Published in final edited form as:

Pediatr Transplant. 2008 November ; 12(7): 737–747. doi:10.1111/j.1399-3046.2008.01018.x.

The Proteogenomic Path towards Biomarker Discovery

Tara K Sigdel, PhD and Minnie M. Sarwal, MD, MRCP, PhD

Department of Pediatrics - Nephrology Stanford University Medical School Stanford University, Stanford, 94305

Abstract

The desire for biomarkers for diagnosis and prognosis of diseases has never been greater. With the availability of genome data and an increased availability of proteome data, the discovery of biomarkers has become increasingly feasible. However, the task is daunting and requires collaborations among researchers working in the fields of transplantation, immunology, genetics, molecular biology, biostatistics, and bioinformatics. With the advancement of high throughput *omic* techniques such as genomics and proteomics (collectively known as *proteogenomics*), efforts have been made to develop diagnostic tools from new and to-be discovered biomarkers. Yet biomarker validation, particularly in organ transplantation, remains challenging because of the lack of a true gold standard for diagnostic categories and analytical bottlenecks that face high-throughput data deconvolution. Even though microarray technique is relatively mature, proteomics is still growing with regards to data normalization and analysis methods. Study design, sample selection, and rigorous data analysis are the critical issues for biomarker discovery using high-throughput proteogenomic technologies that combine the use and strengths of both genomics and proteomics. In this review, we look into the current status and latest developments in the field of biomarker discovery using genomics and proteomics related to organ transplantation, with an emphasis on the evolution of proteomic technologies.

Keywords

Biomarker discovery; proteogenomics; genomics; proteomics; microarray; transplantation; acute rejection; peptidomics

A Biomarker: What, why and why not?

A *biomarker* is a gene, protein/peptide or metabolite present in a biological system. It is indicative of a physiological or pathological state that can be recognized or monitored. Monitoring the biomarker thus provides a means for monitoring the disease condition and assists in its diagnosis and prognosis. The additional value of biomarkers lies in the fact that they can correlate with the underlying disease state. Thus, when assessed in biological fluids distant from the actual site of injury, it provides a non-invasive means to follow the underlying disease. A well-known biomarker is serum creatinine, which is a surrogate biomarker for underlying graft injury. Nevertheless, serum creatinine does not meet the criteria for an *ideal biomarker* for monitoring renal transplant patients as it lacks high specificity (serum creatinine is dependent on muscle mass and hydration status, in the absence of graft injury) and sensitivity (serum creatinine can be elevated with multiple causes of intrinsic and extrinsic graft injury). There is an unmet need of non-invasive which is more specific and sensitive to replace the renal transplant biopsy as the gold-standard.

Biomarkers for monitoring graft injury and survival course in the field of solid organ transplantation can be sought with the help of high throughput proteogenomic techniques (Figure 1).

These biomarkers could be a single or a panel of mRNA transcripts, proteins, or peptide fragments. The emerging field of metabolomics can also examine a metabolite or a panel of these molecules. An ideal biomarker in the field of organ transplant would be able to predict outcomes prior to and after transplantation; predict the onset and the severity of specific events such as acute and chronic graft rejection and infection; as well as predict the specific injury responses to immunosuppressive medications. Another need in transplantation is the ability to customize the dose (or load) of immunosuppression for the patient- i.e., finding the individual threshold between rejection and infection so that the individual can avoid allorecognition and retain normal infectious immunity. The examination of patients with spontaneous operational tolerance [1,2] has allowed for identification of a highly regulated gene set in peripheral blood. This may also bring biomarkers into clinical monitoring for customizing patient immunosuppression.

The *hunt for biomarkers* in the field of organ transplantation has been extraordinarily difficult. The primary problem lies with not having a true gold standard for the classification of different etiologies of graft dysfunction. The transplant biopsy faces many dilemmas: (1) it samples a fraction of the kidney while the underlying process of injury is often patchy (e.g. drug related injury, acute rejection); (2) the fraction of sample processed for biomarker studies is separate and often disparate from the tissue sample processed for histology; (3) the limited amount of tissue received from human transplant studies limits dissection of samples to examine specific cellular compartments or specific infiltrating cells; (4) the heterogeneity of the underlying process may not be recognized by pathology [3]; (5) there is inter-observer variability in pathological diagnosis [4]; (6) chronic graft injury is often pooled together but etiologies for that injury may be highly variable and may be driven by different mechanisms. This may thus require different biomarkers for graft monitoring. This lack of a true gold standard in transplantation thus creates the first hurdle in high throughput technology to analyze genome wide differences in genes or proteins. If the input data is flawed, can we expect meaningful output measures with relevant clinical correlations? An additional problem lies in the heterogeneity of the patients, and the source of tissues for the biomarker studies.

With the evolution of immunosuppression over the last decade, we have been flooded with multiple choices of immunosuppressive drugs. Thus, biomarkers in transplantation have to transcend differences in immunosuppressive drugs in order to be correlative with the health of the transplant. As the proteogenomic technologies are also expensive and time consuming, application of these strategies to very large patient sample numbers a daunting task. Selection of biomarker molecules is a critical step when building predictors of disease state based on gene expression data. Even though several univariate and multivariate gene selection approaches have been proposed, it is hard to make a fair comparison of the available results. This is a result of biased validation set and relatively small datasets [5]. Such small patient group size is one of the main reasons why gene data from a training group may not apply to a different validation patient sample set, when the second set is different in multiple parameters including demographics and treatment. An approach that combines data from individual research groups conducted on similar disease conditions and uses it for analyses could remedy this issue.

