doi:10.1016/j.jmb.2012.06.021 J. Mol. Biol. (2012) 422, 595–606

Contents lists available at www.sciencedirect.com

Journal of Molecular Biology

journal homepage: http://ees.elsevier.com.jmb

PERSPECTIVE

Life's Simple Measures: Unlocking the Proteome

Edward Brody*, Larry Gold, Mike Mehan, Rachel Ostroff, John Rohloff, Jeff Walker and Dom Zichi

SomaLogic, Inc., 2945 Wilderness Place, Boulder, CO 80301, USA

Received 5 June 2012;
accepted 12 June 2012
Available online
19 June 2012

Edited by M. Yaniv

Keywords:
SOMAmers;
aptamers;
proteomics;
cancer;
diagnostics

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.

Examining body fluids as an aid to medical diagnostics is at least as old as Western civilization. Hippocrates spilled patients' urine on the ground to see whether or not it attracted insects. In the 17th century, the English physician Thomas Willis distinguished diabetes mellitus (a disease of the pancreas) from diabetes insipidus (a disease of the pituitary) by determining whether the patients' urine tasted sweet (mellitus) or bland (insipidus). Blood biomarker use grew rapidly in the middle to late 20th century, with varying degrees of success. As the use of biomarkers progressed in medical practice, so did the scientific criteria for their acceptance and reimbursement. Today, when the DNA sequence of each person's genome may soon be within technical and economic reach, it is important to examine carefully the modern role of biomarkers in blood, urine, cerebrospinal fluid, and so on, as a robust means to answer important diagnostic questions. Genomics analysis will be a great leap forward in determining an individual's likelihood of contracting a particular malady, and to some extent, this is already happening. Women carriers of either BRCA1 or BRCA2 mutations have an approximately 10-fold greater risk than the carriers of the wild-type allele of having breast or ovarian cancer. Knowing this can be useful, but it also leads to drastic procedures of avoidance, such as prophylactic bilateral mastectomy and oophorectomy. Clearly, what is missing in the genetic information is the knowledge of the onset of the disease, so that appropriate, perhaps less drastic, therapy can be employed early enough to affect cures. Measuring proteins in blood has been useful but limited in answering medical questions about disease onset, and it is our goal to turn the quantification of most human proteins in blood

*Corresponding author. E-mail address: ebrody@somalogic.com.

Abbreviations used: SOMAmer, slow off-rate modified aptamer; LLOQ, lower limit of quantification; PCA, principal component analysis; NSCLC, non-small cell lung cancer; ROC, receiver operating characteristic; NB, Naïve Bayes; RF, random forest.

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

Target protein	dT	Benzyl-dU	Isobutyl-dU	Tryptamino-dU
4-1BB[a]	Failed[b]	6×10^{-9}	Failed	4×10^{-9}
B7[a]	Failed	1×10^{-8}	Failed	7×10^{-9}
B7-2[a]	Failed	Failed	Failed	6×10^{-9}
CTLA-4[a]	Failed	Failed	Failed	1×10^{-9}
sE-Selectin[a]	Failed	Failed	Failed	2×10^{-9}
Fractalkine/CXC3L-1	Failed	Failed	Failed	5×10^{-11}
GA733-1 protein[a]	9×10^{-9}	3×10^{-9}	5×10^{-9}	5×10^{-10}
gp130, soluble[a]	Failed	6×10^{-9}	2×10^{-8}	1×10^{-9}
HMG-1	Failed	Failed	2×10^{-8}	5×10^{-9}
IR	Failed	2×10^{-9}	1×10^{-8}	2×10^{-10}
Osteoprotegrin[a]	Failed	5×10^{-9}	9×10^{-9}	2×10^{-10}
PAI-1	Failed	4×10^{-10}	9×10^{-10}	2×10^{-10}
P-Cadherin[a]	Failed	4×10^{-9}	5×10^{-9}	3×10^{-9}
sLeptin R[a]	Failed	2×10^{-9}	Failed	5×10^{-10}

[a] 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.

