation does not spin out most of the white cells.$^{15}$ We have found proteins that are released into blood as a result of white cell lysis and their variable presence can masquerade as biomarkers.$^{16,17}$ A similar situation exists with respect to platelet activation during a blood draw. Platelet activation can occur to varying degrees depending on the bore of the needle used in venipuncture, the speed (shear forces) with which blood is withdrawn, the temperature during centrifugation, and a number of other factors.$^{18}$ Equally important, since platelets are much smaller than cells, is the force generated during centrifugation to ensure that most platelets are removed from the serum. Even when a standard protocol is followed, and a small percentage of platelets have been activated, the amount of protein spilled into serum can still be variable if the initial platelet concentration is variable. An example of cell lysis confounding biomarker analysis is shown in Fig. 6. In this study, we use principal component analysis (PCA)$^{19}$ to identify two vectors, a biology vector and a cell lysis vector represented on the $x$- and $y$-axes in Fig. 6a. We shall discuss the details of PCA below. In this experiment, blood was drawn from healthy volunteers, one post-menopausal female and three males. Samples were allowed to sit at room temperature for either 0.5 h, 1 h, 2 h, 4 h, or 20 h before centrifugation. As shown in Fig. 6a, the results in these vectors depended both on the gender of the volunteer and on the time from arm to spin of the samples. In Fig. 6b, we see that such differences, had they not been identified by PCA, could easily have been misidentified as biomarkers for a disease if the gender/cell lysis parameters were different in the two arms of a control–case study. We have identified a number of components that can potentially confound our search for real biomarkers, and we use them systematically to compare different groups of samples. Again, a robust algorithm for disease detection will weed out these potential confounding factors and eliminate them from our analyses. A case in point is an algorithm developed for lung cancer detection based on samples from four different sites. Here, site-to-site variation comparing control groups to each other or cancer groups to each other (as opposed to measuring control groups to cancer groups) gives a measure of proteins that will not be reliable as legitimate cancer markers (Fig. 7). Introducing the PCA of pre-analytical variables has increased our ability to identify and eliminate potential false markers.

It is evident that rigorous bioinformatics analysis of clinical data and protein quantification is necessary to exploit protein measurements and transform...
Fig. 6. Multidimensional vectors of real biology and pre-analytic variables. (a) Plots of the first two principal components of a study composed of blood drawn from four healthy volunteers: three males and one post-menopausal female. Samples were allowed to clot at room temperature for either 0.5 h, 1 h, 2 h, 4 h, or 20 h. The 80 plasma samples (20 per person) were run on our 650-plex proteomics assay. PCA revealed a biology vector on principal component 1 (x-axis) and a cell lysis vector on principal component 2 (y-axis). The left plot shows the values of the samples projected into the first two principal components. The samples are colored by the clotting time: 0.5 h, purple; 1 h, blue; 2 h, green; 4 h, orange; or 20 h, red. The right plot shows the values of the PCA loadings (proteins). The clustering of the samples reveals two distinct patterns of variation. On the first principal component (x-axis), the samples cluster into four columns, corresponding to the four individuals in the study. Proteins with high loadings in the plot on the right include FSH and LH, which are gender hormones known to be different between males and females. The second component reveals a pattern associated with the experimental condition of clotting time. Proteins with high loadings in the plot on the right include a set of intracellular neutrophil proteins that increase concentration in blood in response to cell lysis. It is clear that the results from this analysis depend both on the biology vector (male versus female) and on the cell lysis vector (time of draw to centrifugation). (b) The same components are represented on the x- and y-axes. If the cases, say, cancer patients, had blood draws that allowed blood to sit more than an hour or so at room temperature before centrifugation, and the benign control patients had blood draws that centrifuged samples with a short wait time, the resulting change on the y-axis would have mistakenly identified the protein(s) being measured as cancer biomarkers.
them into answers to medical questions. We have used multiple approaches. In addition to controlling for pre-analytic variability introduced by differences in sample handling, there are other biological confounding factors that must be considered. For example, renal clearance efficiency decreases with age, which results in an increase in protein concentration. This age-dependent bias must be corrected for to prevent misidentification of disease biomarkers. These types of biases are removed by using linear and nonlinear models to remove the variation associated with aging from the protein measurements that are affected by decreased renal function. With sources of confounding variation controlled for or removed in a clinical data set, we perform biomarker discovery using a combination of univariate and multivariate techniques. We begin by performing Kolmogorov–Smirnov tests and selecting a set of significant markers after controlling for multiple comparisons using false discovery rate correction. To gain a better understanding of the differential expression defined by the set of significant markers, we perform PCA and attempt to separate them into clusters. Finally, to create a diagnostic model, we perform backward elimination using the random forest (RF) classifier. We begin by building a model with the entire set of significant biomarkers and assess how well the markers performed using the Gini importance calculated by the model. We remove the least important marker, build a new model, and repeat the process until only a single marker remains. To avoid overfitting, we try to select a small panel of biomarkers that achieves comparable performance to the entire set of significant markers. A smaller panel of markers is also more practical for the development of a diagnostic test.