In a recent report Kong *et al* [6] applied a non-parametric meta-analysis approach for combining independent microarray datasets pertaining to chronic allograft nephropathy (CAN). The work used non-parametric meta-analysis approach on two CAN studies, and

identified 309 distinct genes that expressed differently in CAN. With the help of Fisher's exact test, the study found 6 KEGG pathways to be over-represented among the identified genes [6]. Another approach that utilized a unified framework for finding differentially expressed genes (DEG) has been shown to be better than other gene selection methods [7]. A framework for this kind of approach can use the following modules: (i) gene ranking – using two gene selection algorithms, namely, a) two-way clustering and b) combined adaptive ranking to rank the genes, ii) significance analysis of genes – converting the gene ranks into p-values (iii) validation – using - three fold validations of the obtained DEGs selected by false discovery rates (FDR) analysis [7].

From transcription to translation

Microarrays can interrogate gene expression profiles in various systems that range from bacteria to humans. Microarray technology is still evolving, and has progressed from early technology using nylon membranes to the current state of the art using glass wafers. As a result, many studies are being done ranging from basic science to clinical applications in cancer, diabetes and acute rejection of allograft in transplant. The publication of tens of thousands of research articles using the microarray technique since the appearance of the first paper in 1995 demonstrates its influence and impact on current biological research [8].

Several commercial arrays are currently available for researchers. Among these widely used platforms are: Affymetrix geneChips® (Affymetrix Inc, Santa Clara, CA), Agilent oligonucleotide microarrays (Agilent technologies, Palo Alto, CA), Illumina (Illumina Inc., San Diego, CA), and Nimblegen (Nutley, NJ). It is now widely accepted that each of these platforms have their own strengths and weaknesses. The researcher must choose the appropriate platform based on his/her priorities. We have seen the array technology platform along with several other genomic techniques evolve in recent years. For example, single nucleotide polymorphism (SNP) arrays detect SNPs of human genome [9], aberrations in methylation patterns [10], alterations in gene copy-number [11] alternative RNA splicing [12] are now available. Details on issues in the most recently published microarray studies relevant to human organ transplantation and experimental models in transplantation, are summarized in recent reviews [13-16].

In this review, our primary focus will be on proteomic technologies and their applications to organ transplantation. The field of proteomics involves identification and characterization of the proteins encoded by the genome. Traditionally, methods such as western blot and immunohistochemistry (IHC) were applied to detect and quantify proteins. Such techniques relied on the availability of respective antibodies, and required a relatively large protein quantity. Sophistication in chromatographic techniques and the emergence of powerful mass spectrometers has allowed for monitoring and analysis of thousands of proteins in a relatively short period of time with highly sensitive detection. The mass spectrometers provide mass to charge (m/z) ratio for the protein or the peptide being analyzed. The m/z obtained from mass spectrometers then is used to identify and quantify proteins and peptides using bioinformatics tools.

The power of two techniques

The gene expression profile for a particular gene may not necessarily correlate with the corresponding level of protein [17]. However, when possible, it adds to the power of discovery if the two or more methods can verify the presence or absence of the altered level of mRNA and its corresponding protein. After the initial phase of screening of mRNA transcript biomarker and protein biomarker, an effort to integrate such data can shed light on many questions currently unanswered. One such example is shown in Figure 2. The urinary β 2-microglobulin has been reported as a potential biomarker by Oetting *et al* [18]. When we

analyzed the microarray data published from our lab by Sarwal *et al* in 2003 [3] the expression level of β 2-microglobulin gene was significantly increased among patients with specific kinds of acute rejections (AR-I and AR-II, [3]. When we immunoassayed the level of intact β 2-microglobulin in an independent set of patients with AR (n = 20) and stable graft function (n = 20), the change in the level in between AR and stable graft function was found to be significant (P <0.02) (Figure 2). In a novel attempt using integrative proteomics and gene expression mapping, Ou *et al* identified 179 biomarker overlapping protein candidates using 2DE/MS and microarray experiments using breast cancer cell lines [19]. Out of 179 initially identified proteins, 9 genes were confirmed when tested with microarray on breast tissue. Among the 9 markers validated on breast tissue, 1 biomarker was re-confirmed and 3 novel biomarkers were identified for the first time [19]. This case and that of β 2-microglobulin are examples of how a concerted effort utilizing both genomics and proteomics could be beneficial. The approach is dubbed “proteogenomics”.

Proteomics: Evolving Technologies

Since proteins are involved in different cellular processes, an understanding of proteins inside the cell provides an insight into the cellular events. The proteomics approach provides an unbiased high-throughput approach to identify differentially expressed proteins in the healthy and disease states. For this very reason, proteomic analysis for biomarker discovery has been extensively applied to many fields of biomedical research, including oncology, diabetes, renal and urine related diseases, and solid organ transplantation [20-25]. A dramatic increase in the publications using proteomics has been observed similar to publications using microarrays.

The publication of human genome data along with the use of mass spectrometric techniques such as Matrix Assisted Laser Absorption Desorption Ionization Time of Flight (MALDI TOF) revolutionized our ability to identify proteins with the help of the database search engines. Analysis of tissue samples has led to the patterns of tissue-specific proteomes, or the set of proteins expressed in the cells of a particular organ system. Body fluids, such as cerebrospinal fluid (CSF), synovial fluid, and nipple fluid aspirates, have been used to identify protein signatures that differentiate samples with disease from healthy controls, and provide candidate proteins as surrogate biomarkers of disease. Blood and urine become more appealing candidates for their potential application in the clinical setting. Multiple groups have demonstrated the feasibility of urine proteomic studies by identifying reproducible patterns of normal urinary protein expression [20,23,26,27].