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 500 pM. In other words, an agent that detects a growth factor at 1 pM in blood has to do so in a 10^9 excess of albumin.[1] Very few binding reagents, in 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 3×10^{-8}) 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 dexyUTP-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

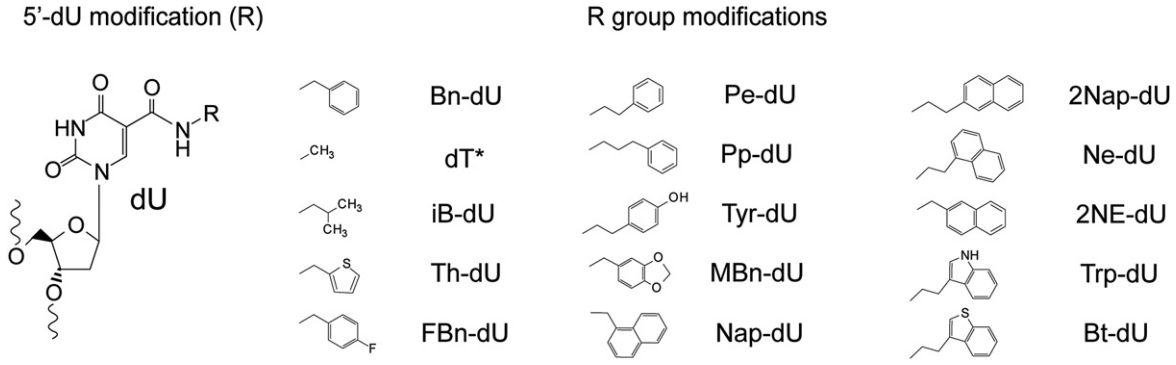

Fig. 1. SOMAmer building blocks:dU-5-carboxamides. The chemical structure of CE-phosphoramidites and nucleoside triphosphates are shown on the left, where R represents the various derivatives of the carboxamide substituent on the 5-position of deoxyuridine. The structure of the R groups that we employ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

capacities of these nucleic acid libraries.[6] 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.[6]

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 tri- and tetranucleotide 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.[7] 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.[8]

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.[9] 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.[9] 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 in the

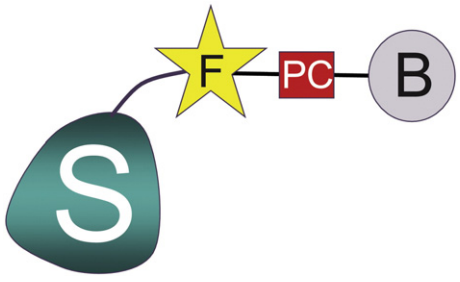

Fig. 2. Decorated SOMAmers are used in the assay. The SOMAmer structure includes the single-stranded DNA SOMAmer shown in blue. These are usually 80 nucleotides long, but some have been truncated to 40–60 nucleotides. At the 5′ end of each SOMAmer are, covalently linked through an aliphatic linking chain, a fluorophore (F), usually the Cy3 dye, then an *o*-nitrobenzyl ether moiety, which is photocleavable UV light (PC), and then a biotin (B), for capture by streptavidin beads.

