CHAPTER 7.

Platforms for biomarker analysis using high-throughput approaches in genomics, transcriptomics, proteomics, metabolomics, and bioinformatics

B. Alex Merrick, Robert E. London, Pierre R. Bushel, Sherry F. Grissom, and Richard S. Paules

Summary

Global biological responses that reflect disease or exposure biology are kinetic and highly dynamic phenomena. While high-throughput DNA sequencing continues to drive genomics, the possibility of more broadly measuring changes in gene expression has been a recent development manifested by a diversity of technical platforms. Such technologies measure transcripts, proteins and small biological molecules, or metabolites, and respectively define the fields of transcriptomics, proteomics and metabolomics that can be performed at a cell-, tissue-, or organism-wide basis. Bioinformatics is the discipline that derives knowledge from the large quantity and diversity of biological, genetic, genomic and gene expression data by integrating computer science, mathematics, statistics and graphic arts. Gene, protein and metabolite expression profiles can be thought of as snapshots of the current, poorly-mapped molecular landscape. The ultimate aim of genomic platforms is to fully map this landscape to more completely describe all of the biological interactions within a living system, during disease and toxicity, and define the behaviour and relationships of all the components of a biological system. The development of databases and knowledge bases will support the integration of data from multiple domains, as well as computational modelling. This chapter will describe the technical platform

methods involving DNA sequencing, mass spectrometry, nuclear magnetic resonance combined with separation systems, and bioinformatics to derive genomic and gene expression data and include the relevant bioinformatic tools for analysis. These genomic, or omics platforms should have wide application to epidemiological studies.

Introduction

The sequencing of the human genome stands as one of the major scientific achievements of the twentieth century. It embodies a defining moment in modern biology by which most high-throughput technologies are compared for size scope, and complexity. Beginning in 1990, it took roughly a decade for the first draft of the human genome to be completed. By 2003, about 99% of all gene-containing regions were described, numbering about 20 500 genes (1), although some regions of the genome, such as centromeres, telomeres and gene deserts, continue to undergo characterization and study. Data from the human genome project has provided a generalized human map of the three billion nucleotides comprising the DNA of a few human subjects. However, studies on the variations (polymorphisms) in human DNA sequences are currently underway; samples from 270 individuals of multiethnic backgrounds are being used in a consortium called the International HapMap Project for haplotype mapping (http://www.hapmap.org) (2). The goal is to identify the patterns of single nucleotide polymorphism (SNP) groups, called haplotypes or haps, among individual human beings. In addition, interpretation of the human genome has been greatly enhanced by the DNA sequencing of many other genomes that allow comparison of genetic organization, evolution and function. Nearly 300 genomes have been completely sequenced and range from unicellular organisms, like *E. coli* and *S. cervisiae*, to model invertebrate organisms, such as *Drosophila melanogaster* and *C. elegans*, to several mammalian species for which the completed and ongoing genome projects are all available online (3).

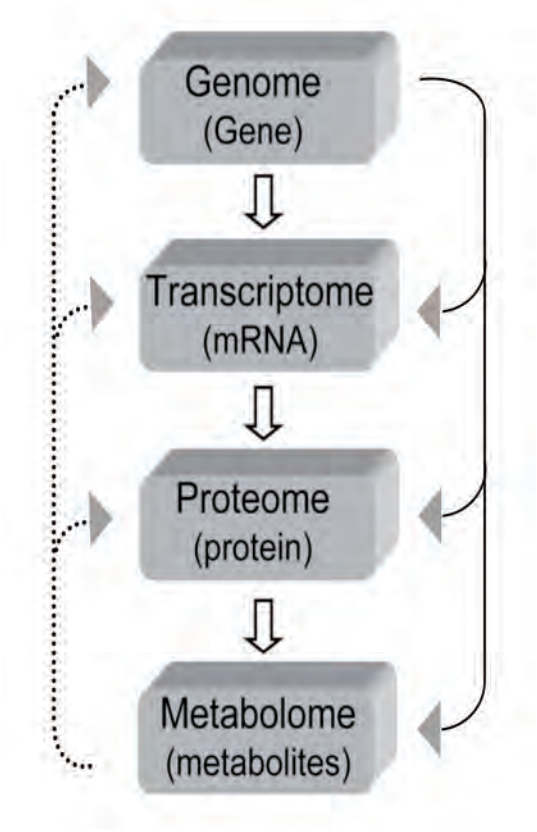
Although the conception of the idea for sequencing the human genome is relatively recent, the project could not have occurred without the preceding decades of biological and technological developments. Particularly noteworthy of these contributions are early cytogenetics and chromosomal studies at the beginning of the twentieth century by Morgan and colleagues, the discovery of DNA structure by Watson and Crick in 1953, DNA cloning in 1973 by Berg and Cohen, the DNA sequencing reaction in 1975 by Sanger, reverse transcriptase in 1970 and restriction endonucleases in 1971, and the polymerase chain reaction (PCR) by Mullis in 1983 (4). A new century now begins with an era of "omics," those fields describing a multitude of genomic functions aimed at further deciphering the biological meaning of sequences in the human genome.

The purpose of this chapter is to describe the technical platforms in genomics, transcriptomics, proteomics, metabolomics and bioinformatics that could be useful in epidemiologic studies. These analytical platforms favour high sample throughput and generation of large data sets.

Omes and omics

Gene expression constantly changes during health, adaptation, toxicity, disease and aging. While the genetic blueprint of an individual is relatively static, the various levels of gene expression to form and operate a complex organism are dynamically regulated, structurally complex and spatially determined. At any point in time, only a portion of a genome is expressed in specific cells and tissues. At the mRNA level of gene expression, the transcriptome represents all genes transcribed at any one moment, and the proteome is the complement of proteins making up cells and tissues. Small molecules and metabolites comprise the metabolome. The global study of each gene expression level is suffixed with "omics," such as transcriptomics, proteomics and metabolomics. Figure 7.1 suggests a sequence of gene expression based on genomic DNA sequences that are dynamically reflected in changes of transcripts, proteins and metabolites. Each level of gene expression (represented by upward, curved, dotted lines) has the opportunity to feed back and influence other levels reflective of highly integrated, multicellular processes in cells, tissues and organisms. Studies of each gene expression area utilize very different technical platforms to maximize large scale coverage of the transcriptome, proteome and metabolome. Technical platforms may involve mass parallel analysis using robotics, miniaturization, automation and computer processing. The integration of the many levels of gene expression is often referred to as systems biology by bioinformatics or computational biology. Bioinformatics represents an applied field of mathematics to biochemistry and molecular biology using statistics, computer science and artificial intelligence to design algorithms to derive biological meaning from gene expression data.

Figure 7.1. The interdependence of the cellular metabolome, proteome, transcriptome, and genome. Each type of characterization provides a functional indication of the activity of the proceeding set of molecules (solid lines). Conversely, there will be some degree of feedback regulation built into the system (dotted lines).



Genomics

Chromosomal abnormalities are responsible for many developmental defects and malignancies, and include rearrangements in genomic DNA or changes in copy number, such as deletions, duplications and amplifications. Identification of genomic changes and mutations that underlie disease rely on comparisons of DNA sequences between affected and unaffected individuals. Finding disease-causing chromosomal abnormalities by genomic analysis is confounded by the fact that many sequence polymorphisms are functionally irrelevant and produce no observable biological consequence. Detection of many disease-causing mutations, such as those in *p53* in the Li-Fraumeni syndrome (5) and ATM in ataxia telangiectasia (6), have been found by sectional resequencing of genomic DNA or PCR-amplified DNA or RNA. However, *de novo* sequencing of individuals using automated Sanger-based capillary electrophoresis systems has so far been practical for only small regions of the human genome-containing candidate genes.

Recent advances in nucleic acid sequencing technologies using massive parallel sequencing, called next-generation sequencing, now allow sequencing of much larger genomic intervals (7). Sequencing of entire genomes can take place within a matter of several weeks, in a comprehensive search for chromosomal aberrations and mutations that affect phenotype. DNA sequencing does have inherent advantages in achieving single-base resolution and importantly for *de novo* analysis of samples without the prior knowledge of existing DNA sequence required for fabricated sequence platforms (8). New sequencing technologies that are high-throughput and low-cost while maintaining high accuracy and completeness are in continued development (9). New platforms often integrate real-time (RT) PCR and may incorporate microelectrophoresis, sequencing by hybridization, mass spectrometry, high-density oligonucleotide arrays, or incorporation of nanopore technology to bring within sight the goal of routine human genome sequencing (10) for personalized medicine (11).

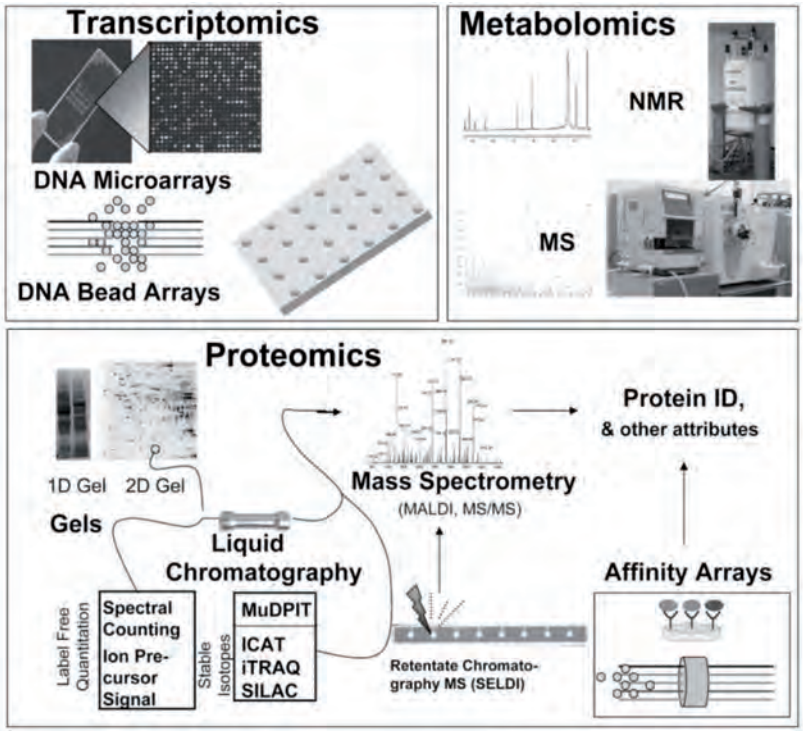
There are several alternatives to whole-genome assessment of chromosomal abnormalities that do not involve traditional DNA sequencing. Alternative platforms involve selective genotyping of very focused genomic loci (short sections of DNA) that are potentially related to disease susceptibility. Haplotype mapping data have been extremely useful in providing candidate genomic regions with high polymorphic variation. Hundreds of thousands of loci can be very rapidly genotyped using BeadArray platforms, a technology based upon direct hybridization of whole-genome-amplified (WGA) genomic DNA to BeadArrays of locus-specific, 50mer oligonucleotide sequences (12). As such, genome-wide association studies (GWAS) comprise an important evolving field in genetic epidemiology in which more than 450 GWAS have been published, and the associations of greater than 2000 single nucleotide polymorphisms (SNPs) or genetic

loci have been reported so far (13). Equally as important are high-density oligonucleotide microarrays for SNP detection in linkage analysis of susceptibility genes often used in cancer studies (14) or pharmacogenomics (15). In addition, fluorescence in situ hybridization (FISH) provides a visual map to examine all the chromosomes of a patient for abnormalities using fluorescent probes for specific genes or whole chromosome probes (16). Although high-throughput sequencing methodologies have been developed to accommodate the demand for sequence output, they consume large amounts of a valuable and potentially limiting genomic DNA. WGA can potentially remove DNA as a limiting factor for genomic analyses (17) that include multiple displacement amplification (MDA), primer extension preamplification (PEP), and degenerate oligonucleotide primed PCR (DOP) (18). However, genomic amplification technologies generate a certain level of replication error in sequence, which should be considered during verification studies. In summary, bead or chip DNA arrays or FISH platforms exemplify whole-genome technologies for high-throughput, compared to DNA sequencing for detection of genetic variation that can be linked to disease-based foci, protein, biomarkers and pharmacogenomic responses.