Gel based and gel free methods

Different proteomic methods are available for biomarker discovery efforts. Proteomic methods are broadly classified into gel-based and gel-free methods. Since the first publication from James O'Farrell using 2D gel electrophoresis (2 DE) in 1975, it remains one of the most popular platforms among the different proteomics platforms [28]. The increased reproducibility achieved with the use of immobilized pH gradient (IPG) strips and the use of mass spectrometry revolutionized the way proteins were resolved and identified. A new addition to conventional 2D gel electrophoresis is two dimensional difference gel electrophoresis (2D DIGE) in which as many as three fluorescently labeled protein samples can be analyzed in a single gel, thereby reducing gel-to-gel variation and increasing reproducibility [29]. Figure 3 presents an example of a DIGE experiment designed to identify possible protein markers for an AR episode. A DIGE gel was run to see the treatment response on the urinary proteins of renal patients who had undergone through biopsy proven acute rejection (AR) episode. The pooled urinary protein samples from AR patients were labeled with Cy3-Dye (green), and the pooled urine samples from post-AR

patients were labeled with Cy5-Dye (red). The green spots represent the proteins with an elevated level in AR urine, whereas, the red spots represent the proteins with an elevated level at 3 months post-AR urine (Sigdel and Sarwal unpublished data). Proteins in red dye (highlighted in the inset A) can serve as potential markers for treatment response as their concentration increased dramatically after the treatment, and the green spots (seen in inset B) can serve as a potential biomarker for AR, as this protein disappears as a result of AR treatment response, and concentration for these proteins went down as treatment response. An inherent problem with this technology is the detection limit.

Among the available gel free technologies, surface enhanced laser desorption/ionization-time of flight mass spectrometry (SELDI-TOF) has been used extensively in biomarker discovery studies [13,21,22]. This is rapid, easy to scale up, and generates patterns of peaks representing proteins and protein fragments, which then need to be identified using other technologies. Even though SELDI has been well standardized and commercialized for general use, and it is limited in its ability to discover the identity of key proteins in the disease processes directly. The second inherent problem with SELDI is its inability to look into a wider spectrum of proteome (i.e. > 15 kDa and < 4 kDa). These two limitations (namely protein identification and the relatively narrow working range) have forced researchers to seek other techniques that are better in both protein identification and in working range.

Development of electrospray ionization (ESI) and Matrix-assisted laser desorption/ionization (MALDI) assisted in the application of mass spectrometry to analyze biological samples [30]. Several MALDI-TOF instruments are currently being used for the analysis of different biological samples. This non-exhaustive approach helps researchers to revisit samples loaded onto the MALDI plate more than once. Its broader m/z range because of its nature of ionization and relative longer ionization time have some disadvantages with the MALDI method as compared to ESI [31]. With its robustness and low detection limit, this gel-free method of proteomics, which is based on liquid chromatography coupled with mass spectrometry (LC-MS), has become increasingly popular in efforts to discover biomarker molecules [32]. The last ten years have seen a significant improvement in the development and use of LC-MS-based proteomics in biomarker discovery efforts.

For protein profiling purposes, several LC-MS platforms are being employed. Each platform uses some kind of protein fractionation or enrichment step. This step may include depletion of highly abundant proteins or the enrichment of proteins using a method of choice. Based on the nature of the sample and the need of the study, researchers may either use a custom designed depletion method or one of many commercially available depletion kits listed in the next section. Several strategies for urinary protein enrichment have been suggested. For our urinary proteomic study, we found centrifugal filtration followed by the depletion of the six most abundant plasma proteins from renal transplant patient useful [33]. The protein level of enrichment or depletion is usually followed by enzymatic digestion which is followed by another step either to enrich or fractionate the complex peptide mixture. A technique called “multidimensional protein identification technology” (mudPIT) analyzes small proteolytically digested peptides from complex biological samples and has become a popular “bottom-up” approach [34]. Here a 2D liquid chromatography separation is used prior to mass spectrometric analysis. By virtue of their charge and hydrophobicity, peptides then are separated, analyzed and identified by tandem MS. The first dimension is normally a strong cation exchange (SCX) column followed by a reverse phase chromatography (RP). Isoelectric focusing (IEF) [35], free flow electrophoresis (FFE) [36] are other methods that employ peptide fractionation purpose. Specific peptide enrichment strategies are found to be useful in fractionating the complex samples such as Cysteinyll [37] or glycopeptides enrichment [38]. Application of such sample depletion, enrichment or fractionation has

facilitated analysis of a complex protein mixture, which has been popularly known as “Shotgun Proteomics” [39]. The availability of highly sensitive mass spectrometers such as triple quadrupole MS instruments, and newer 2D linear ion trap instruments such as LTQ FT and LTQ-orbitrap, have facilitated peptide identification of thousands of proteins.

The ability to analyze intact proteins without any enzymatic digestion by the use of high-end mass spectrometers such as hybrid linear ion trap (LTQ)-orbitrap [40] and Fourier-transform ion cyclotron resonance mass spectrometry (FT ICR) has been reported which can provide an alternative to the gel based “top down” approach.

Quantitative proteomics using stable isotopes

Accurate quantification of proteins or peptides is not easy and often requires some method that utilizes tagging of samples with some stable isotope. Several approaches exist that have been used in quantitative proteomics. (1) Stable isotope labeling by amino acids in cell culture (SILAC). SILAC is a simple approach for incorporation of a label into proteins for MS-based quantitative proteomics. SILAC relies on metabolic incorporation of a given 'light' or 'heavy' form of the amino acid into the proteins being synthesized in two cell populations [41]. (2) Gygi et al introduced Isotope Coded Affinity Tags (ICAT) 1999 [42]. ICAT analysis provides information of relative amounts of cysteine-containing peptides in enzymatically digested protein samples. The relative abundance of tryptic peptide is then used to quantify the expression of the corresponding protein. This technique was later commercialized by Applied Biosystems as ICAT®. (3) Yet another addition among these techniques is iTRAQ™ Reagents by Applied Biosystems which uses isobaric labels which, upon fragmentation in MS/MS, gave rise to four unique reporter ions (m/z 114-117) [43]. This technique is based upon chemically tagging the N-terminus of peptides generated from the protein digests with the label consisting of reporter and balancer groups. The reporter is quantitatively cleaved during collision-induced dissociation (CID) to the quantity of a single peptide of known mass with as many as four different samples. This unique ability of analyzing more than two samples has been successfully exploited in different clinical samples [44,45]. A new iTRAQ kit is now commercially available, and helps analyze as many as 8 samples per experiment. (4) $^{18}\text{O}/^{16}\text{O}$ labeling method uses [^{18}O] and [^{16}O] labeled water for tryptic digestion of two samples [46].