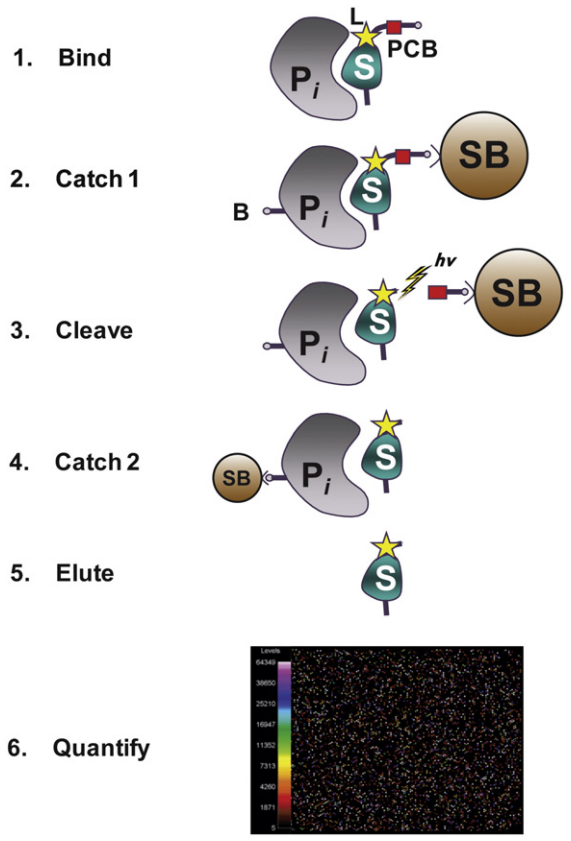

Fig. 3. Details of the SomaLogic proteomics assay. (1) Bind. The mix of up to 1100 SOMAmers (at present; moreover, we have no reason to believe this will not scale up to the 5000 SOMAmers we anticipate will be in our assay by 2014) is added to the sample (serum. plasma, cerebrospinal fluid, tissue lysate, etc.). Each SOMAmer (S) carrying the moieties indicated in Fig. 2 (F, fluorophore, represented by the yellow star; L, photocleavable group, represented by the red box; B, biotin, represented by the small white circle). Each SOMAmer binds primarily to its cognate protein (Pi) to the extent limited by its equilibrium K_d. (2) Catch1. Bound protein–SOMAmer complexes are captured onto streptavidin-coated beads (SB) by the photocleavable biotin at the 5′ end of the SOMAmer. Unbound proteins are washed away, and the kinetic challenge with a nonspecific anionic competitor is carried out during the washes to release noncognate proteins that had been bound at the equilibrium binding step. Protein biotinylation is carried out on the bound proteins. (3) Cleave. The photocleavage step is carried out with UV light, releasing the SOMAmer–protein complexes back into solution. (4) Catch 2. The SOMAmer–protein complexes are captured onto new streptavidin magnetic beads by the biotin groups on the protein, and unbound SOMAmers are washed away. The SOMAmers now on beads are only those bound to their cognate proteins. (5) Elute. The SOMAmers are eluted off their cognate proteins back into solution. The SOMAmers in solution are now an accurate measure of how many cognate proteins were in the original sample. (6) Quantitative hybridization to a custom DNA microarray containing probes complementary to each SOMAmer (plus hybridization controls) is now carried out. Biochemical details of all these steps can be found in Ref. 5.

separation of two steps in the assay. Figure 3 describes how the SOMAmers are used to measure cognate proteins. In solution, the diluted serum or plasma is incubated with the mix of up to 1100 SOMAmers. Next, the protein–SOMAmer complexes are bound through the biotin to streptavidin beads. The unbound proteins are washed away. Then, the bound proteins on the complexes are biotinylated, after which the kinetic challenge is begun by diluting into a polyanionic competitor. Next, the protein–SOMAmer complexes are released back into solution by photocleavage of the *o*-nitrobenzyl ether with UV light. At this stage, the protein–SOMAmer complexes are captured with streptavidin beads by binding to the biotinylated proteins in each complex. The free SOMAmers are washed away. As a result, we are left with beads bound to a 1-to-1 complex of protein and SOMAmer. Now, one 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.[5] 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

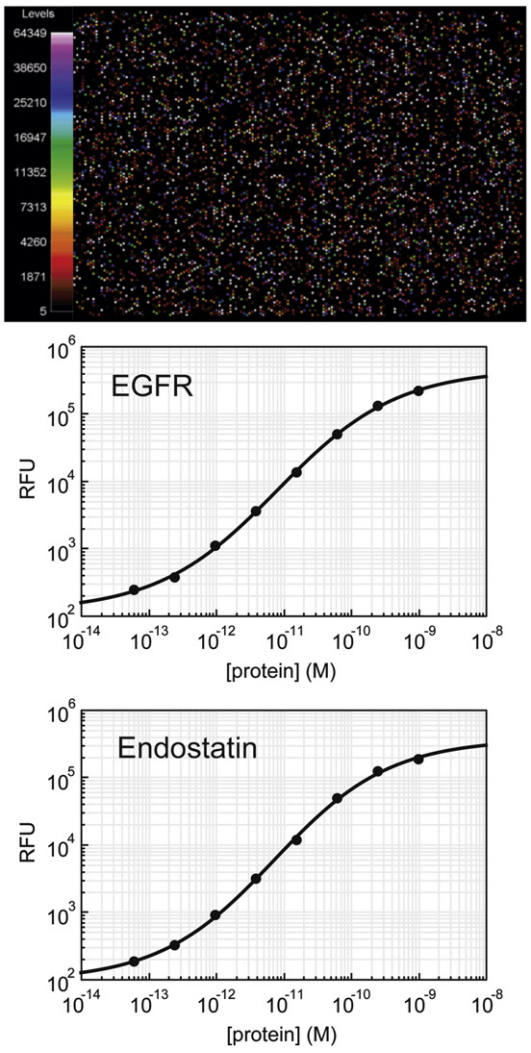

Fig. 4. SOMAmers quantified by hybridization. (a) An example of the Agilent custom array reflecting hybridization of SOMAmers to their complementary probes. There are 15,000 spots per array, and we use 10 spots per SOMAmer to get statistically robust measurements of SOMAmer concentrations. (b) A dose–response curve of EGFR, one of our biomarkers for lung cancer, in buffer. The *x*-axis represents the molar concentration of this protein and the *y*-axis represents the SOMAmer concentration measured in relative fluorescence units (RFUs). The limit of detection is 100 fM, and the measurement is linear over a 3.8 logarithm dynamic range. (c) A dose–response curve for Endostatin, another of our lung cancer biomarkers. Measurement is as in (b). The limit of detection is 70 fM, and the dynamic range is 4.1 logarithms.