Transcriptomics

Transcriptomics studies the full, global complement of mRNA molecules expressed in cells and tissues. Some 20 500 genes are present within the human genome, of which about 10–15 000 are expressed at any one time in any particular tissue (19). Many expressed genes are necessary to perform basic functions of the cell, regardless of cell type or tissue, but a proportion of the expressed genes contribute to a cell's unique phenotype and specialized functions. Beginning around 1989, DNA microarrays, consisting of thousands of high-density cDNAs or oligonucleotides on support surfaces (called chips), were introduced and have evolved into powerful and versatile platforms for transcriptomic analysis (20–23). Spotted microarrays, either cDNAs or oligonucleotides, have been used extensively since the late 1990s, particularly by academia-based research scientists (see Figure 7.2). Commercial oligonucleotide arrays provide highly reproducible platforms representing the entire genome. Oligonucleotides from 25–70 bp in length are arrayed by either spotting pre-synthesized oligonucleotides directly onto glass, chemically synthesizing directly onto glass substrates (e.g. Agilent Technologies, Inc.),

Figure 7.2. Platforms for gene expression in transcriptomics, proteomics and metabolomics. Transcriptomic platforms are cDNA or oligonucleotides bound to glass slides or microbeads for analysis of mRNA. Metabolomic platforms are nuclear magnetic resonance (NMR) or mass spectrometry (MS) instruments for small biologic molecules or metabolites. Proteomic platforms can be gel-based or liquid chromatography-based (e.g. linear column gradients or multidimensional chromatography (MuDPIT)) for separation of proteins before identification (ID) by mass spectrometry. Use of stable isotopes greatly facilitates protein quantitation (ICAT, isotope coded affinity tags; iTRAQ, isobaric tags for relative and absolute quantitation; SILAC, stable isotope labelling by amino acids in culture), while non-isotopic or label-free methods can also be used that include spectral counting and ion precursor signal measurement. Retentate chromatography mass spectrometry (i.e. surface enhanced laser detection ionization (SELDI)) has been used for rapid profiling of biofluid samples using chemically reactive surfaces for separation and MALDI for generating protein mass spectra. Alternatives for MS-based proteomics involve affinity arrays, such as antibody microarrays or fluorescently tagged antibody bound bead suspensions (i.e. Luminex technology).



or by synthesizing directly onto quartz wafers by photolithographic technology (e.g. Affymetrix, Inc.) (24). In addition, oligonucleotides can be covalently linked with microbeads that can then be used in a 96-well microtiter dish format or on a glass substrate (e.g. Illumina, Inc.). With ever increasing technological advances, microarrays have progressed from chips with only several hundred probes to modern DNA chips reflecting expression of thousands, to even millions, of features per array.

The strength of any gene expression analysis, and the ability to determine expression profiles as potential biomarkers of exposure or effect, is dependent on proper experimental design and careful execution to minimize sources of variance and error and maximize useful biological information. Thus it is critical, in any microarray experiment, to design proper controls to include with samples of biological interest. Proper sampling and integrity of the RNA obtained from those samples are vital in determining the success of the analysis. Once RNA is isolated, proper labelling, hybridization, washing and scanning can dramatically influence the integrity of the resulting data. Since transcript expression is generally expressed in relative terms of a fold change of a particular gene expressed in one sample relative to a value in a control or normal sample, it is critical that the investigator structure the experiment in such a way as to minimize variations other than the one variable to be tested. Fortunately, commercial microarray providers have added increasingly more stringent quality control measurements in the production facilities. This has resulted in high reproducibility and low variation in microarrays coming from a manufacturer. Investigators have been allowed to shift resources away from multiple analyses of single samples to focusing on expanding the numbers of experimental samples. In turn, this has resulted in significant improvements in the confidence of results from microarray experiments. In fact, several large consortium efforts have demonstrated that comparable biological affects could be revealed in carefully controlled experiments in which multiple commercial microarray platforms, as well as rigorously quality-controlled spotted cDNA arrays, were used to analyse the same biological material (25–27).

Additional technological improvements have allowed for the reduction in the starting amounts of mRNA required in the labelling processes for the commercial platforms. Most labelling protocols used a single, PCR-based linear amplification of sample mRNA, which is used to incorporate a nucleotide conjugated with a fluorescent dye, biotin, or some other chemical modification. This amplification step has reduced the starting material for a sample to be analysed to only a few micrograms of total mRNA or less. Furthermore, protocols have been developed for additional rounds of PCR-based amplification of starting mRNA samples that make it possible to analyse very small quantities in the range of nanograms, and even picograms, of mRNA. These developments have facilitated gene expression profiling of samples derived from laser capture microdissection (LCM), for example, as well as biopsy samples, and other clinically derived samples that are limited in quantity. In addition, recent technological developments, particularly using bead-based microarrays (e.g. Illumina BeadChip), have opened up the possibility of using formalin-fixed, paraffin-embedded material for gene expression analysis.

An accessible, biological material fluid of principal interest to several clinical research scientists is blood. Many researchers are interested in testing the utility of gene expression profiling of peripheral blood leukocytes to generate biomarkers as surrogates for other tissues or organs affected in disease or injury processes (28). The utility of this approach has been demonstrated in studies of inflammatory responses and diseases in both animal models and humans, of neurological disorders, of angiomyolipoma (AML) and renal cell cancers, and of cardiac injury (29–36). Recent studies have used gene expression analysis of blood samples to generate molecular profiles as biomarkers of exposure and exposure-induced injury to arsenic, benzene, tobacco smoke and hepatotoxic levels of acetaminophen (37–41).

A common application in transcriptomics, useful in epidemiology studies, is to compare transcript outputs between normal and diseased tissues in what has been termed transcript profiling or expression profiling. Transcript expression studies can query all known or predicted genes in an organism, providing an abundance of information that represents a snapshot of the expression status of a tissue at any given time. One can gain considerable insight into molecular mechanisms from properly structured microarray experiments, both on the level of individual genes and on the level of biological pathways and processes. The potential for mRNA degradation makes expression profiling most applicable to freshly isolated tissues, cultured cells or flash-frozen tissue sections, but not paraffin-embedded tissue. While microarray approaches

can be used to interrogate the entire genome on a single microarray chip, focused arrays representing distinct gene subsets have been used to focus upon changes in specific pathways or processes. These include both glass slide-based microarrays (e.g. the National Institute of Environmental Health Sciences' Human ToxChip (42)), and PCR-based gene expression analyses (e.g. SuperArrays).

DNA arrays can also reflect epigenetic effects upon gene expression. Epigenetics is defined as heritable changes in gene expression that are not due to DNA sequence alterations. Methylation is the most common epigenetic change and is detected by bisulfite conversion, methylation-sensitive restriction enzymes, methyl-binding proteins, and anti-methylcytosine antibodies. Combining these techniques with DNA microarrays and high-throughput sequencing has made the mapping of DNA methylation feasible on a genome-wide scale. Genomic DNA methylation occurs particularly at cytosines in clusters of cytosine-guanine dinucleotides, or CpG islands (p is the phosphodiester bond between C and G bases). Methylation of CpG islands in promoter regions frequently results in gene silencing, which normally occurs during development (43), but is often observed as an early alteration in some cancers by causing inactivation of tumour suppressors genes, such as von Hippel-Lindau disease (VHL), inhibitor of cyclin-dependent kinase 4a ($p16^{INK4a}$), and breast cancer gene 1 (*BRCA1*) (44).

DNA microarrays have also been developed for expression beyond profiling. In addition to SNP and comparative genomic hybridization (CGH) applications mentioned in the previous section, genome-wide localization of transcription factor binding sites can be accomplished by chromatin immunoprecipitation (ChIP) analysed on a microarray chip that forms the so-called ChIP-on-chip technique (45). The method can be innovatively combined with different types of DNA arrays, such as SNP chips, to form "ChIP-on-SNP" (46). The future for array technologies will also bring about a revolution in clinical DNA diagnostics (47), develop pharmaceuticals in pharmacogenomics (48), and personalized medicine (49).

Proteomics

The field for describing protein expression on a global scale is proteomics, which aims to detail the structure and functions of all proteins in an organism over time. The wide application of proteomics has generated great interest in many established disciplines of exposure biology and medicine, including the field of epidemiology (50,51). Chemical or toxicant exposure can bind to or modify proteins, produce changes in protein expression, and dysregulate critical biological pathways and processes that lead to toxicity and disease, which in theory should be detectable by proteomic analysis. Primary aims in proteomic analysis are the discovery of key modified proteins, the determination of affected pathways, and the development of biomarkers for association with and eventual prediction of disease.

The complexity of a proteome, represented by the total protein expression of a specific cell, organ, tissue or biofluid, presents numerous challenges for comprehensive analysis. Proteins are more complex than nucleic acids, and therefore proteomic analysis involves measurement of just some of the many attributes of proteins during any single expression analysis (52). Proteins exhibit many attributes of interest to biomarker development in epidemiology studies, including determination of protein sequence identity, quantity, post-translational modifications (PTM), protein–protein interactions, structure and function. Some of the challenges in proteomic analysis include: defining the identities and quantities of an entire proteome in a particular spatial location, such as serum or subcellular structures like mitochondria; the existence of multiple protein forms and complexes; the evolving structural and functional annotations of the human and rodent proteomes; and integration of proteomics data with transcriptomics or other expression data. Primary aims of proteomic analysis are to achieve maximal proteome identification, quantitative high-throughput protein measurement, timely analysis, and discovery-oriented platforms. Proteomic platforms represent combinations of technologies that describe protein attributes by the separation, quantitation and identification of all proteins in a biological sample. Proteomic analysis includes four broad categories of proteomic platforms: mass spectrometry has played a central role in proteomic platform development in large part because of its sensitive and versatile ability to identify proteins; the ability to separate proteins greatly determines the designation of platform type by gel-based separation or liquid chromatophic separation linked to mass spectrometry (53); solid phase adsorption, based on partitioning of peptides and proteins due to specific chemical properties, has been exploited in rententate chromatography combined with mass spectrometry; and finally, affinity chromatography, which

sorts and identifies proteins in one reaction, is exemplified by use of antibodies in various formats (54). The following proteomic platforms represent some of the primary technologies being used for separating, identifying, and quantifying proteins during toxicoproteomic studies (Cf. Figure 7.2).