Apart from these isotope labeling methods; label free methods have also been used. Capillary electrophoresis coupled mass spectrometry (CE MS) has successfully been utilized in predicting the clinical outcome of congenital unilateral ureteropelvic junction obstruction in newborns as well as a biomarker discovery for urothelial cancer [22,24]. A similar method, utilizing the peptides first fractionated by HPLC and followed by MALDI TOF analysis (LC MALDI), has been used to discover oral fluid biomarkers for oral cancer [47]. A summary of the applications of proteomics in the field of transplantation is summarized in Table 1. A summary of currently popular proteomic methods and their strengths and concerns have been summarized in Table 2.

Avoiding noise in proteogenomic datasets

The course of biomarker discovery can be the one (with microarray) or the other (with proteomics) or both as shown in Figure 4. Given the heterogeneity of patient population and complexity of any subset of genome or proteome being interrogated, the task of biomarker search is daunting. To add to the complication, as reported by Mueller et al in a recent report, that the transcriptomic footprint could be continuous instead of dichotomous [48]. The presence of thousands of proteins and their concentration can span a dynamic range of at least 10 orders of magnitude, and only adds to the already complex proteomic sample [32]. Having a low signal to noise ratio is a well-recognized problem in these experiments.

Sophistication in automated data analysis software has resulted in minimizing the problem of background noise from hybridization and wash variables, but biological noise still remains a problem, specifically in samples where the signal level for the disease specific signatures can be relatively low. This was recently shown in a genomics study geared towards the discovery of peripheral blood biomarkers for graft rejection using whole blood analysis from PAXgene tubes [49]. The study reports how globin genes may interfere with biomarker discovery efforts for allograft rejection in peripheral blood samples.

Gel based proteomic techniques have a number of inherent issues of reproducibility and their high detection limit impedes the effort of detecting the low abundance protein biomarkers. The problem caused by masking of low abundance potential protein biomarkers in blood by high abundance plasma proteins was realized which led to the design and implementation of commercially generated depletion columns. Multiple Affinity Removal System columns Human 14, Human 7, Human 6 from Agilent Technologies designed to deplete 14, 7 and 6, the most abundant proteins among albumin, IgG, antitrypsin, IgA, transferrin, haptoglobin, fibrinogen, alpha2-macroglobulin, alpha1-acid glycoprotein, IgM, apolipoprotein A-I, apolipoprotein A-II, complement C3, and transthyretin (<http://www.chem.agilent.com>). The ProteoPrep® 20 Plasma Immunodepletion Kit from Sigma-Aldrich is designed to specifically remove the 20 most abundant proteins from human plasma or serum (<http://www.sigmaaldrich.com>). The ProteomeLab IgY-12 High Capacity Proteome Partitioning kits from Beckman Coulter, Fullerton, CA are specifically designed to remove twelve highly abundant proteins (Albumin, IgG, Transferrin, Fibrinogen, IgA, α 2-macroglobulin, IgM, α 1-Antitrypsin, Haptoglobin, Orosomucoid, Apolipoprotein A-I, Apolipoprotein A-II) from the human/primate biological fluids such as serum and plasma. Apart from this issue of masking low abundance proteins, overloading of protein needs to be addressed. Several strategies have been tested that range from 1D fractionation to multidimensional fractionation. As mentioned in the previous section, depletion of high abundance proteins from blood and urine samples, enrichment of urinary proteins, and use of different fractionation techniques, such as SCX followed by RP, FFE and IEF, have been used to address the issue. Specific peptides enrichment methods such as Cysteinylation and glycopeptides have also been successfully used for this purpose.

Supercomputing of proteogenomic datasets

One of the strengths of the omic tools is that they have tremendous power to generate a large volume of data dealing with multiple samples at the same time very effectively. This strength has been very useful in a broad spectrum, ranging from the research trying to elucidate basic cellular events to translational research aimed at discovering biomarkers for diseases. In one such work, Shyamsunder *et al* performed microarray analysis on 115 human tissue samples representing 35 different tissue types, and using cDNA microarrays that represented approximately 26,000 different human genes [50]. The data obtained was not only useful in seeing a tissue specific gene expression across different tissue types, but also provided researchers with a set of deposited publicly available data for future analysis with slightly different queries. Such publicly available data repositories have already been used to ask questions pertaining to different health conditions. Using publicly available microarray, genetics, proteomics and gene knock-down data from human and other species, English *et al* reanalyzed data from individual experiments related to identification of obesity related biomarkers [51]. The authors report that such an approach in reanalyzing the data obtained from different high throughput experiments significantly improved the sensitivity, specificity and precision of the prediction of obesity-associated genes. In a recently published work Brouard, Mansfield *et al* [1] identified a panel of 49 peripheral blood transcriptional biomarker genes that are associated with renal allograft tolerance from a study that included 75 renal-transplant patients and 16 healthy individuals. The panel of

genes has provided us with a so-called “tolerance footprint”. A new clinical study, with an even larger cohort of patients, will have to be pursued in order to understand more about the onset of graft tolerance among transplant patients. With the use of the proteomic approach, a similar panel of proteins that can potentially be used as therapeutic targets has been discovered for multiple sclerosis (MS). It is selective for specific pathological stages [52]. This study used six multiple sclerosis (MS) patients and two controls to identify the targets. Liu *et al*, looking for serum biomarkers for colorectal cancer using SELDI used a discovery set (n = 122) and a test set (n = 99) [53]. Mao *et al* used a SELDI fingerprint analysis with 49 subjects for discovery and 24 subjects for an independent test set for renal allograft sub-clinical rejection [54]. Since these techniques are being developed presently, there is no existing preliminary data that can be utilized for sample size prediction, or power calculation purposes. With more data being generated, efficient data integration and analysis is expected to add to the biomarker discovery effort.