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.

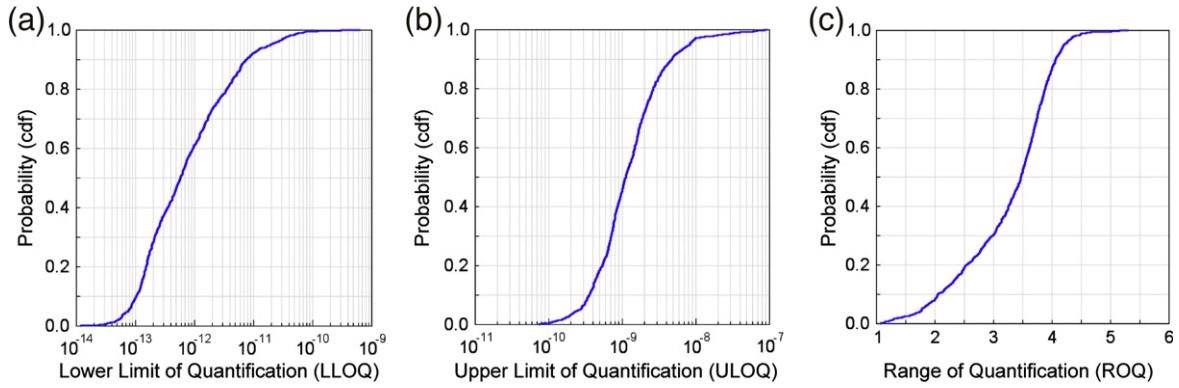

Fig. 5. SOMAmer quantification range. LLOQ measurements for 1033 SOMAmers. (a) LLOQ is defined by the lowest concentration of protein that still gives under a 20% coefficient of variation. See Ref. 5 for details of these calculations. The LLOQs (x-axis) are plotted as a probability density function (y-axis) for 1033 SOMAmers. The median LLOQ is 0.5 pM. (b) Upper limit of quantification (ULOQ) measurements for 1033 SOMAmers. The ULOQ is defined by the highest concentration of protein that still gives under a 20% coefficient of variation. See Ref. 5 for details of these calculations. The median ULOQ for 1033 SOMAmers is 1.5 nM. (c). Median quantification range. The quantification range for each SOMAmer is the ULOQ minus the LLOQ. It is measured as the logarithm of this difference. The median quantification range is 3.5 logs.

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 °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