We have successfully derived useful algorithms from SOMAmer measurements to answer important medical questions. The following are among the questions for which we have at least discovery and verification studies: Does this heavy smoker have lung cancer? Does this person exposed to asbestos over his or her lifetime have mesothelioma? Does this person at risk for pancreatic cancer have the disease? Is this person who has already had evidence

---

**Fig. 7.** Identical collection protocol at four different sites. Differences between sample collection sites that are likely due to sample collection differences. SMV represents Sample Mapping Vector as determined by protein measurements known to be associated with, in these cases, either neutrophil lysis or complement activation. The boxplots show the difference between lung cancer and healthy controls stratified by the four collection sites in the study. The left plot shows an intracellular neutrophil protein that is elevated in serum and plasma in the presence of cell lysis. The right plot shows a complement protein that increases in the presence of complement activation. This demonstrates how differential expression introduced by differences in sample collection protocol and compliance can be inadvertently identified as disease markers.

**Fig. 8.** ROC curve of the SomaLogic algorithm for NSCLC. After deriving an algorithm from our first discovery study, we recalculated the algorithm, removing case-control bias using our PCA vector analysis, and expanded the menu to 1000 protein measurements in a blinded verification study. The ROC curves plot on the x-axis 1-specificity, which is the false-positive rate, against, on the y-axis, sensitivity, which is the true-positive rate. The AUC, area under the curve, a measure of total accuracy, is 0.81 for stages I and II NSCLC and 0.89 for stages III and IV NSCLC.
of serious cardiovascular disease likely to have a second serious event in the 6 years to come? What pathological stage is the tumor in this person with renal cell carcinoma? Is this person with Alzheimer’s disease likely to deteriorate rapidly or slowly? There are many other results at various stages of the discovery process. There are also a few questions for which we have not yet detected answers. It is our hope that as we add more and more SOMAmers to the mix and continue to improve our measurements in the assay, these medical questions can also be answered with biomarkers.

We shall present two examples of how SOMAmers are used to answer medical questions, both in oncology. One is a blood test for detection of lung cancer in heavy smokers and the other is the detection of mesothelioma in asbestos-exposed individuals. Lung cancer is the leading cause of cancer...
deaths, because 84% of cases are diagnosed at an advanced stage. Worldwide, in 2008, 1.5 million people were diagnosed and 1.3 million died—a survival rate unchanged since 1960. However, patients who are diagnosed at an early stage and have surgery experience an 86% overall 5-year survival. New diagnostics are therefore needed to identify early-stage lung cancer. We have examined the serum of 1326 patients, 291 of whom had non-small cell lung cancer (NSCLC) (the type that accounts for >80% of all lung cancer). This population was compared to 565 patients known by computed tomography to have benign lung nodules and 470 patients who had similar smoking histories to the other two groups (about half of this group would be assumed to have benign nodules had a computed tomography scan been done). These samples were collected from four independent sites, and one of the first results we obtained was that there were strong site-to-site differences (Fig. 7 and Ref. 16). The aforementioned analysis of pre-analytic variables shows clearly that sample handling differed among the four sites and that case–control sample handling differences were seen in at least one site; these could easily have been misinterpreted as cancer markers (Fig. 6 and Ref. 16).