2D PAGE and DIGE

Two-dimensional polyacrylamide gel electrophoresis (2D PAGE) systems have been combined with mass spectrometry in an established and adaptable platform. Since 1975, 2D PAGE has been the most commonly used proteomic platform to separate and comparatively quantitate protein samples (55). Current state-of-the-art 2D gels use immobilized pH gradient (IPG) gels to separate proteins by charge. They are then resolved by mass spectrometry using sodium dodecyl sulfate (SDS) gel electrophoresis for effective separation of complex protein samples in µg to mg quantities. Either visible stains, such as Coomassie Blue or silver, or fluorescent staining are used for sensitive protein detection. After electronic alignment (registration) of stained proteins in 2D gels by image analysis software, intensities of identical protein spots are compared among treatment groups and a ratio (fold change) is calculated for each protein. A relatively new variation of the 2D PAGE technique, difference gel electrophoresis (DIGE), allows an investigator to measure three samples per gel that have been labelled with Cy2, Cy3 and Cy5 fluorescent dyes, which reduces some of the error associated with electronic registration during multiple gel alignment. This strategy allows for direct comparison of samples on one gel for better reproducibility and quantitation than conventional image analysis for comparison of multiple 2D gels. Thus, separation of proteins by 2D gels, using single stains or multiple fluors (i.e. DIGE), can be combined with mass spectrometry for ready protein identification to form a versatile and discovery-oriented platform for use in proteomic studies (56). In addition, some protein samples are sufficiently limited in their protein content that a simple size separation (one dimension) by SDS–PAGE can be used to identify protein bands of interest by mass spectrometry in 1D-Gel-MS.

Multidimensional LC-MS/MS

Proteomic platforms incorporating liquid chromatography (LC) as the primary means of separation (versus gel-based separations) have become the preferred means of analysis. There are many different types of LC separations often termed "multidimensional." In this proteomic platform, LC is used to separate protein digests by exploiting different biophysical properties of proteins before identification by tandem mass spectrometry (MS/MS). One of the most notable multidimensional LC-MS/MS platforms is Multidimensional Protein Identification Technology (MuDPIT). MuDPIT attempts to identify all proteins in a sample by two-dimensional separation of protein digests by charge (strong anion exchange matrix) and hydrophobicity (C18 column) with online LC immediately before entry into a tandem mass spectrometer (MS/MS) for protein identification (57). The platform has also been called shotgun proteomics, as entire protein lysates are trypsin digested into thousands of peptide fragments without any prior fractionation before separation and identification. Advantages of this newer platform are the potential for detection and identification of low abundance proteins that may not be observed in gel staining-based methods. One drawback is that LC-MS/MS platforms, like MuDPIT, are only semiquantitative and somewhat low-throughput in capacity.

The issue of protein quantitation in proteomics is an important one, since changes in protein expression may be a matter of altering existing gene and protein expression rather than turning them on (induction) or off (repression), which makes quantitation of proteins crucial in normal and diseased or control and experimental states. The use of stable isotopes, as detailed in the section below, takes advantage of the high resolution power of mass spectrometers to discriminate protein samples stable-isotopically labelled for quantitative comparison. However, rapid developments are being made using "label-free" approaches to quantitation if sample mass spectral data are sufficiently detailed (58). The two main approaches used are ion precursor signal intensities and spectral counting. Though they deliver relative sample quantitation, versus absolute protein measurement, they are simpler and less costly than stable isotopic methods, but not as precise. Label-free protein quantitation methods should be considered at the outset of designing LC-MS/MS proteomic studies and weighed against the considerable advantages and choices of stable isotopic approaches.

Stable isotope LC-MS/MS platforms: ICAT, iTRAQ and SILAC

A primary goal of proteomics is to comprehensively analyse all proteins in a sample, or as many

as possible. However, the prospect of quantifying protein levels for comparison among protein samples has been a difficult aspect of proteomic analysis. Protein quantitation can be considered either in relative terms as a proportion of treatment (test) samples compared to control samples, or in absolute terms as the number of molecules (moles) or concentration (molarity). Internal standards are useful, but not realistic, for complex protein samples of unknown composition in most proteomic studies. Since many proteomic platforms are based in mass spectrometry, comparison of intensity signals seems the most direct means for comparative measurements; however, intensities are subject to many interfering factors. Quantitation by mass spectrometry has generally been regarded as semiquantitative under the best of circumstances.

The use of stable isotopes for tagging proteins has made great strides in proteomics for determining the relative amounts of proteins among samples (59). Stable isotopes of an element differ in mass due to the number of neutrons, but have the same elemental and chemical characteristics as the element. Stable isotopes are not radioactive. Common stable elements and their stable isotopes are ^{1}H and ^{2}H; ^{12}C and ^{13}C; ^{14}N and ^{15}N; ^{16}O and ^{18}O; ^{32}S and ^{34}S. A unique feature of high-resolution mass spectrometers is the ability to finely distinguish between small differences in mass, even to the point of resolving the relative abundance of stable isotopes in otherwise identical samples. Several proteomic platforms for protein quantitation and identification have been built around the use of isotopic tagging of proteins (isotope coded affinity tagging (ICAT), peptides (isotope tag for relative and absolute quantitation (iTRAQ), or metabolic incorporation of isotopically tagged amino acids in cell culture (stable isotope labelling with amino acids in cell culture (SILAC)). The applications of stable isotopes in proteomics have been recently reviewed for their sensitive detection of proteins in a quantitative and comparative fashion (60). As mentioned above, continuing improvements in spectral counting for use in LC-MS/MS platforms should have wide utility as a versatile, isotope-free method of protein quantitation when stable isotope use is not feasible (61).

SELDI-TOF mass spectrometry

Retentate chromatography-mass spectrometry (RC-MS) is a high-throughput proteomic platform that creates a laser-based mass spectrum (based on matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometry) from a chemically-absorptive surface. The principle of this approach is the adsorptive retention of a subset of sample proteins on a thin, chromatographic support (i.e. hydrophobic, normal phase, weak cation exchange, strong anion exchange or immobilized metal affinity supports). The absorptive surfaces are placed on thin metal chips which can be inserted into a MALDI-type mass spectrometer. The laser rapidly desorbs proteins from each sample on a metal chip to create a mass spectrum profile. RC-MS can be performed upon any protein sample, but thus far this platform has found greatest utility in the analysis of serum and plasma for disease biomarker discovery (62). The lead commercial platform of RC-MS proteomic platforms is the surface-enhanced laser desorption ionization time-of-flight mass spectrometry (SELDI-TOF-MS) instrument (63). Analysis of samples is relatively rapid (100/day), and only a few µl of sample is necessary. A downside is that protein identification of peaks is not readily accomplished without additional conventional separation and analysis.

Antibody arrays

Protein microarrays represent a promising new proteomic tool that closely emulates the design for parallel analysis of DNA microarray technology (64). Protein microarray formats can be divided into two types of multianalyte sensing formats: forward phase arrays and reverse phase arrays. In the forward phase array format, the analyte is captured from the solution phase by different capture molecules, such as an antibody immobilized on a substratum (i.e. glass slides). In contrast, the reverse array format immobilizes the individual test samples in each array spot so that, for example, hundreds of different patient blood or tissue samples are arrayed and probed with one detection protein, such as an antibody and a single analyte endpoint are measured for comparison across multiple samples. Such microarray studies have been carried out for metastatic ovarian cancer (65).

Many different types of capture molecules can be arrayed, including peptides (i.e. peptide substrates for kinases on phosphorylation arrays), proteins (protein–protein interaction arrays) and oligonucleotides (i.e. transcription factor binding arrays to oligonucleotides), but the most prevalent are antibody arrays in the forward phase format. Antibody arrays can directly separate proteins from complex biological fluids like plasma, serum or cell lysates by affinity binding to specific antigenic sites on target proteins.

Generally, current commercial antibody array platforms fall into three classes based on the targeted proteins: cytokine/chemokine arrays, cellular function protein arrays and cell signalling arrays. Although not all proteins for any given cell type or biofluid (i.e. blood, serum, plasma, urine, cerebral spinal fluid) are currently represented on antibody arrays, they do provide a rapid screen for protein alterations that may be relevant to tissue injury or disease (54). Antibodies can be placed in ordered array on glass slides or on a fluorescent microbead format (i.e. Luminex technology) for multiplexed separation, identification and quantitation (66). For example, in a study investigating the chemotherapeutic and radiotherapy of patients with rectal cancer, 40 tumour samples were analysed by DNA microarray and plasma samples were analysed by antibody (Luminex) bead microarray platforms. Using a kernel-based method with Least Squares Support Vector Machines to predict rectal cancer regression grade, investigators found that combining and integrating of microarray and proteomics data improved predictive power leading to the best model based on five genes and 10 proteins with an accuracy of 91.7%, sensitivity of 96.2% and specificity of 80% (67). In a different approach, a molecular epidemiological study used SELDI-TOF MS for *in vivo* studies of humans exposed to benzene. By using two sets of 10 exposed and 10 unexposed subjects, researchers identified with chemically-reactive surfaces and validated with antibody-coated chips three differentially expressed proteins in the serum of benzene-exposed individuals, two of which were identified as PF4 and CTAP-III, both members of the CXC-chemokine family (68). The same can be done with peptides (instead of antibodies) as the affinity ligand. This method was applied in the development of two diagnostic antibodies against avian influenza detection for epidemiologic studies, in which the epitopes of two monoclonal antibodies (mAbs) against avian influenza nucleoprotein (NP) were found using truncated NP recombinant proteins and peptide array techniques (69).

Future developments in proteomics will see incorporation of more sophisticated methods of quantitation in proteomic analysis (70), combining higher data density LC-MS/MS platforms with stable isotope labelled peptides, spectral counting, and parallel use of complementary proteomic platforms, such as tissue arrays (71). Study designs that remove abundant proteins from biofluids, enrich subcellular structures, and include cell-specific isolation from heterogeneous tissues will greatly increase differential expression capabilities. Advancement in mechanistic insights and biomarker development using proteomics will be furthered by completely defining plasma (serum) proteome and circulating microparticles in humans and rodent species as accessible biofluids (72). Some of the representative biomarkers and patterns of protein and message expression are shown in Table 7.1. Reviews on using proteomics to develop biomarkers and further mechanistic insights have been published (73–75).