Other issues

One of the criticisms with the *omic* efforts in biomarker discovery is that very few biomarker molecules have been successfully applied in a clinical setting. There are several factors that contribute to this slow pace of clinical use for the identified potential biomarkers. (1) Lack of a gold standard in the classification of the disease condition and the lack of a proper normalization method to categorize the samples accurately. (2) There are different techniques available that monitor the level of mRNA transcripts or proteins. Different platforms have their own strengths and weaknesses. It has been an issue among researchers for sometime now about the methodology for normalizing the data across platforms without losing its biological relevance. For the microarray, the issue of platforms has been discussed extensively [55,56]. (3) The methods available for use in discovery are slow and are limited by the number of samples run in a given time. Since the validation and verification step requires a much larger set of samples, it has now become the roadblock between the potential biomarkers for their verification and subsequent clinical use. (4) Because of the sensitivity of the methods being used and the method of sample preparation process (extraction, amplification, hybridization etc) there is no consensus on the threshold of significance. Usually, methods being employed for discovery use expensive and sophisticated instruments not suitable for a clinical setting since they are relatively expensive and require skilled personnel for their operation. A threshold set for such a technique may not hold true when the biomarker reaches the clinical testing method using a different technique.

Concluding Remarks

Due to the availability of human genome data and the relentless effort put towards building mapping and proteome databases such as Human Proteinpedia (www.humanproteinpedia.org) and human proteome initiative (<http://www.expasy.org/sprot/hpi>), proteogenomics presents one of the most attractive approaches towards discovering biomarkers for health issues including organ transplantation. Even though the discovery process is daunting because of the overwhelming presence of confounding factors and is compared to “a needle in proteome haystack”. It is perhaps, the continuous nature of the disease presence rather than a dichotomous signal [48,57], the need for a better diagnostic and prognostic tool and treatment course has always remained a strong motivator. If the current trend of improvement continues, clinicians will soon have the means available to help them select a better therapeutic strategy specific to individual patients and a disease phenotype. Stringent attention to sample selection, processing, confounding variables, and data analysis will improve our chances for success over past attempts. Proper identification of biomarkers would make assessment of treatment

efficacy and susceptibility to adverse events easier. The high throughput nature of proteogenomic approach has facilitated accumulation of large amount of informative data that will be instrumental in the discovery step, which will eventually complete the circle from data gathering and hypothesis generation to hypothesis validation and clinical application.

Acknowledgments

The work was supported by NIH Grant RO1-AI-061739, awarded to MS, and Child health Research Program and Deans Fellowship to TS. We are grateful to Mary Hansen and Jon Martin for proofreading the manuscript and helpful suggestions for manuscript revision.

List of Abbreviations

2DE	2 dimensional gel electrophoresis
2DE/MS	2 dimensional gel electrophoresis and mass spectrometry
IPG	immobilized pH gradient
SNP	single nucleotide polymorphism
DIGE	difference gel electrophoresis
LC-MS	liquid chromatography mass spectrometry
FT ICR	Fourier-transform ion cyclotron resonance mass spectrometry
mudPIT	Multidimensional Protein Identification Technology
SELDI TOF	Surface-enhanced laser desorption/ionization time-of-flight mass spectrometry
ICAT	Isotope-coded affinity tagging
iTRAQ	isobaric tags for relative and absolute quantification
MALDI TOF	Matrix Assisted Laser Desorption /Ionization- Time of Flight
CE MALDI	capillary electrophoresis Matrix Assisted Laser Desorption /Ionization
CSF	cerebrospinal fluid
SCX	strong cation exchange
RP	reverse phase
CAN	chronic allograft nephropathy
DEG	differentially expressed genes
SNP	single nucleotide polymorphism
FDR	false discovery rate

References

1. Brouard S, et al. Identification of a peripheral blood transcriptional biomarker panel associated with operational renal allograft tolerance. *Proc Natl Acad Sci U S A*. 2007; 104(39):15448–53. [PubMed: 17873064]
2. Sarwal M, Pascual J. Immunosuppression minimization in pediatric transplantation. *Am J Transplant*. 2007; 7(10):2227–35. [PubMed: 17711553]
3. Sarwal M, et al. Molecular heterogeneity in acute renal allograft rejection identified by DNA microarray profiling. *N Engl J Med*. 2003; 349(2):125–38. [PubMed: 12853585]