The analysis, both by the Naïve Bayes (NB) approach using Kolmogorov–Smirnov statistics and by the multidimensional approach using PCA, eliminated the false markers when decision algorithms were generated (Figs. 6 and 7). The diagnostic accuracy of this blood test for NSCLC is shown as a receiver operating characteristic (ROC) curve measuring sensitivity and specificity in Fig. 8.

As we previously mentioned, we have extended our analyses to tissue samples, and we have published encouraging preliminary results on lung cancer tissue to complement our blood-based assays. Analysis of proteins in NSCLC tissue was compared to protein levels in both normal tissue adjacent to the surgical margins and normal tissue distant from the extirpated cancer. Figure 9a shows no significant differences in proteins in adjacent normal tissue compared to distant normal tissue. In other words, we see no evidence of a cancer “field effect,” at least for the 820 human proteins analyzed here. In contrast, we find many changes in protein levels when we compare tumor tissue homogenates to either adjacent or distal normal lung homogenates (Fig. 9a and b). We identified 11 proteins with a more than fourfold difference between cancer and controls, and 53 proteins with a greater than twofold difference. Many of these proteins have been implicated in cell–cell interactions as well as cell–matrix interactions. Among these were MMP-7 and MMP-12, both of which were blood biomarkers of NSCLC. We are presently obtaining more tissue samples to match our blood analyses in all of our oncology studies.

Fig. 10. ROC curve of the SomaLogic algorithm for malignant pleural mesothelioma (MPM). ROC curves are explained in the legend to Fig. 8. Here, sensitivity and specificity of diagnosis of MPM as compared to patients equally exposed to asbestos but who do not have MPM are plotted on the curve. Here, the algorithm comprised 13 proteins derived from a discovery set, which included 60 patients with MPM and 60 patients who were asbestos exposed. The AUC is greater than 99%. A blinded set of 19 patients with MPM and 20 asbestos-exposed patients also gave an AUC greater than 99%.

The final example is the detection of mesothelioma (a malignant cancer of the pleural linings) in patients with a history of extensive exposure to asbestos, ordinarily through jobs as pipe coverers, miners, shipyard workers, among others. The incidence of mesothelioma is relatively low, about 3000 new cases per year in the United States, but the number of potential cases is quite high. It is estimated that almost 30 million people in the United States have been occupationally exposed to high levels of asbestos. Our study compared sera from mesothelioma patients with sera from asbestos-exposed patients without malignant disease. Again, we divided samples into discovery and verification groups from disease and control groups, and measured protein levels for 820 proteins. Bioinformatics analysis was performed both by NB algorithm construction and by RF algorithm construction. Many mesothelioma markers were common to both types of analysis. These included proteins that were up-regulated in mesothelioma and a smaller number that were down-regulated in this cancer. Measuring both sensitivity and specificity in the NB and RF approaches gave very high accuracy of detection. In the NB approach (6-marker algorithm) the area under the curve for the ROC analysis was over 0.95. In the RF approach (13-marker algorithm), the area under the curve was greater than 0.99 (Fig. 10). Some of the biomarkers for mesothelioma were known to be implicated in cell growth or in cell–cell or cell–matrix interactions.23
In Fig. 11, we present a summary of 10 different medical or technological questions for which we have been able to provide verified biomarker answers. We emphasize that these represent only 10 of perhaps 60 or 70 questions that we have been able to give answers in terms of biomarkers. The 10 were selected to demonstrate the power of this methodology. Color coding indicates whether the questions were in oncology (blood), red bars; oncology (tissue), turquoise bars; cardiovascular risk, blue bars; and sample handling questions, green bars. It is clear that some markers are the same when different types of cancer are analyzed (the example of protein C9 is indicated), yet the total pattern of biomarkers is distinct for each type of protein (different markers for NSCLC and mesothelioma are indicated). Some cardiovascular risk markers (mainly in inflammation proteins) are also cancer markers, but a whole set of unique proteins specify cardiovascular risk. Finally, we know when to suspect a false marker when one of the proteins represented by the green bars appears to be a marker in control–case studies and will merit special analysis to see whether it really contributes to the disease-specific set of protein biomarkers. We think this constitutes the beginning of a new, robust way to answer important medical questions using our new proteomic approach.

References