Metabolomics

Analogous to the genomic characterization of cellular DNA and the proteomic characterization of the set of proteins expressed at a given time, cells also can be

Table 7.1. Recently identified biomarkers and signatures of toxicity using transcriptomics and proteomics

Biomarker/Signature	Identification	Condition	Reference No.
KIM-1	DNA microarray	Renal toxicity	(145)
Adipsin	DNA microarray	GI toxicity, functional gamma secretase inhibitors (FGSIs)	(146)
CXC-chemokines	SELDI	Benzene exposure	(68)
Troponin I,T	2D gel-MS	Myocardial ischemia, infarction	(147)
Aminopeptidase-P Annexin A1	2D gel-MS	Radioimmunotherapy	(148)
12 Lipid-gene signature	DNA microarray	Drug-induced phospholipidosis	(149)
Multigene blood signature	DNA microarray	Systemic sepsis	(32)

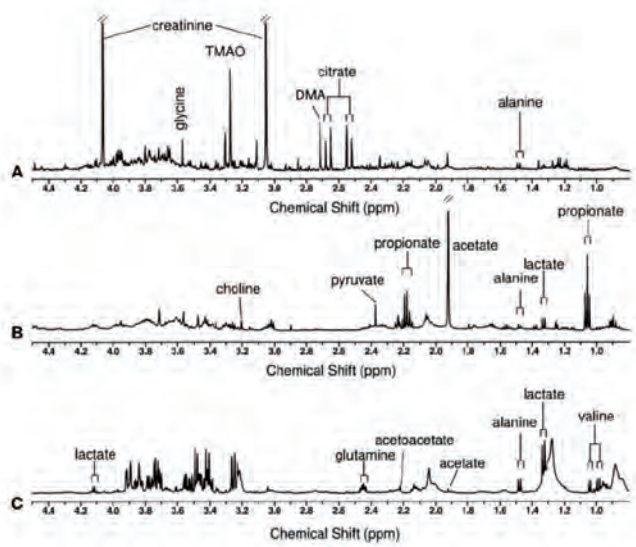
characterized in terms of the set of low-molecular-weight metabolites (typically < 1500 D) that comprise the cellular "metabolome." The cellular metabolome provides a functional readout of the cellular proteome (Cf. Figure 7.1). Although the analysis of homogeneous cell populations in culture, receiving identical nutrients and oxygenation, and exposed to the same levels of excreted waste products, represents the most ideal system for metabolomic characterization, the approach has been extended to the analysis of extracts and fluids derived from higher organisms. Urinary and blood metabolites have been among the most frequent targets for metabolomic characterization, but analyses of other fluids, such as cerebrospinal fluid (CSF), bronchoalveolar lavage fluid (76) and saliva (77), and of cellular extracts also have been performed. Typical ^{1}H NMR spectra illustrating the different metabolite composition of urine, blood, and saliva are shown in Figure 7.3. Currently, metabolomic characterization is being used for a wide range of objectives in human nutrition and toxicology, and for the development of pharmaceuticals and agricultural products. The underlying objectives of these studies include: discovery of metabolite signatures as prognostic indicators, diagnostics or biomarkers of disease states; establishing toxicological markers for drug development and environmental toxicology; understanding mechanisms of metabolic diseases; and correlation of metabolite phenotypes (metabotype) with genotype and environmental input (e.g. nutrition).

The screening of neonates for genetic disorders in intermediary metabolism is an application that predates the more recent interest in metabolomic characterization. The recent reviews of analytical approaches for clinical diagnosis of metabolic disorders (78,79) summarize methods for metabolite analysis and provide good examples of the application of MS to metabolomic analysis (Cf. Figure 7.2). Mass spectrometric analysis typically requires preparation of the metabolic components using either gas chromatography (GC) after chemical derivatization, or LC, with the newer method of ultra-performance liquid chromatography (UPLC) increasingly used. The use of capillary electrophoresis (CE) coupled to MS also has shown some promise. It was reported that a combination of approaches for metabolite extraction produced over 10 000 unique metabolite features, indicating both the complexity of the human metabolome and the potential of metabolomics in biomarker discovery (80).

Other more specialized techniques, such as Fourier transform infrared (FTIR) spectroscopy and arrayed electrochemical detection, have been used in some cases (81,82). The main limitation of FTIR is the low level of detailed molecular identification that can be achieved. In one study, MS was also employed for metabolite identification (82).

The extensive application of nuclear magnetic resonance (NMR) for metabolic characterization of urine and other body fluids has been developed primarily by Nicholson and coworkers (83–85). The primary advantage of the NMR approach (Cf. Figure 7.2) is the lack of required preparatory separations, which in turn leads to an unbiased and potentially quantitative measure of the constituents of the sample. Alternatively, while in principle it

Figure 7.3. Typical 1H NMR spectra illustrating the metabolite composition of urine (A), saliva (B) and plasma (C). Used with permission from (77).



TMAO, trimethylamine oxide; DMA, dimethylamine.

is possible to obtain a molecular mass corresponding to each unfragmented metabolite observed in a mass spectrum, in an NMR spectrum, molecular information is distributed among the resonances of a compound, and the position of these resonances can depend critically on pH, salt concentration, divalent ions and other physical parameters. Additional caveats discussed in various reviews include loss of more volatile metabolites, metabolite contributions derived from intestinal bacteria (86), potential bacterial growth in stored fluids, and other factors (84,87). Some have proposed using dimethylamine and its nitroso metabolite as biomarkers for small bowel bacterial overgrowth (86), while others have suggested monitoring 4-hydroxyphenylacetate as a potential screening method for small bowel disease and bacterial overgrowth syndromes (88). Significant levels of ethanol in plasma or urinary samples generally indicate either bacterial contamination of the sample or small bowel bacterial overgrowth (87). Such conditions typically accompany renal failure or other serious illnesses.

In studies of chemical toxins or pharmaceutics, the xenobiotic and its metabolites and conjugates typically constitute an important source of variation of the metabolome. While of interest from a metabolic perspective, these compounds are not directly indicative of organ toxicity or therapeutic response, so that in general these metabolites are not relevant to the study. One approach to dealing with this issue involves stable isotope labelling of the compounds under study, so that the compound and metabolites derived from it will exhibit characteristic features in the NMR or mass spectrum. Alternatively, if the test compound and all of its metabolites and conjugates can be identified, these can simply be ignored or eliminated from the analysis. One general source of toxicity resulting from the administration of high levels of test compounds is the depletion of sulfur-containing amino acids that results from the excretion of glutathione and cysteine conjugates. Thus, it was reported that rats receiving high levels of acetaminophen excreted significant amounts of pyroglutamic acid. This effect of excess acetaminophen was prevented/reversed by supplementation with methionine (89). The glutathione analogue ophthalmic acid recently has been found to accumulate after high dosage with acetaminophen, and may also function as a biomarker for oxidative stress and glutathione depletion (90).

Various multivariate analyses of metabolite composition have been applied to detect differences among subject groups, such as those receiving different treatments or different chemical exposures. This type of analysis, termed “metabonomics” by the Nicholson group, most frequently utilizes principal component analysis (PCA) of the spectral data to reveal clustering behaviour that differentiates treated from control subjects. The clusters of data points in PC plots reveal the uniformity of the control and treated groups, as well as the extent to which the treated group yields a distinct metabolic phenotype. Since the axes of the PC plot are dependent on the data set and do not correspond to independent variables, the ability of different laboratories to utilize the published information is limited. Interestingly, a recent study that evaluated statistical methodology for the analysis of gene expression data found that the use of PCA to reveal clustering behaviour generally degrades cluster quality, and concluded that PCA was only useful in special cases (91). Hence, there is a critical need for the identification of metabolic biomarkers that provide universally quantifiable indications of organ function and toxicity.

Another critical issue related to the identification of biomarkers is the need to separate acute and chronic effects of illnesses or toxins. It is not unusual for a particular metabolite to become elevated during the acute phase of a toxic response, but to become depressed as the chronic effects become significant. Alternatively, in the absence of chronic effects, the metabolite level may return to pretreatment values. For this reason, plots of data obtained at different times after dosage, or trajectory plots for individual subjects, can provide critical information on the time-dependent response. Several of the issues discussed above are illustrated in studies identifying the association of the oxidative stress biomarker 8-oxoguanosine with Parkinson's disease (92). In this study, elevation of the 8-oxoguanosine level was observed in cerebrospinal fluid (CSF), but not in the serum of patients with Parkinson. Further, there was a significant negative correlation between the level of the biomarker and the duration of the disease. Finally, as indicated in Table 7.2, 8-oxoguanosine is also elevated in other conditions, e.g. amyotrophic lateral sclerosis (93). Another important limitation on the identification of some biomarkers relates to the chemical reactivity, which can deplete the free metabolite pool and lead to heterogeneous adduct formation and difficulties of detection. Homocysteine, which has long been linked to cardiovascular disease, provides one example (94).

Much of the early NMR work evaluating the effects of various toxins noted changes in the levels of tricarboxylic acid cycle (TCA) metabolites and other abundant molecules that may be present in the nutrient source and that are not organ-specific (85). More recently, several more specific metabolomic biomarkers have been correlated with various diseases or treatments (Table 7.2). Notably, most of the biomarkers that have been related to metastatic growth are proteins, but new metabolomic markers continue to be developed. The metabolite 12(S)-hydroxyeicosatetraenoic acid (12(S)-HETE) has been demonstrated to play a pivotal role in experimental melanoma invasion and metastasis, suggesting that 12-lipoxygenase expression may be important in early human melanoma carcinogenesis (95–97). Changes in phosphate-containing metabolites and in phospholipid composition have been correlated with tumour stage and metastatic spread. In studies of extracts from tumour cell lines, elevations in phosphorylcholine or other membrane-related phosphomonoesters have frequently been observed, and suggested to be correlated with metastatic potential (98,99). Increases have been observed in aspartyl-4-phosphate

Table 7.2. Biomarkers recently identified using metabolomics

Biomarker	Sample Analysis	Condition	Reference No.
NAN; 2-PY	NMR – human and rodent urine	Type 2 diabetes mellitus	(110)
NAN; 2-PY	NMR – rat urine	peroxosome proliferation	(111,112)
ADMA	MS –human blood plasma	Renal failure; atherosclerosis	(108,109)
Ophthalmic acid	MS – mouse serum, liver extract	Acetaminophen-induced hepatotoxicity	(90)
Pyroglutamic acid	NMR – rat urine	APAP-induced deficiency of sulfur- amino acids	(89)
3-nitrotyrosine	HPLC - human CSF	Amyotrophic lateral sclerosis	(150)
8-oxoguanosine	HPLC – human CSF	Alzheimer's disease	(93)
8-oxoguanosine	HPLC – human CSF	Parkinson's disease	(92)
Modified nucleosides	LC-IT-MS of human urine	Breast cancer	(102)
12(S)-HETE[a]	HPLC - tumor cell extracts	Human melanoma	(95-97)
Aspartyl-4-phosphate	DESI-MS/NMR – murine urine	Lung cancer/ tumour growth	(100)
Phosphorylcholine	31P NMR – cell extracts	Breast cancer cell extracts	(98)
Depressed lysophosphatidyl choline levels	31P NMR – blood plasma	Renal cell carcinoma	(101)
Elevated xanthine, hypoxanthine, urate	GC-MS – human urine	Lesch-Nyhan syndrome	(103)
Glc-Gal-pyridinoline	HPLC - human urine	Synovial degradation – RA	(104-106)
4-hydroxyphenyl acetate	GC-LC – human urine	SBBO	(88)
Dimethylamine, nitrosodimethylamine	GC – human serum; GC – whole blood	SBBO	(86)

[a] In most reported studies, concentration or enzymatic activity of 12-lipoxygenase, rather than 12(S)-HETE, has been determined.
2-PY, N-methyl-2-pyridone-5-carboxamide; ADMA, NG,NG-dimethylarginine; APAP, Acetaminophen; DESI, Desorption electrospray ionization; GC-LC, Gas chromatography-liquid chromatography; GC-MS, Gas chromatography-mass spectrometry; HPLC, High performance liquid chromatography; LC-IT-MS, Liquid chromatography-ion trap-mass spectrometry; NAN, N-methylnicotinamide; NMR, Nuclear magnetic resonance; RA, rheumatoid arthritis; SBBO, small bowel bacterial overgrowth.

that may correlate with tumour growth (100). Phosphorus-31 NMR studies of blood plasma derived from patients with advanced renal cell carcinoma have been found to exhibit depressed levels of lysophosphatidylcholine (101). A significantly improved discrimination of breast cancer patients based on metabolomic analysis of modified nucleosides present in urine was recently reported (102).