4. Furness PN, et al. International variation in histologic grading is large, and persistent feedback does not improve reproducibility. *Am J Surg Pathol*. 2003; 27(6):805–10. [PubMed: 12766585]
5. Lai C, et al. A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. *BMC Bioinformatics*. 2006; 7:235. [PubMed: 16670007]
6. Kong X, Mas V, Archer KJ. A non-parametric meta-analysis approach for combining independent microarray datasets: application using two microarray datasets pertaining to chronic allograft nephropathy. *BMC Genomics*. 2008; 9(1):98. [PubMed: 18302764]
7. Shaik JS, Yeasin M. A unified framework for finding differentially expressed genes from microarray experiments. *BMC Bioinformatics*. 2007; 8:347. [PubMed: 17877806]
8. Schena M, et al. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*. 1995; 270(5235):467–70. [PubMed: 7569999]
9. Syvanen AC. Toward genome-wide SNP genotyping. *Nat Genet*. 2005; 37(Suppl):S5–10. [PubMed: 15920530]
10. Schumacher A, et al. Microarray-based DNA methylation profiling: technology and applications. *Nucleic Acids Res*. 2006; 34(2):528–42. [PubMed: 16428248]
11. Pollack JR, et al. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet*. 1999; 23(1):41–6. [PubMed: 10471496]
12. Religio A, et al. Alternative splicing microarrays reveal functional expression of neuron-specific regulators in Hodgkin lymphoma cells. *J Biol Chem*. 2005; 280(6):4779–84. [PubMed: 15546866]
13. Weintraub LA, Sarwal MM. Microarrays: a monitoring tool for transplant patients? *Transpl Int*. 2006; 19(10):775–88. [PubMed: 16961768]
14. Sarwal MM, Li L. Designer genes: Filling the gap in transplantation. *Transplantation*. 2006; 82(10):1261–72. [PubMed: 17130773]
15. Li L, et al. Interference of globin genes with biomarker discovery for allograft rejection in peripheral blood samples. *Physiol Genomics*. 2008; 32(2):190–7. [PubMed: 17971501]
16. Ying L, Sarwal M. In Praise of Arrays. *Pediatr Nephrol*. 2008 in press.
17. Greenbaum D, et al. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol*. 2003; 4(9):117. [PubMed: 12952525]
18. Oetting WS, et al. Urinary beta2-microglobulin is associated with acute renal allograft rejection. *Am J Kidney Dis*. 2006; 47(5):898–904. [PubMed: 16632030]
19. Ou K, et al. Novel Breast Cancer Biomarkers Identified by Integrative Proteomic and Gene Expression Mapping. *J Proteome Res*. 2008
20. Reichelt O, et al. Prediction of renal allograft rejection by urinary protein analysis using ProteinChip Arrays (surface-enhanced laser desorption/ionization time-of-flight mass spectrometry). *Urology*. 2006; 67(3):472–5. [PubMed: 16527560]
21. Oates JC, et al. Prediction of urinary protein markers in lupus nephritis. *Kidney Int*. 2005; 68(6):2588–92. [PubMed: 16316334]
22. Theodorescu D, et al. Discovery and validation of new protein biomarkers for urothelial cancer: a prospective analysis. *Lancet Oncol*. 2006; 7(3):230–40. [PubMed: 16510332]
23. Ye B, et al. Proteomic-based discovery and characterization of glycosylated eosinophil-derived neurotoxin and COOH-terminal osteopontin fragments for ovarian cancer in urine. *Clin Cancer Res*. 2006; 12(2):432–41. [PubMed: 16428483]
24. Decramer S, et al. Predicting the clinical outcome of congenital unilateral ureteropelvic junction obstruction in newborn by urinary proteome analysis. *Nat Med*. 2006; 12(4):398–400. [PubMed: 16550189]
25. Schaub S, et al. Proteomic-based detection of urine proteins associated with acute renal allograft rejection. *J Am Soc Nephrol*. 2004; 15(1):219–27. [PubMed: 14694176]
26. Vitzthum F, et al. Proteomics: from basic research to diagnostic application. A review of requirements & needs. *J Proteome Res*. 2005; 4(4):1086–97. [PubMed: 16083257]
27. Yokomizo K, et al. Urinary protein analysis in pre- and postoperative cancer patients. *J Clin Lab Anal*. 2005; 19(6):253–9. [PubMed: 16302210]
28. O'Farrell PH. High resolution two-dimensional electrophoresis of proteins. *J Biol Chem*. 1975; 250(10):4007–21. [PubMed: 236308]

29. Unlu M, Morgan ME, Minden JS. Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. *Electrophoresis*. 1997; 18(11):2071–7. [PubMed: 9420172]
30. Cox J, Mann M. Is proteomics the new genomics? *Cell*. 2007; 130(3):395–8. [PubMed: 17693247]
31. Bogdanov B, Smith RD. Proteomics by FTICR mass spectrometry: top down and bottom up. *Mass Spectrom Rev*. 2005; 24(2):168–200. [PubMed: 15389855]
32. Qian WJ, et al. Advances and challenges in liquid chromatography-mass spectrometry-based proteomics profiling for clinical applications. *Mol Cell Proteomics*. 2006; 5(10):1727–44. [PubMed: 16887931]
33. Sigdel TK, et al. Optimizing protein recovery for urinary proteomics, a tool to monitor renal transplantation. *Clin Transplant*. 2008
34. Washburn MP, Wolters D, Yates JR 3rd. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol*. 2001; 19(3):242–7. [PubMed: 11231557]
35. Heller M, et al. Two-stage Off-Gel isoelectric focusing: protein followed by peptide fractionation and application to proteome analysis of human plasma. *Electrophoresis*. 2005; 26(6):1174–88. [PubMed: 15704250]
36. Weber G, Wildgruber R. Free-flow electrophoresis system for proteomics applications. *Methods Mol Biol*. 2008; 384:703–16. [PubMed: 18392590]
37. Liu T, et al. The use of a quantitative cysteinyl-peptide enrichment technology for high-throughput quantitative proteomics. *Methods Mol Biol*. 2007; 359:107–24. [PubMed: 17484113]
38. Thaysen-Andersen M, et al. Investigating the biomarker potential of glycoproteins using comparative glycoproteomics - application to tissue inhibitor of metalloproteinases-1. *Biochim Biophys Acta*. 2008; 1784(3):455–63. [PubMed: 18206988]
39. Diller R, et al. Therapy and prognosis of tumors of the genitourinary tract after kidney transplantation. *Transplant Proc*. 2005; 37(5):2089–92. [PubMed: 15964347]
40. Macek B, et al. Top-down protein sequencing and MS3 on a hybrid linear quadrupole ion trap-orbitrap mass spectrometer. *Mol Cell Proteomics*. 2006; 5(5):949–58. [PubMed: 16478717]
41. Ong SE, et al. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics*. 2002; 1(5):376–86. [PubMed: 12118079]
42. Gygi SP, et al. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol*. 1999; 17(10):994–9. [PubMed: 10504701]
43. Ross PL, et al. Multiplexed Protein Quantitation in *Saccharomyces cerevisiae* Using Amine-reactive Isobaric Tagging Reagents. *Mol Cell Proteomics*. 2004; 3(12):1154–1169. [PubMed: 15385600]
44. Griffiths SD, et al. The use of isobaric tag peptide labeling (iTRAQ) and mass spectrometry to examine rare, primitive hematopoietic cells from patients with chronic myeloid leukemia. *Mol Biotechnol*. 2007; 36(2):81–9. [PubMed: 17914187]
45. Abdi F, et al. Detection of biomarkers with a multiplex quantitative proteomic platform in cerebrospinal fluid of patients with neurodegenerative disorders. *J Alzheimers Dis*. 2006; 9(3):293–348. [PubMed: 16914840]
46. Mirgorodskaya OA, et al. Quantitation of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry using (18)O-labeled internal standards. *Rapid Commun Mass Spectrom*. 2000; 14(14):1226–32. [PubMed: 10918372]
47. Hu S, et al. Discovery of oral fluid biomarkers for human oral cancer by mass spectrometry. *Cancer Genomics Proteomics*. 2007; 4(2):55–64. [PubMed: 17804867]
48. teh Mueller TF, et al. Microarray analysis of rejection in human kidney transplants using pathogenesis-based transcript sets. *Am J Transplant*. 2007; 7(12):2712–22. [PubMed: 17941957]
49. Abdurrahman Z, et al. Sirolimus is not always responsible for new-onset proteinuria after conversion for chronic allograft nephropathy. *Pediatr Transplant*. 2007; 11(3):336–9. [PubMed: 17430494]
50. Shyamsundar R, et al. A DNA microarray survey of gene expression in normal human tissues. *Genome Biol*. 2005; 6(3):R22. [PubMed: 15774023]