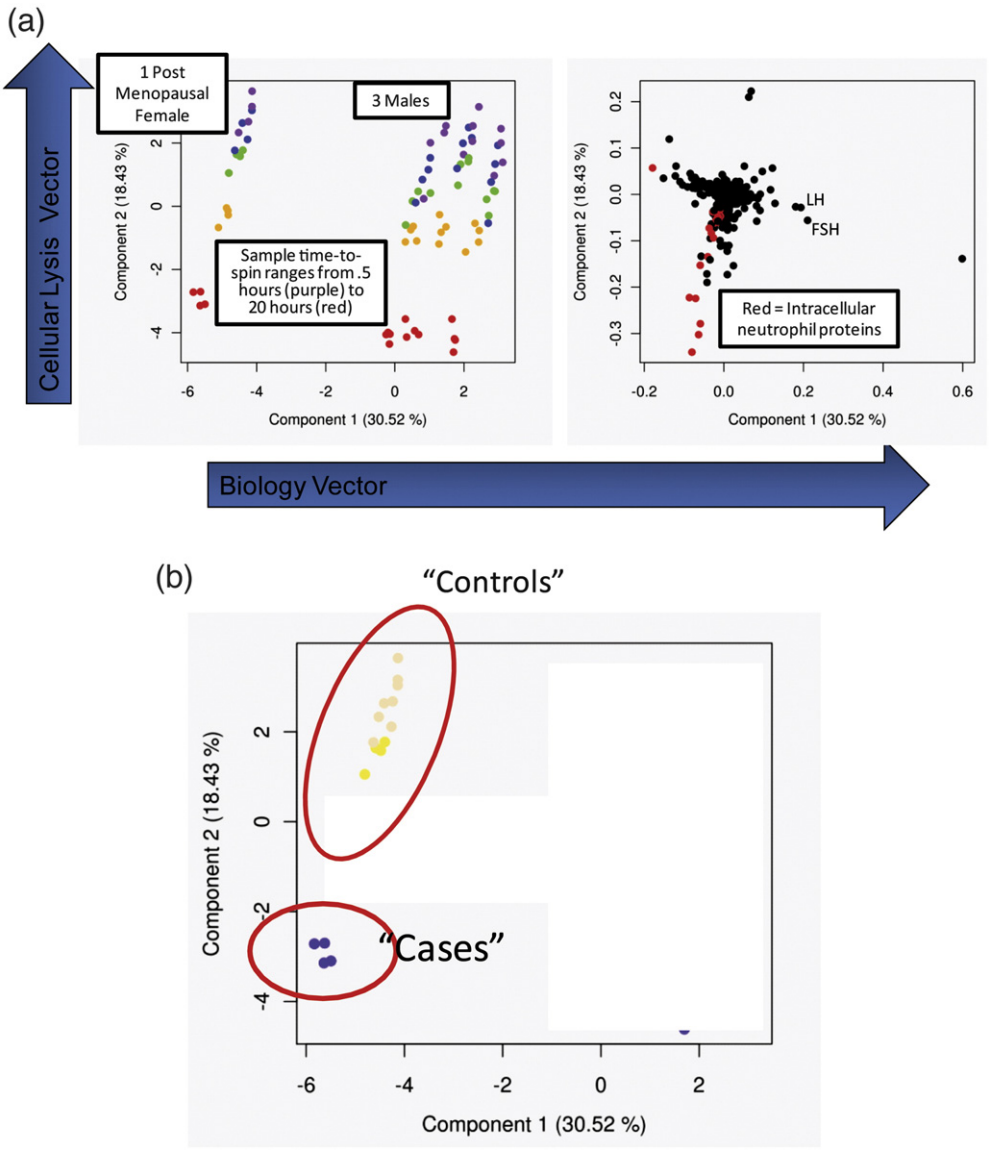

Fig. 6. Multidimensional vectors of real biology and pre-analytic variables. (a) Plots of the first two principal components of a study composed of blood drawn from four healthy volunteers: three males and one post-menopausal female. Samples were allowed to clot at room temperature for either 0.5 h, 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.

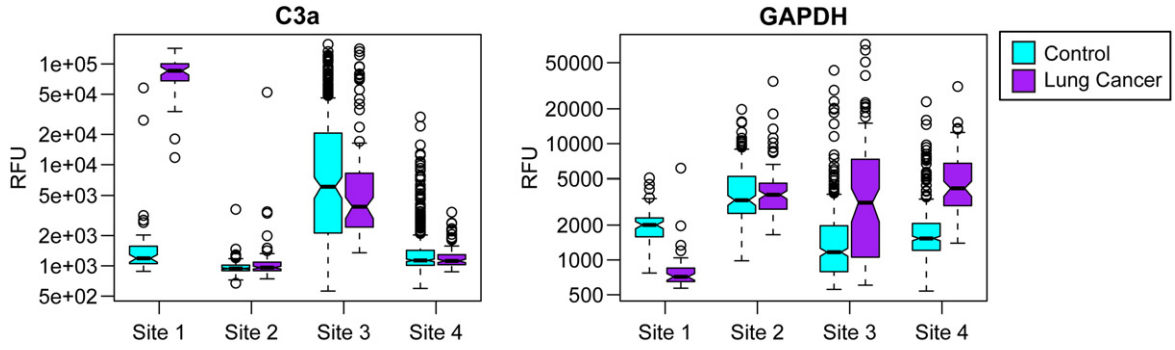

Fig. 7. Identical collection protocol at four different sites. Differences between sample collection sites that are likely due to sample collection differences. SMV represents Sample Mapping Vector as determined by protein measurements known to be associated with, in these cases, either neutrophil lysis 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.