A metabolomic approach using a combination of gas chromatography and MS has identified elevations in the levels of the metabolites xanthine, hypoxanthine, urate and guanine in patients with Lesch-Nyhan syndrome (103). The urinary cross link product, Glc-Gal-pyridinoline (104), has been identified as a biomarker for synovial degradation observed in osteoarthritis (105); treatment with ibuprofen lowers the level of this excreted metabolite (106). Endogenously formed N^G,N^G-dimethylarginine, also referred to as asymmetric dimethylarginine (ADMA), is a potent inhibitor of nitric oxide synthase (107). Plasma levels increase as a consequence of renal failure (108), and ADMA has been identified as a biomarker for atherosclerosis (109). Metabolomic analyses have identified several pyridine derivatives in urine from diabetic rats (110), and the same derivatives have shown up as biomarkers of peroxisome proliferation (Table 7.2) (111,112). The presence of these compounds indicates a perturbation of the tryptophan-nicotinamide adenine dinucleotide (NAD) pathway.

As is typical for new technologies, some reports of putative biomarkers have proven controversial. Early identification of the partly-characterized metabolites CFSUM1 and CFSUM2, associated with chronic fatigue syndrome, were subsequently demonstrated to arise from incompletely derivatized pyroglutamic acid and serine (113,114), and the quantitative abnormalities of these metabolites in urine from patients with chronic fatigue syndrome/myalgic encephalomyelitis has been reported to be artefactual. Early analyses supporting the use of ^{1}H NMR of blood sera to diagnose coronary artery disease (115) have subsequently been found to be more equivocal than originally suggested and to compare unfavourably with angiography-based diagnosis (116).

As more specific biomarkers are identified, the power of this approach will continue to evolve, providing useful diagnostic information for pathological, environmental, toxicological, pharmaceutical and nutritional research, as well as enhancing the value of metabolomic analysis for basic research into mechanisms of toxicity. Future developments in mass spectrometry platforms in metabolomics will increase the detectable coverage of the metabolome in clinical specimens and experimental species, and permit better identification of metabolites in the process of converting raw data to biological knowledge (117).

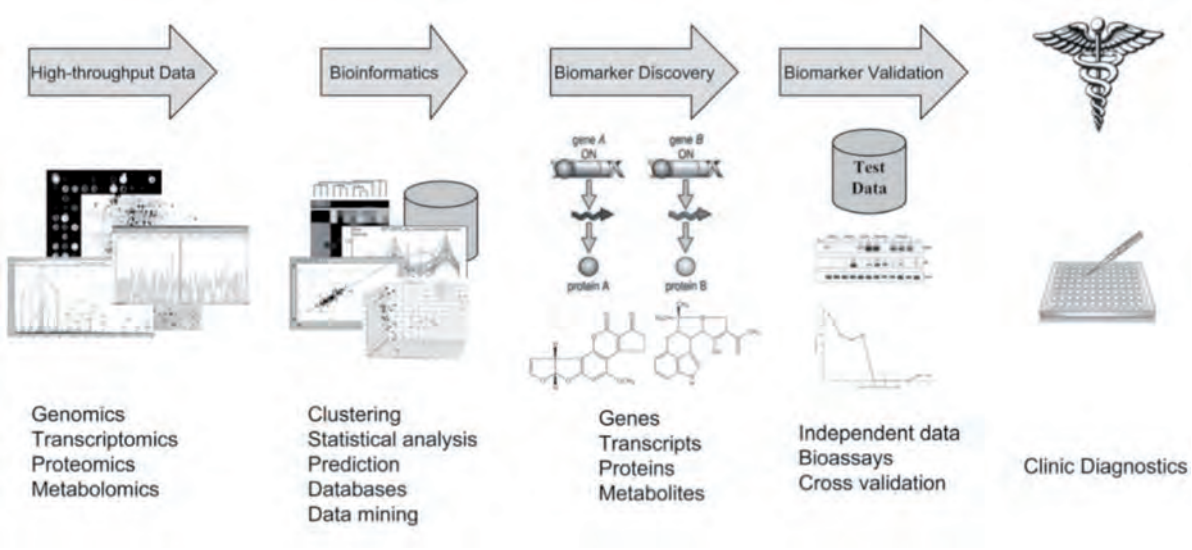
Bioinformatics

The wealth of data generated through high-throughput omics approaches has become increasingly complex and too vast for conventional biomarker analysis strategies. Bioinformatics has played a crucial role in biomarker discovery and validation (Figure 7.4). It is a multidisciplinary field involving biology, computer science, mathematics and statistics to derive knowledge from biological, genomics and genetic data (118,119). Database systems, computational algorithms, statistical models, data mining methods and other analytical tools are typically employed in a bioinformatics framework to effectively manage, analyse and summarize the plethora of data. For example, proteomics approaches, such as SELDI-TOF and mass spectrometry in conjunction with bioinformatics tools, have greatly facilitated the discovery of new and better serum biomarkers to detect cancer (120).

The bioinformatics processes to translate omics data into clinically useful biomarkers can comprise a myriad of steps, beginning with initial analysis of the data to validation of the biomarkers (121). This multistep process typically involves discovery, data integration, predictive modeling, and delivery of the biomarkers to the clinic in a format to facilitate implementation (122). Bioinformatic analyses may take different approaches with several checkpoints along the way. A flowchart is described for the application of bioinformatics strategies to improve the identification of candidate biomarkers from cancer genome-wide expression analyses (123). The process proceeds with acquisition of gene expression data from cancer tissues, followed by the identification of candidate genes as biomarkers. The next step entails meta-mining public cancer data sets of the same type of pathophysiology to reduce the biomarker false-positive rate. The last step before use of the biomarkers in clinical trials involves validation of the candidates by RT–PCR, ELISA, tissue arrays, immunohistochemistry, and other types of bioassays.

What are these bioinformatics tools and processes and how are they used to aid in the discovery and validation of disease biomarkers? It is helpful to appreciate that a useful biomarker must be objective, highly accurate and very reliable in determining disease states and

Figure 7.4. Sequential scheme for the integration of omics with bioinformatics in the biomarker discovery process



assessing risk. In other words, they must generalize well (i.e. extend) to broad cases, exhibit significance in their reporting, and be precise in their utility. Unfortunately, omics data possesses systematic variation due to the experimental error in the data acquisition process (124). There are technical limitations in the sensitivity of detecting biomarkers that are lowly expressed or non-abundant; however, biomarkers do not need to be highly expressed or in large abundance. Thus, the challenge in biomarker identification is successfully mining omics data with inherent error to find the features that reliably, accurately and objectively relate to the pathophysiology of a disease (125).

Data normalization and dimension reduction (data condensing) techniques have been widely used to preprocess omics data before biomarker discovery. Standard ways of dealing with data normalization have been adopted for omics data. Robust Multiarray Average (RMA), loess and quantile normalization methods have seemed to pass the test of time (126–128). A systematic variation normalization (SVN) approach was developed specifically to remove systematic error from microarray gene expression data (129). Baseline subtraction, signal smoothing and normalization methodologies were employed in the preprocessing of mass spectrometry proteomics data to reduce the noise and to make the analysis of spectral data comparable (130). In general, once the omics data is made unbiased and adjusted to make fair comparisons across samples, all the data can be used to mine for biomarkers or be filtered first to remove uninformative or redundant information. Some believe that the inclusion of uninformative features in the biomarker selection process will severely degrade the performance of the predictor model (131). Thus, it has been suggested to remove variables that do not contribute to a biological response of interest before the selection of biomarkers. Filtering of the data can be based on a signal or relative level, fold change, a confidence level, standard deviation from the mean of the distribution, P-values, mutual information, full or partial correlations, or more elaborate methods. The underlying omics data should be rich enough to be narrowed down to a core set of features that best represent the biology of the system for biomarker selection. Caution must be taken with respect to the preprocessing of omics data for biomarker identification, as the use of a particular combination of methods will surely add variability to the set of indicators selected. In other words, the preprocessing and filtering steps are sources of variability in and of themselves. The US Food and Drug Administration-led MicroArray Quality Control Phase II (MAQC-II) consortium set out to address this by trying to understand the limitations of various bioinformatics data analysis methods in developing and validating microarray-based predictive models, and determining if best practices for development and validation of predictive models based on microarray gene expression and genotyping data can be derived for biomarker discovery and personalized medicine (132).

The development and utilization of classification and prediction methods for analysis of omics data have transcended the process

for identifying biomarkers from molecular signatures. One of the most widely used methods for classification is clustering. The process works by using a dissimilarity measure for the feature profiles in the omics data to iteratively form groups of samples that are tightly clustered. Features that cluster well together and can distinguish between groups of samples that differ in pathophysiology are considered to be potential biomarkers. Clustering and an F-test-like score based on within- and between-sample gene expression variance measures were used effectively to identify an intrinsic gene subset (i.e. a molecular portrait) that has a high predictive score for human breast tumours (133). More sophisticated bioinformatics methods have been developed to identify potential biomarkers. A hybrid approach was developed based on the genetic algorithm (GA) and k-nearest neighbours (KNN) classifier that is capable of identifying gene and protein molecular signatures of diseases based on microarray and proteomics data, respectively (134,135). The GA serves as a search tool to choose small subsets of predictors, whereas the KNN functions as a non-parametric (no distribution model assumed) pattern recognition method to evaluate the discriminative ability of the subsets. More recently, a hybrid approach for biomarker discovery from microarray gene expression data was developed to distinguish between types of cancer (136). This approach is based on Fisher's ratio (a measure for the linear discriminative power of variables) to select features "wrapped" with a classifier (hence, the procedure is called FR-Wrapper) to perform predictions. With these hybrid approaches, the two main objectives in biomarker discovery are met: 1) the identification of a small set of relevant indicators with minimum redundancy, and 2) the validation of the predictors using a classifier and cross-validation strategy. To balance false-positives and false-negatives in the selection of biomarkers, a clever method was proposed to use common peaks in mass spectrometry data as the predictive indicators (137). The procedure applies AdaBoost (a form of ensemble classifier training) to perform the classification and to select the informative common peaks.