51. Basic-Jukic N, et al. Arterial hypertension in renal transplant recipients. *Acta Med Croatica*. 2007; 61(2):171–6. [PubMed: 17585473]
52. Han MH, et al. Proteomic analysis of active multiple sclerosis lesions reveals therapeutic targets. *Nature*. 2008
53. Liu XP, et al. A serum proteomic pattern for the detection of colorectal adenocarcinoma using surface enhanced laser desorption and ionization mass spectrometry. *Cancer Invest*. 2006; 24(8): 747–53. [PubMed: 17162557]
54. Mao Y, et al. Diagnosis of renal allograft subclinical rejection by urine protein fingerprint analysis. *Transpl Immunol*. 2008; 18(3):255–9. [PubMed: 18047934]
55. Shi L, et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol*. 2006; 24(9):1151–61. [PubMed: 16964229]
56. Shippy R, et al. Using RNA sample titrations to assess microarray platform performance and normalization techniques. *Nat Biotechnol*. 2006; 24(9):1123–31. [PubMed: 16964226]
57. Pillutla RC, et al. Target validation and drug discovery using genomic and protein-protein interaction technologies. *Expert Opin Ther Targets*. 2002; 6(4):517–31. [PubMed: 12223066]
58. Barrett T, et al. NCBI GEO: mining millions of expression profiles--database and tools. *Nucleic Acids Res*. 2005; 33(Database issue):D562–6. [PubMed: 15608262]
59. Chen R, Li L, Butte AJ. AILUN: reannotating gene expression data automatically. *Nat Methods*. 2007; 4(11):879. [PubMed: 17971777]
60. Zhou H, et al. Exosomal Fetuin-A identified by proteomics: a novel urinary biomarker for detecting acute kidney injury. *Kidney Int*. 2006; 70(10):1847–57. [PubMed: 17021608]
61. Holly MK, et al. Biomarker and drug-target discovery using proteomics in a new rat model of sepsis-induced acute renal failure. *Kidney Int*. 2006; 70(3):496–506. [PubMed: 16760904]
62. Pan TL, et al. Expression, by functional proteomics, of spontaneous tolerance in rat orthotopic liver transplantation. *Immunology*. 2004; 113(1):57–64. [PubMed: 15312136]
63. Dear JW, et al. Liver proteomics for therapeutic drug discovery: Inhibition of the cyclophilin receptor CD147 attenuates sepsis-induced acute renal failure. *Crit Care Med*. 2007; 35(10):2319–28. [PubMed: 17944020]
64. Huguet S, et al. Identification by proteomic tool of atypical anti-liver/kidney microsome autoantibodies targets in de novo autoimmune hepatitis after liver transplantation. *Ann N Y Acad Sci*. 2007; 1109:345–57. [PubMed: 17785324]
65. Emadali A, et al. Proteomic analysis of ischemia-reperfusion injury upon human liver transplantation reveals the protective role of IQGAP1. *Mol Cell Proteomics*. 2006; 5(7):1300–13. [PubMed: 16622255]
66. Emadali A, et al. Proteomic analysis of tyrosine phosphorylation during human liver transplantation. *Proteome Sci*. 2007; 5:1. [PubMed: 17199894]
67. O'Riordan E, et al. Characterization of urinary peptide biomarkers of acute rejection in renal allografts. *Am J Transplant*. 2007; 7(4):930–40. [PubMed: 17331118]
68. Jahnukainen T, et al. Proteomic analysis of urine in kidney transplant patients with BK virus nephropathy. *J Am Soc Nephrol*. 2006; 17(11):3248–56. [PubMed: 17035609]
69. Schaub S, et al. Proteomic-based identification of cleaved urinary beta2-microglobulin as a potential marker for acute tubular injury in renal allografts. *Am J Transplant*. 2005; 5(4 Pt 1):729–38. [PubMed: 15760396]
70. Nelsestuen GL, et al. Proteomic identification of human neutrophil alpha-defensins in chronic lung allograft rejection. *Proteomics*. 2005; 5(6):1705–13. [PubMed: 15800973]
71. Wolters DA, Washburn MP, Yates JR 3rd. An automated multidimensional protein identification technology for shotgun proteomics. *Anal Chem*. 2001; 73(23):5683–90. [PubMed: 11774908]
72. Johnson T, et al. A CE-MALDI interface based on the use of prestructured sample supports. *Anal Chem*. 2001; 73(8):1670–5. [PubMed: 11338578]
73. Floyd PD, et al. Characterization of peptides from Aplysia using microbore liquid chromatography with matrix-assisted laser desorption/ionization time-of-flight mass spectrometry guided purification. *J Chromatogr A*. 1999; 830(1):105–13. [PubMed: 10023620]

Figure 1.