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.[19] 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.[20] 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

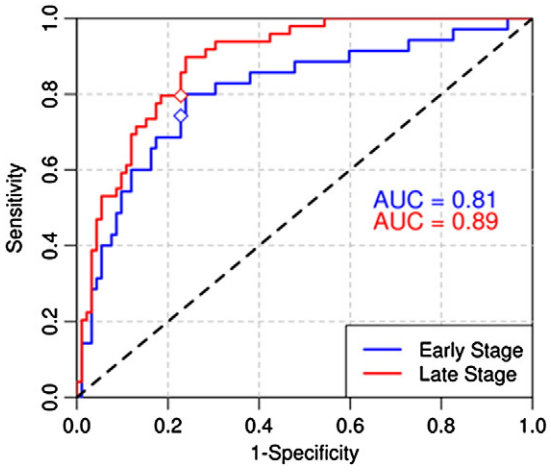

Fig. 8. ROC curve of the SomaLogic algorithm for NSCLC. After deriving an algorithm from our first discovery study,[16] we recalculated the algorithm, removing case–control bias using our PCA vector analysis, and expanded the menu to 1000 protein measurements in a blinded verification study. The ROC 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.

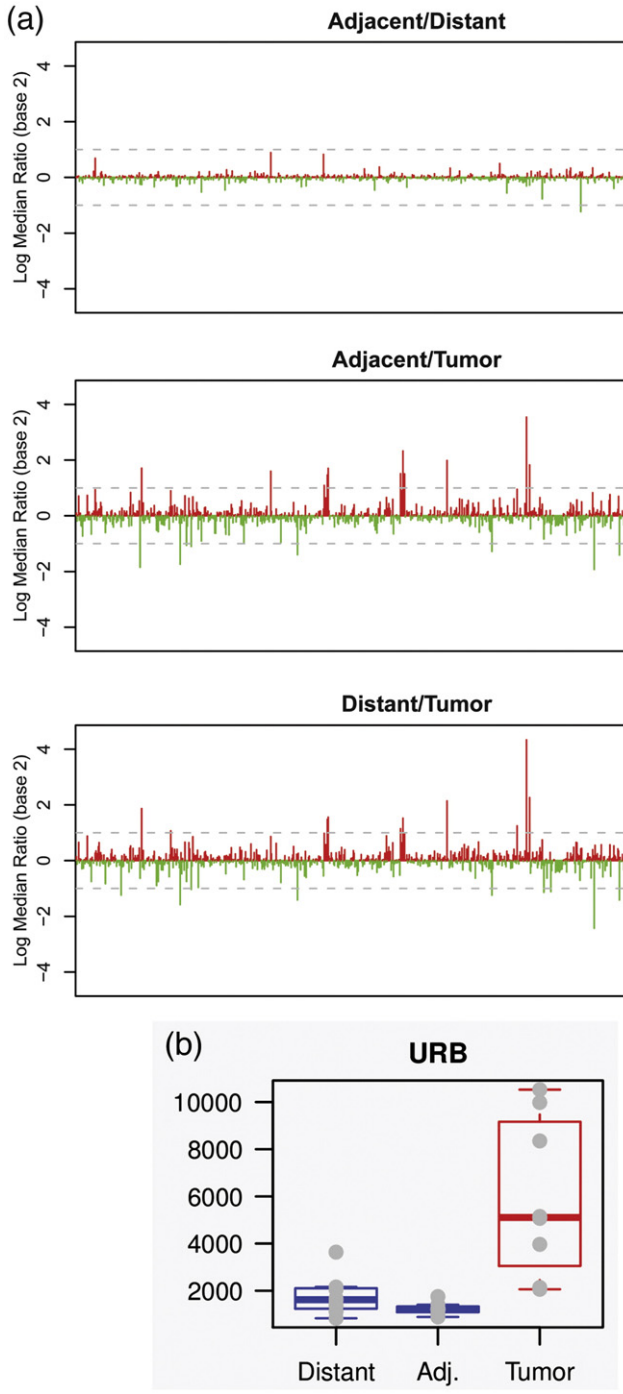

Fig. 9. Protein expression changes in lung tumor tissue. (a) Lung tumor tissue from eight resected tumors of NSCLC were frozen within 5–10 min of resection. Samples were taken from within each cancer, from adjacent healthy tissue (within 1 cm), and from distal healthy lung tissue. Homogenization was done as described in Ref. 12, and the sample was analyzed for 813 proteins in our standard assay.[5,12] Normalization was done to total protein concentration in each sample. Relative protein concentrations for these proteins are displayed for distant/adjacent normal lung tissue (a), tumor/adjacent normal tissue (b), and tumor/distal normal tissue (c). The *x*-axis represents one bar per protein. The *y*-axis is the log 2 median ratios. The dotted lines represent twofold change. Red bars represent up-regulation and green bars represent down-regulation. There are 36 proteins with consistent twofold or greater differences with either proximal or distal normal lung tissue. Twenty of these are up-regulated in the tumor and 16 are down-regulated. (b) An example of a lung cancer tissue marker. In the assays cited above, an example (URB) is shown distributed across the eight samples analyzed. The protein levels of each of the eight tumors (measured in relative fluorescence units) is shown on the *y*-axis in red, with each symbol representing a different patient. The same symbols are used in blue for the adjacent healthy tissue and in green for the distal healthy tissue. Horizontal black bars represent median values for the three sets.