Bioinformatics approaches to discover biomarkers can take on more sophisticated implementations. For instance, a dependence (interaction) network modeling scheme was suggested for identifying biomarkers from groups of genes or proteins (138). Very clear differences were observed in the dependence networks for cancer and non-cancer samples. On the other hand, a gene selection algorithm was used based on Gaussian processes to discover consistent gene expression patterns associated with ordinal clinical phenotypes (139). The method was able to identify subsets of genes as potential biomarkers for colon and prostate cancers. The integration of time-course microarray gene expression data with cytotoxicity measurements, by way of a partial least squares objective criterion, has been shown to be useful for identifying biomarkers in primary rat hepatocytes exposed to cadmium (140). The approach demonstrates the value of integrating omics data with associated biological data to glean more information about the biomarker's diagnostic utility. More recently, a bioinformatics approach was introduced that takes into account the inherent correlation of genes when using gene expression data for biomarker discovery (141). Finally, a host of techniques and software to integrate omics data are summarized, to shed additional light on the complex molecular interactions that take place on a systems biology level (142).

Repositories of omics data from various studies that can be queried may bring about improved means for detecting biomarkers of a clinical process or phenotype than one which is isolated or from a small group of data sets (143). This realization has motivated the generators of omics data to store them in repositories for meta-mining purposes. Figure 7.5 represents a brief list of some of the publicly accessible databases that store, distribute and permit querying of omics data (a more comprehensive list is presented in (142)). A plasma proteome database at the Institute of Bioinformatics that stores comparisons of human plasma protein concentration levels along with their isoforms in normal and disease states should be useful for discovery of novel biomarkers (144). Another database, ONCOMINE, stores a collection of curated cancer gene expression profiles integrated with a therapeutic target database and biological resources, such as Gene Ontology, so that the data can be mined for putative biomarkers (123).

The use of bioinformatics tools has increased mechanistic understanding and development of biomarkers in the analysis of massive genomics, proteomics and metabolomics data (Cf. Figure 7.4). Bioinformatics techniques will continue to be useful in organizing and extracting candidate biomarkers for chemical exposures and disease for epidemiology, clinical and experimental studies. However, mere access to sophisticated bioinformatics tools will be insufficient to grapple with

the identification of biomarkers from omics data (Figure 7.5). An ongoing and vigorous debate has emerged over the use and reproducibility of bioinformatics approaches and omics data for biomarker discovery and clinical applications (145,146). Clearly there is a need for rigorous quality control in the field of bioinformatics for the use of omics type data in clinical, diagnostic and regulatory settings (Lyle Burgoon, personal communication). A fundamental understanding of the inherent problems and issues with omics data, and knowing how, where and when to apply which type of bioinformatics approach, are essential to effectively translating omics biomarkers into clinically useful diagnostic tools and epidemiological markers.

Figure 7.5. Online bioinformatics resources for various omics fields, including genomics, transcriptomics, proteomics, and metabolomics

Data Type	Resource	Host	Description	URL
Genomics	Genomes OnLine Database (GOLD)	Genomesonline.org	Completed and ongoing genome projects	http://www.genomesonline.org/index
	GenBank	NIH-NCBI	Genetic sequence database	http://www.ncbi.nlm.nih.gov/Genbank
	EMBL-Bank	EMBL-EBI	Nucleotide sequence database	http://www.ebi.ac.uk/embl/
Transcriptomics	Gene Expression Omnibus (GEO)	NIH-NCBI	Gene expression/ molecular abundance repository supporting MIAME-compliant data	http://www.ncbi.nlm.nih.gov/geo
	ArrayExpress	EMBL-EBI	Repository for MIAME-compliant microarray data	http://www.ebi.ac.uk/arrayexpress
	ArrayTrack™	FDA-NCTR	An integrated solution for managing, analysing and interpreting MIAME-compliant microarray gene expression data from pharmaco- and toxicogenomics studies	http://www.fda.gov/ScienceResearch/BioinformaticsTools/Arraytrack/default.htm
	Stanford Microarray Database (SMD)	Stanford University	Microarray-based MIAME-compliant gene expression	http://genome-www.stanford.edu/microarray
	ONCOMINE	University of Michigan	Database for mining and examining gene expression in cancer	http://www.oncomine.org
	Environment, Drugs and Gene Expression (EDGE)	University of Wisconsin-Madison	Toxicology-related gene expression data for mapping transcriptional changes from chemical exposure	http://edge.oncology.wisc.edu/
Proteomics	World-2DPAGE	Expert Protein Analysis System (ExPASy)	A dynamic portal to query simultaneously worldwide proteomics databases	http://ca.expasy.org/world-2dpage
	Plasma Protein Database	Pandey Lab and Institute of Bioinformatics	A comprehensive resource for all human plasma proteins along with their isoforms	http://www.plasmaproteomedatabase.org

Data Type	Resource	Host	Description	URL
	Open Proteomics Database (OPD)	University of Texas at Austin	For storing and disseminating mass spectrometry-based proteomics data	http://apropos.icmb.utexas.edu/OPD
Metabolomics	METLIN Metabolite Database	The Scripps Research Institute	A repository for mass spectral metabolite data	http://metlin.scripps.edu
	Human Metabolome Database (HMDB)	Genome Alberta and Genome Canada	Database of small metabolites found in the human body containing or linking chemical, clinical and molecular biology/biochemistry data	http://www.hmdb.ca/
Integrated	Chemical Effects in Biological Systems (CEBS)	NIH-NIEHS	Integrates study design, clinical pathology and histopathology, gene expression and proteomics data from all studies and enables discrimination of critical study factors	http://cebs.niehs.nih.gov/
	PharmGKB	Stanford	Collects, encodes and disseminates knowledge about the impact of human genetic variations on drug responses. Provides curated primary genotype and phenotype data, annotated gene variants, and gene-drug-disease relationships via literature review, and summarizes important PGx genes and drug pathways	http://www.pharmgkb.org/
	Kyoto encyclopaedia of genes and geomics pathways (KEGG)	Kyoto University, Japan	KEGG is an integrated database that relates genes to metabolic pathways; outlines functional relationships among genes; disease-associated genes; pharmaceuticals-targeted genes	http://www.genome.jp/kegg

References

1. Mundy C (2001). The human genome project: a historical perspective. *Pharmacogenomics*, 2:37–49.doi:10.1517/14622416.2.1.37 PMID: 11258196

2. Thorisson GA, Smith AV, Krishnan L, Stein LD (2005). The International HapMap Project Web site. *Genome Res,* 15:1592–1593. doi:10.1101/gr.4413105 PMID:16251469

3. Liolios K, Tavernarakis N, Hugenholtz P, Kyrpides NC (2006). The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res,* 34 Database issue;D332–D334.doi:10.1093/nar/gkj145 PMID:16381880

4. Csako G (2006). Present and future of rapid and/or high-throughput methods for nucleic acid testing. *Clin Chim Acta,* 363:6–31.doi:10.1016/j.cccn.2005.07.009 PMID:161027 38

5. Royds JA, Iacopetta B (2006). p53 and disease: when the guardian angel fails. *Cell Death Differ,* 13:1017–1026.doi:10.1038/sj.cdd.4401913 PMID:16557268

6. Taylor AM, Byrd PJ (2005). Molecular pathology of ataxia telangiectasia. *J Clin Pathol,* 58:1009–1015.doi:10.1136/jcp.2005.026062 PMID:16189143

7. Nowrousian M (2010). Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems. *Eukaryot Cell,* 9:1300–1310.doi:10.1128/EC.00123-10 PMID: 20601439

8. Rando OJ (2007). Chromatin structure in the genomics era. *Trends Genet*, 23:67–73. doi:10.1016/j.tig.2006.12.002 PMID:17188397

9. Metzker ML (2010). Sequencing technologies - the next generation. *Nat Rev Genet,* 11:31–46.doi:10.1038/nrg2626 PMID: 19997069

10. Bentley DR (2006). Whole-genome re-sequencing. *Curr Opin Genet Dev,* 16:545–552.doi:10.1016/j.gde.2006.10.009 PMID:17055251

11. Bates S (2010). Progress towards personalized medicine. *Drug Discov Today*, 15:115–120.doi:10.1016/j.drudis.2009.11.001 PMID:19914397

12. Steemers FJ, Gunderson KL (2007). Whole genome genotyping technologies on the BeadArray platform. *Biotechnol J*, 2:41–49. doi:10.1002/biot.200600213 PMID:17225249

13. Ku CS, Loy EY, Pawitan Y, Chia KS (2010). The pursuit of genome-wide association studies: where are we now? *J Hum Genet,* 55:195–206.doi:10.1038/jhg.2010.19 PMID: 20300123

14. Middeldorp A, Jagmohan-Changur S, Helmer Q *et al.* (2007). A procedure for the detection of linkage with high density SNP arrays in a large pedigree with colorectal cancer. *BMC Cancer,* 7:6.doi:10.1186/1471-2407-7-6 PMID:17222328

15. Giacomini KM, Brett CM, Altman RB *et al.*; Pharmacogenetics Research Network (2007). The pharmacogenetics research network: from SNP discovery to clinical drug response. *Clin Pharmacol Ther,* 81:328–345. doi:10.1038/sj.clpt.6100087 PMID:17339863

16. Mateuca R, Lombaert N, Aka PV *et al.* (2006). Chromosomal changes: induction, detection methods and applicability in human biomonitoring. *Biochimie*, 88:1515–1531.doi:10.1016/j.biochi.2006.07.004 PMID:16919864

17. Gunderson KL, Steemers FJ, Ren H *et al.* (2006). Whole-genome genotyping. *Methods Enzymol,* 410:359–376.doi:10.1016/S0076-6879(06)10017-8 PMID:16938560

18. Pinard R, de Winter A, Sarkis GJ *et al.* (2006). Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics,* 7:216. doi:10.1186/1471-2164-7-216 PMID:16928277

19. Jongeneel CV, Iseli C, Stevenson BJ *et al.* (2003). Comprehensive sampling of gene expression in human cell lines with massively parallel signature sequencing. *Proc Natl Acad Sci USA,* 100:4702–4705.doi:10.1073/pnas.0831040100 PMID:12671075

20. Brown PO, Botstein D (1999). Exploring the new world of the genome with DNA microarrays. *Nat Genet,* 21 Suppl;33–37.doi: 10.1038/4462 PMID:9915498

21. Davis TN (2004). Protein localization in proteomics. *Curr Opin Chem Biol,* 8:49–53. doi:10.1016/j.cbpa.2003.11.003 PMID:15036156