An array of proteomic and genomic techniques that are being used for biomarker discovery. 2D gel/DIGE – 2 dimensional gel electrophoresis and difference gel electrophoresis; qRT-PCR – real-time quantitative PCR; LC-MS/MS – liquid chromatography mass spectrometry/mass spectrometry; SELDI MS - Surface-enhanced laser desorption/ionization mass spectrometry; LC MALDI – Liquid Chromatography Matrix Assisted Laser Desorption/Ionization.

Figure 2.

Proteogenomics: Linking microarray and proteomic data. $\beta 2$ -microglobulin has been reported as a biomarker for acute renal allograft rejection. Panel A shows significant rise in the gene expression level of β -microglobulin in some rejection subtypes (AR I and AR II) from previously published microarray data [3], Panel B shows significant rise in $\beta 2$ -microglobulin as measured by ELISA (Sigdel and Sarwal, unpublished data), and Panel C, a work published by Oetting *et al* using MALDI TOF [18]. This shows a similar and significant trend of $\beta 2$ -microglobulin in the graft and urine during the process of acute transplant rejection.

Figure 3. The use of 2D difference gel electrophoresis (DIGE) is an attractive tool to identify protein biomarker candidates that are differentially present in healthy and disease condition
A DIGE gel was run to see the treatment response on the urinary proteins of renal patients who had undergone through biopsy proven acute rejection (AR) episode. The urinary protein samples from AR and post-AR were labeled with Cy3-Dye (green) and Cy5-Dye (red) respectively. The green spots are proteins elevated in AR whereas the red spots are proteins elevated level in post-AR urine. Proteins in red and green (highlighted in the inset A and B) can be used as a marker of treatment response and AR respectively.

Figure 4.

A schematic for possible pathways for the discovery pipeline for biomarkers for disease using proteogenomics. The ability to cross-annotate microarray and proteomic platforms, using publicly available tools (e.g. GEO [58], SymAtlas: (<http://symatlas.gnf.org>), AILUN, [59]) will allow for the ability to identify significant biomarkers by different technologies, within similar diseases.

Table 1

A list of protein biomarkers reported by the use of various proteomic approaches.

Organism	Disease	Specimen	Biomarkers	Techniques	Year	Reference
Human	Kidney Injury	Urine	26 proteins	2DE/MS	2006	[60]
Mouse	Acute renal failure	Urine	Merin-1- α	DIGE	2006	[61]
Mouse	Liver transplant tolerance	serum	Haptoglobin	2DE/MS	2004	[62]
Mouse	Sepsis induced acute renal failure	Liver	Cyclophilin	DIGE MS	2007	[63]
Rat	de novo autoimmune hepatitis	Liver	carbonic anhydrase III, glutathione S-transferase (GST), and β -1 subunit of proteasome	2DE	2007	[64]
Human	IRI	Liver	IQGAP1	SDS PAGE, LC MS	2007	[65]
Human	IRI	Liver	NCK-1	SDS PAGE, LC MS		[66]
Human	AR for renal transplant	Urine	β -defensin-1 (4.7 kDa) and α -1-antichymotrypsin (4.4 kDa),	SELDI, MS/MS	2007	[67]
Human	BK virus-associated renal allograft nephropathy (BKVAN)	Urine	m/z values of 5,872, 11,311, 11,929, 12,727, and 13,349 kD	SELDI	2006	[68]
Human	Acute rejection of renal graft	Urine	β 2-microglobulin	SELDI	2005	[69]
Human	Chronic lung allograft rejection	bronchoalveolar lavage fluid (BALF)	human neutrophil peptides (HNP)	MALDI MS, MS/MS	2005	[70]

Table 2

An Evaluation of Different Proteomic Techniques:

	Methods	Strengths	Weaknesses
Gel based	<ul style="list-style-type: none"> 2D gel electrophoresis (2DE)[28] 	<ul style="list-style-type: none"> Simple experiment setup Low cost and easy to run Easy identification of proteins and run on the western blot 	<ul style="list-style-type: none"> Low sensitivity Poor reproducibility Poor detection of low abundance proteins Needs relatively more sample
	<ul style="list-style-type: none"> 2D DIGE [29] 	<ul style="list-style-type: none"> Needs relatively less sample compared to 2DE More sensitive compared to 2DE Better detection limit and better reproducibility Quantitative 	<ul style="list-style-type: none"> Requires relatively more sample than for LC MS Expensive and needs special user expertise
Gel free	<ul style="list-style-type: none"> LC MS (shotgun: no fractionation) 	<ul style="list-style-type: none"> Fast and simple 	<ul style="list-style-type: none"> Suffers from the dynamic range of protein concentration in the sample Exhaustive (samples once used can not be revisited)
	<ul style="list-style-type: none"> LC MS (MudPIT- with fractionation) [71] 	<ul style="list-style-type: none"> Very sensitive and easy protein identification 	<ul style="list-style-type: none"> Suffers from false positive protein IDs Exhaustive (samples once used can not be revisited)
	<ul style="list-style-type: none"> CE MALDI [24,72]/ LC MALDI [73] 	<ul style="list-style-type: none"> Samples can be revisited for reanalysis 	<ul style="list-style-type: none"> Quantitative nature of the assay if the samples are not run immediately after the CE or HPLC run