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.[13] Worldwide, in 2008, 1.5 million people were diagnosed and 1.3 million died[14]—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.[13,15] 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).[16] 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.[12] 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",[18] 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.[16] We are presently obtaining more tissue samples to match our blood analyses in all of our oncology studies.

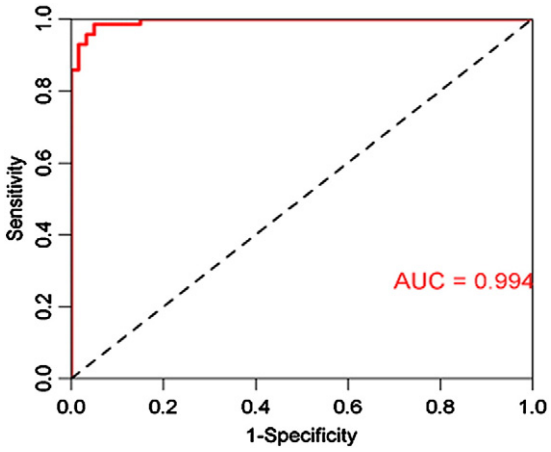

Fig. 10. ROC curve of the SomaLogic algorithm for malignant pleural mesothelioma (MPM). ROC curves are explained in the legend to Fig. 8. Here, sensitivity and specificity of diagnosis of MPM as compared to patients equally exposed to asbestos but who do not have MPM are plotted on the curve. Here, the 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.[20] 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.[21] 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[16] 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.[22]

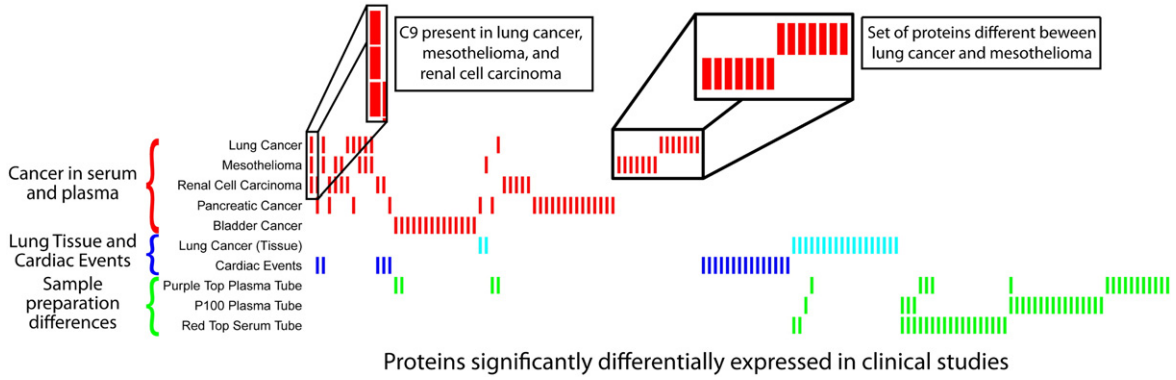

Fig. 11. Summary of biomarkers identified through discovery efforts. A visual depiction of the proteins on the SomaLogic menu that were differentially expressed in various clinical studies. Each column corresponds to a protein and the presence of a rectangle indicates that it was significantly differentially expressed after false discovery rate correction with a p value of 0.01. The number of significant proteins per study was limited to 20 to balance differences in the power of the studies. The figure is split into three main sections, each with a different color. The red rectangles correspond to cancer studies performed in serum and plasma. The blue rectangles correspond to a study of lung cancer tissue homogenate compared to healthy adjacent tissue as well as a study of cardiovascular events. The green rectangles correspond to a series of studies on the effect of different sample collection protocols. The time from draw to spin and freeze were varied in three tube types to determine which proteins were sensitive to sample preparation.