22. Lockhart DJ, Dong H, Byrne MC *et al.* (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol,* 14:1675–1680.doi:10.1038/nbt1296-1675 PMID:9634850

23. Schena M, Shalon D, Davis RW, Brown PO (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467–470. doi:10.1126/science.270.5235.467 PMID:7569999

24. Fodor SP, Read JL, Pirrung MC *et al.* (1991). Light-directed, spatially addressable parallel chemical synthesis. *Science*, 251:767–773.doi:10.1126/science.1990438 PMID:1990438

25. Bammler T, Beyer RP, Bhattacharya S *et al.*; Members of the Toxicogenomics Research Consortium (2005). Standardizing global gene expression analysis between laboratories and across platforms. *Nat Methods,* 2:351–356.doi:10.1038/nmeth0605-477a PMID:15846362

26. Irizarry RA, Warren D, Spencer F *et al.* (2005). Multiple-laboratory comparison of microarray platforms. *Nat Methods,* 2:345–350.doi:10.1038/nmeth756 PMID:15846361

27. Ulrich RG, Rockett JC, Gibson GG, Pettit SD (2004). Overview of an interlaboratory collaboration on evaluating the effects of model hepatotoxicants on hepatic gene expression. *Environ Health Perspect,* 112:423–427. doi:10.1289/ehp.6675 PMID:15033591

28. Whitney AR, Diehn M, Popper SJ *et al.* (2003). Individuality and variation in gene expression patterns in human blood. *Proc Natl Acad Sci USA,* 100:1896–1901.doi:10.1073/pnas.252784499 PMID:12578971

29. Bennett L, Palucka AK, Arce E *et al.* (2003). Interferon and granulopoiesis signatures in systemic lupus erythematosus blood. *J Exp Med,* 197:711–723.doi:10.1084/jem.20021553 PMID:12642603

30. Borovecki F, Lovrecic L, Zhou J *et al.* (2005). Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease. *Proc Natl Acad Sci USA,* 102:11023–11028.doi:10.1073/pnas.0504921102 PMID:16043692

31. Burczynski ME, Twine NC, Dukart G *et al.* (2005). Transcriptional profiles in peripheral blood mononuclear cells prognostic of clinical outcomes in patients with advanced renal cell carcinoma. *Clin Cancer Res,* 11:1181–1189. PMID:15709187

32. Fannin RD, Auman JT, Bruno ME *et al.* (2005). Differential gene expression profiling in whole blood during acute systemic inflammation in lipopolysaccharide-treated rats. *Physiol Genomics,* 21:92–104. doi:10.1152/physiolgenomics.00190.2004 PMID:15781589

33. Heller RA, Schena M, Chai A *et al.* (1997). Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc Natl Acad Sci USA,* 94:2150–2155.doi:10.1073/pnas.94.6.2150 PMID:9122163

34. Liew CC, Dzau VJ (2004). Molecular genetics and genomics of heart failure. *Nat Rev Genet,* 5:811–825.doi:10.1038/nrg1470 PMID:15520791

35. Sullivan PF, Fan C, Perou CM (2006). Evaluating the comparability of gene expression in blood and brain. *Am J Med Genet* B Neuropsychiatr Genet, 141B:261–268. doi:10.1002/ajmg.b.30272 PMID:16526044

36. Tang Y, Lu A, Aronow BJ, Sharp FR (2001). Blood genomic responses differ after stroke, seizures, hypoglycemia, and hypoxia: blood genomic fingerprints of disease. *Ann Neurol,* 50:699–707.doi:10.1002/ana.10042 PMID:11761467

37. Argos M, Kibriya MG, Parvez F *et al.* (2006). Gene expression profiles in peripheral lymphocytes by arsenic exposure and skin lesion status in a Bangladeshi population. *Cancer Epidemiol Biomarkers Prev,* 15:1367–1375.doi:10.1158/1055-9965.EPI-06-0106 PMID:16835338

38. Bushel PR, Heinloth AN, Li J *et al.* (2007). Blood gene expression signatures predict exposure levels. *Proc Natl Acad Sci USA,* 104:18211–18216.doi:10.1073/pnas.0706987104 PMID:17984051

39. Forrest MS, Lan Q, Hubbard AE *et al.* (2005). Discovery of novel biomarkers by microarray analysis of peripheral blood mononuclear cell gene expression in benzene-exposed workers. *Environ Health Perspect,* 113:801–807.doi:10.1289/ehp.7635 PMID:15929907

40. Lampe JW, Stepaniants SB, Mao M *et al.* (2004). Signatures of environmental exposures using peripheral leukocyte gene expression: tobacco smoke. *Cancer Epidemiol Biomarkers Prev,* 13:445–453. PMID:15006922

41. Fannin RD, Russo M, O'Connell TM *et al.* (2010). Acetaminophen dosing of humans results in blood transcriptome and metabolome changes consistent with impaired oxidative phosphorylation. *Hepatology*, 51:227–236. PMID:19918972

42. Afshari CA, Nuwaysir EF, Barrett JC (1999). Application of complementary DNA microarray technology to carcinogen identification, toxicology, and drug safety evaluation. *Cancer Res,* 59:4759–4760. PMID:10519378

43. Zilberman D, Henikoff S (2007). Genome-wide analysis of DNA methylation patterns. *Development*, 134:3959–3965.doi:10.1242/dev.001131 PMID:17928417

44. Esteller M (2008). Epigenetics in cancer. *N Engl J Med,* 358:1148–1159.doi:10.1056/NEJMra072067 PMID:18337604

45. Hoheisel JD (2006). Microarray technology: beyond transcript profiling and genotype analysis. *Nat Rev Genet,* 7:200–210.doi:10.1038/nrg1809 PMID:16485019

46. McCann JA, Muro EM, Palmer C *et al.* (2007). ChIP on SNP-chip for genome-wide analysis of human histone H4 hyperacetylation. *BMC Genomics,* 8:322.doi:10.1186/1471-2164-8-322 PMID:17868463

47. Beaudet AL, Belmont JW (2008). Array-based DNA diagnostics: let the revolution begin. *Annu Rev Med,* 59:113–129. doi:10.1146/annurev.med.59.012907.101800 PMID:17961075

48. Hardiman G (2008). Applications of microarrays and biochips in pharmacogenomics. *Methods Mol Biol,* 448: 21–30.doi:10.1007/978-1-59745-205-2_2 PMID:18370228

49. Allison M (2008). Is personalized medicine finally arriving? *Nat Biotechnol,* 26:509–517. doi:10.1038/nbt0508-509 PMID:18464779

50. Vineis P, Perera FP (2007). Molecular epidemiology and biomarkers in etiologic cancer research: the new in light of the old. *Cancer Epidemiol Biomarkers Prev,* 16:1954–1965.doi:10.1158/1055-9965.EPI-07-0457 PMID:17932342

51. Wild CP (2009). Environmental exposure measurement in cancer epidemiology. *Mutagenesis*, 24:117–125.doi:10.1093/mutage/gen061 PMID:19033256

52. Wetmore BA, Merrick BA (2004). Toxicoproteomics: proteomics applied to toxicology and pathology. *Toxicol Pathol,* 32:619–642.doi:10.1080/01926230490518244 PMID:15580702

53. Kline KG, Sussman MR (2010). Protein quantitation using isotope-assisted mass spectrometry. *Annu Rev Biophys*, 39:291–308. doi:10.1146/annurev.biophys.093008.131339 PMID:20462376

54. Sanchez-Carbayo M (2010). Antibody array-based technologies for cancer protein profiling and functional proteomic analyses using serum and tissue specimens. *Tumour Biol,* 31:103–112.doi:10.1007/s13277-009-0014-z PMID:20358423

55. Righetti PG, Castagna A, Antonucci F *et al.* (2004). Critical survey of quantitative proteomics in two-dimensional electrophoretic approaches. *J Chromatogr A,* 1051:3–17.doi:10.1016/j.chroma.2004.05.106 PMID:15532550

56. Yates JR 3rd (2004). Mass spectral analysis in proteomics. *Annu Rev Biophys Biomol Struct,* 33:297–316.doi:10.1146/annurev.biophys.33.111502.082538 PMID:15139815

57. Macdonald N, Chevalier S, Tonge R *et al.* (2001). Quantitative proteomic analysis of mouse liver response to the peroxisome proliferator diethylhexylphthalate (DEHP). *Arch Toxicol,* 75:415–424.doi:10.1007/s002040100259 PMID:11693183

58. Neubert H, Bonnert TP, Rumpel K *et al.* (2008). Label-free detection of differential protein expression by LC/MALDI mass spectrometry. *J Proteome Res,* 7:2270–2279. doi:10.1021/pr700705u PMID:18412385

59. Ong SE, Pandey A (2001). An evaluation of the use of two-dimensional gel electrophoresis in proteomics. *Biomol Eng,* 18:195–205.doi:10.1016/S1389-0344(01)00095-8 PMID:11911086

60. Turner SM (2006). Stable isotopes, mass spectrometry, and molecular fluxes: applications to toxicology. *J Pharmacol Toxicol Methods,* 53:75–85.doi:10.1016/j.vascn.2005.08.001 PMID:16213756

61. Mueller LN, Brusniak MY, Mani DR, Aebersold R (2008). An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. *J Proteome Res,* 7:51–61.doi:10.1021/pr700758r PMID:18173218

62. Petricoin EF, Liotta LA (2004). SELDI-TOF-based serum proteomic pattern diagnostics for early detection of cancer. *Curr Opin Biotechnol,* 15:24–30.doi:10.1016/j.copbio.2004.01.005 PMID:15102462

63. Engwegen JY, Gast MC, Schellens JH, Beijnen JH (2006). Clinical proteomics: searching for better tumour markers with SELDI-TOF mass spectrometry. *Trends Pharmacol Sci,* 27:251–259.doi:10.1016/j.tips.2006.03.003 PMID:16600386

64. Cutler P (2003). Protein arrays: the current state-of-the-art. *Proteomics*, 3:3–18. doi:10.1002/pmic.200390007 PMID:12548629

65. Sheehan KM, Calvert VS, Kay EW *et al.* (2005). Use of reverse phase protein microarrays and reference standard development for molecular network analysis of metastatic ovarian carcinoma. *Mol Cell Proteomics,* 4:346–355.doi:10.1074/mcp.T500003-MCP200 PMID:15671044

66. Schwenk JM, Lindberg J, Sundberg M *et al.* (2007). Determination of binding specificities in highly multiplexed bead-based assays for antibody proteomics. *Mol Cell Proteomics,* 6:125–132. PMID:17060675

67. Daemen A, Gevaert O, De Bie T *et al.* (2008). Integrating microarray and proteomics data to predict the response on cetuximab in patients with rectal cancer. *Pac Symp Biocomput,* 166–177. PMID:18229684

68. Vermeulen R, Lan Q, Zhang L *et al.* (2005). Decreased levels of CXC-chemokines in serum of benzene-exposed workers identified by array-based proteomics. *Proc Natl Acad Sci USA,* 102:17041–17046.doi:10.1073/pnas.0508573102 PMID:16286641