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 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

1. Brody, E. N., Gold, L., Lawn, R. M., Walker, J. J. & Zichi, D. (2010). High-content affinity-based proteomics unlocking protein biomarker discovery. *Expert Rev. Mol. Diagn.* **10**, 1013–1022.
2. Tuerk, C. & Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, **249**, 505–510.
3. Ellington, A. D. & Szostak, J. W. (1990). In vitro selection of RNA molecules that bind specific ligands. *Nature*, **346**, 818–822.
4. Gold, L. (1995). Oligonucleotides as research, diagnostic, and therapeutic agents. *J. Biol. Chem.* **270**, 13581–13584.
5. Gold, L., Ayers, D., Bertino, J., Bock, C., Bock, A., Brody, E. N. *et al.* (2010). Aptamer based multiplexed proteomic technology for biomarker discovery. *PLoS One*, **5**, e15004.
6. Vaught, J. D., Bock, C., Carter, J., Fitzwater, T., Otis, M., Schneider, D. *et al.* (2010). Expanding the chemistry of DNA for in vitro selection. *JACS*, **132**, 4141–4151.
7. Mehan, M. R., Ostroff, R., Wilcox, S. K., Steele, F., Schneider, D. *et al.* (2012). Highly multiplexed proteomic platform for biomarker discovery, diagnostics and therapeutics. In *Complement Therapeutics Special Issue* (Lambris, J., Rickelin, D. & Holers, M., eds), Springer, New York City; in press.
8. Keeney, T. R., Bock, C., Gold, L., Kraemer, S., Lollo, B., Nikrad, M. *et al.* (2009). Automation of the SomaLogic proteomics assay: a platform for biomarker discovery. *J. Assoc. Lab. Automation*, **14**, 360–366.
9. Kraemer, S., Vaught, J. D., Bock, C., Gold, L., Katilius, E., Keeney, T. R. *et al.* (2011). From SOMAmer-based biomarker discovery to diagnostic and clinical applications: a SOMAmer-based, streamlined multiplex proteomic assay. *PLoS One, e26332.*
10. Peytavi, R., Raymond, F. R., Gagne, D., Picard, F. J., Jia, G., Zoval, J. *et al.* (2005). Microfluidic device for rapid (<15 min) automated microarray hybridization. *Clin. Chem.* **51**, 1836–1844.

11. Liu, J., Williams, B. A., Gwirtz, R. M., Wold, B. W. & Quake, S. (2006). Enhanced signals and fast nucleic acid hybridization by microfluidic chaotic mixing. *Angew. Chem. Int. Ed.* **45**, 3618–3623.
12. Mehan, M. R., Ayers, D., Thirstrup, D., Xiong, W., Ostroff, R., Brody, E. *et al.* (2012). Protein signature of lung cancer tissues. *PLoS One, e35157.*
13. Jemal, A., Siegal, R., Ward, E., Hao, Y., Xu, J. & Thun, M. J. (2009). Cancer Statistics, 2009. *CA Cancer J. Clin.* **59**, 22–249.
14. World Cancer Report (2008). Boyle, P. & Levin, B., eds. International Agency for Research on Cancer (IARC), Lyon.
15. Okada, M., Nishio, W., Sakamoto, T., Uchino, K., Yuki, T., Nakagawa, A. & Tsubota, N. (2005). Effect of tumor size on prognosis in patients with non-small cell lung cancer: the role of segmentectomy as a type of lesser resection. *J. Thorac. Cardiovasc. Surg.* **129**, 87–93.
16. Ostroff, R. M., Bigbee, W. L., Franklin, W., Gold, L., Mehan, M., Miller, Y. E. *et al.* (2010). Unlocking biomarker discovery: large scale application of aptamer proteomic technology for early detection of lung cancer. *PLoS One*, **5**, e15003.
17. Ostroff, R., Foreman, T., Keeney, T. R., Stratford, S., Walker, J. J. & Zichi, D. (2010). The stability of the circulating human proteins to variations in sample collection and handling procedures measured with an aptamer-based proteomics array. *J. Proteomics*, **73**, 649–666.
18. Stearman, R. S., Dwyer-Nield, L., Grady, M. C., Malkinson, A. M. & Geraci, M. W. (2008). A macrophage gene expression signature defines a field effect in the lung tumor microenvironment. *Cancer Res.* **68**, 34–43.
19. Storey, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc., B*, **64**, 479–498.
20. Breiman, L. (2001). Random forests. *Machine Learning*, **45**, 5–32.
21. Wehrens, R. (2011). *Chemometrics with R.* pp. 43–65, Springer-Verlag, Berlin Heidelberg.
22. Pass, H. I. & Carbone, M. (2009). Current status of screening for malignant pleural mesothelioma. *Thorac. Cardiovasc. Surg.* **21**, 97–104.