69. Yang M, Berhane Y, Salo T *et al.* (2008). Development and application of monoclonal antibodies against avian influenza virus nucleoprotein. *J Virol Methods,* 147:265–274. doi:10.1016/j.jviromet.2007.09.016 PMID:18006085

70. Cho WC (2007). Proteomics technologies and challenges. *Genomics Proteomics Bioinformatics,* 5:77–85.doi:10.1016/S1672-0229(07)60018-7 PMID:17893073

71. Chung JY, Braunschweig T, Tuttle K, Hewitt SM (2007). Tissue microarrays as a platform for proteomic investigation. *J Mol Histol,* 38:123–128.doi:10.1007/s10735-006-9049-2 PMID:16953460

72. Merrick BA (2008). The plasma proteome, adductome and idiosyncratic toxicity in toxicoproteomics research. *Brief Funct Genomic Proteomic,* 7:35–49.doi:10.1093/bfgp/eln004 PMID:18270218

73. Kumar S, Mohan A, Guleria R (2006). Biomarkers in cancer screening, research and detection: present and future: a review. *Biomarkers*, 11:385–405.doi:10.1080/13547500600775011 PMID:16966157

74. Lemley KV (2007). An introduction to biomarkers: *applications to chronic kidney disease. Pediatr Nephrol,* 22:1849–1859. doi:10.1007/s00467-007-0455-9 PMID:17394023

75. Merrick BA, Bruno ME (2004). Genomic and proteomic profiling for biomarkers and signature profiles of toxicity. *Curr Opin Mol Ther,* 6:600–607. PMID:15663324

76. Azmi J, Connelly J, Holmes E *et al.* (2005). Characterization of the biochemical effects of 1-nitronaphthalene in rats using global metabolic profiling by NMR spectroscopy and pattern recognition. *Biomarkers*, 10:401–416. doi:10.1080/13547500500309259 PMID:16308265

77. Walsh MC, Brennan L, Malthouse JP *et al.* (2006). Effect of acute dietary standardization on the urinary, plasma, and salivary metabolomic profiles of healthy humans. *Am J Clin Nutr,* 84:531–539. PMID:16960166

78. Chace DH, Kalas TA, Naylor EW (2002). The application of tandem mass spectrometry to neonatal screening for inherited disorders of intermediary metabolism. *Annu Rev Genomics Hum Genet,* 3:17–45.doi: 10.1146/annurev.genom.3.022502.103213 PMID:12142359

79. Sim KG, Hammond J, Wilcken B (2002). Strategies for the diagnosis of mitochondrial fatty acid beta-oxidation disorders. *Clin Chim Acta,* 323:37–58.doi:10.1016/S0009-8981(02)00182-1 PMID:12135806

80. Want EJ, O'Maille G, Smith CA *et al.* (2006). Solvent-dependent metabolite distribution, clustering, and protein extraction for serum profiling with mass spectrometry. *Anal Chem,* 78:743–752.doi:10.1021/ac051312t PMID:16448047

81. Gamache PH, Meyer DF, Granger MC, Acworth IN (2004). Metabolomic applications of electrochemistry/mass spectrometry. *J Am Soc Mass Spectrom,* 15:1717–1726. doi:10.1016/j.jasms.2004.08.016 PMID:15589749

82. Kaderbhai NN, Broadhurst DI, Ellis DI *et al.* (2003). Functional genomics via metabolic footprinting: monitoring metabolite secretion by Escherichia coli tryptophan metabolism mutants using FT-IR and direct injection electrospray mass spectrometry. *Comp Funct Genomics,* 4:376–391.doi:10.1002/cfg.302 PMID:18629082

83. Lindon JC, Holmes E, Nicholson JK (2006). Metabonomics techniques and applications to pharmaceutical research & development. *Pharm Res,* 23:1075–1088.doi:10.1007/s11095-006-0025-z PMID:16715371

84. Nicholson JK, Connelly J, Lindon JC, Holmes E (2002). Metabonomics: a platform for studying drug toxicity and gene function. *Nat Rev Drug Discov,* 1:153–161.doi:10.1038/nrd728 PMID:12120097

85. Shockcor JP, Holmes E (2002). Metabonomic applications in toxicity screening and disease diagnosis. *Curr Top Med Chem,* 2:35–51.doi:10.2174/1568026023394498 PMID:11899064

86. Dunn S, Simenhoff M, Ahmed K *et al.* (1998). Effect of oral administration of freeze-dried Lactobacillus acidophilus on small bowel bacterial overgrowth in patients with end stage kidney disease: reducing uremic toxins and improving nutrition. *Int Dairy J*, 8:545–553 doi:10.1016/S0958-6946(98)00081-8.

87. London R, Houck D. Introduction to metabolomics and metabolic profiling. In: Hamadeh HK, Afshari CA, editors. Toxicogenomics: principles and applications. Hoboken (NJ): Wiley-LSS; 2004. p. 299–340.

88. Chalmers RA, Valman HB, Liberman MM (1979). Measurement of 4-hydroxyphenylacetic aciduria as a screening test for small-bowel disease. *Clin Chem,* 25:1791–1794. PMID:476929

89. Ghauri FY, McLean AE, Beales D *et al.* (1993). Induction of 5-oxoprolinuria in the rat following chronic feeding with N-acetyl 4-aminophenol (paracetamol). *Biochem Pharmacol,* 46:953–957.doi:10.1016/0006-2952(93)90506-R PMID:8373447

90. Soga T, Baran R, Suematsu M *et al.* (2006). Differential metabolomics reveals ophthalmic acid as an oxidative stress biomarker indicating hepatic glutathione consumption. *J Biol Chem,* 281:16768–16776.doi:10.1074/jbc.M601876200 PMID:16608839

91. Yeung KY, Ruzzo WL (2001). Principal component analysis for clustering gene expression data. *Bioinformatics*, 17:763–774. doi:10.1093/bioinformatics/17.9.763 PMID:11590094

92. Abe T, Isobe C, Murata T *et al.* (2003). Alteration of 8-hydroxyguanosine concentrations in the cerebrospinal fluid and serum from patients with Parkinson's disease. *Neurosci Lett,* 336:105–108.doi:10.1016/S0304-3940(02)01259-4 PMID:12499051

93. Abe T, Tohgi H, Isobe C *et al.* (2002). Remarkable increase in the concentration of 8-hydroxyguanosine in cerebrospinal fluid from patients with Alzheimer's disease. *J Neurosci Res,* 70:447–450.doi:10.1002/jnr.10349 PMID:12391605

94. Nekrassova O, Lawrence NS, Compton RG (2003). Analytical determination of homocysteine: a review. *Talanta*, 60:1085–1095.doi:10.1016/S0039-9140(03)00173-5 PMID:18969134

95. Chen YQ, Duniec ZM, Liu B *et al.* (1994). Endogenous 12(S)-HETE production by tumor cells and its role in metastasis. *Cancer Res*, 54:1574–1579. PMID:7511046

96. Liu B, Marnett LJ, Chaudhary A *et al.* (1994). Biosynthesis of 12(S)-hydroxyeicosatetraenoic acid by B16 amelanotic melanoma cells is a determinant of their metastatic potential. *Lab Invest,* 70:314–323. PMID:8145526

97. Winer I, Normolle DP, Shureiqi I *et al.* (2002). Expression of 12-lipoxygenase as a biomarker for melanoma carcinogenesis. *Melanoma Res,* 12:429–434.doi:10.1097/00008390-200209000-00003 PMID:12394183

98. Aiken NR, Gillies RJ (1996). Phosphomonoester metabolism as a function of cell proliferative status and exogenous precursors. *Anticancer Res,* 16 3B;1393–1397. PMID:8694507

99. Singer S, Souza K, Thilly WG (1995). Pyruvate utilization, phosphocholine and adenosine triphosphate (ATP) are markers of human breast tumor progression: a 31P- and 13C-nuclear magnetic resonance (NMR) spectroscopy study. *Cancer Res,* 55:5140–5145. PMID:7585561

100. Chen H, Pan Z, Talaty N *et al.* (2006). Combining desorption electrospray ionization mass spectrometry and nuclear magnetic resonance for differential metabolomics without sample preparation. *Rapid Commun Mass Spectrom,* 20:1577–1584.doi:10.1002/rcm.2474 PMID:16628593

101. Süllentrop F, Moka D, Neubauer S *et al.* (2002). 31P NMR spectroscopy of blood plasma: determination and quantification of phospholipid classes in patients with renal cell carcinoma. *NMR Biomed,* 15:60–68. doi:10.1002/nbm.758 PMID:11840554

102. Frickenschmidt A, Frohlich H, Bullinger D *et al.* (2008). Metabonomics in cancer diagnosis: mass spectrometry-based profiling of urinary nucleosides from breast cancer patients. *Biomarkers*, 13:435–449. doi:10.1080/13547500802012858 PMID:18484357

103. Ohdoi C, Nyhan WL, Kuhara T (2003). Chemical diagnosis of Lesch-Nyhan syndrome using gas chromatography-mass spectrometry detection. *J Chromatogr B Analyt Technol Biomed Life Sci*, 792:123–130. doi:10.1016/S1570-0232(03)00277-0 PMID:12829005

104. Seibel MJ, Gartenberg F, Silverberg SJ *et al.* (1992). Urinary hydroxypyridinium cross-links of collagen in primary hyperparathyroidism. *J Clin Endocrinol Metab,* 74:481–486.doi:10.1210/jc.74.3.481 PMID:1740480

105. Gineyts E, Garnero P, Delmas PD (2001). Urinary excretion of glucosyl-galactosyl pyridinoline: a specific biochemical marker of synovium degradation. *Rheumatology (Oxford),* 40:315–323.doi:10.1093/rheumatology/40.3.315 PMID:11285380

106. Gineyts E, Mo JA, Ko A *et al.* (2004). Effects of ibuprofen on molecular markers of cartilage and synovium turnover in patients with knee osteoarthritis. *Ann Rheum Dis,* 63:857–861.doi:10.1136/ard.2003.007302 PMID:15194584

107. MacAllister RJ, Whitley GSJ, Vallance P (1994). Effects of guanidino and uremic compounds on nitric oxide pathways. *Kidney Int,* 45:737–742.doi:10.1038/ki.1994.98 PMID:7515129

108. Zoccali C, Bode-Böger SM, Mallamaci F *et al.* (2001). Plasma concentration of asymmetrical dimethylarginine and mortality in patients with end-stage renal disease: a prospective study. *Lancet*, 358:2113–2117.doi:10.1016/S0140-6736(01)07217-8 PMID:11784625

109. Miyazaki H, Matsuoka H, Cooke JP *et al.* (1999). Endogenous nitric oxide synthase inhibitor: a novel marker of atherosclerosis. *Circulation*, 99:1141–1146. PMID:10069780

110. Salek RM, Maguire ML, Bentley E *et al.* (2007). A metabolomic comparison of urinary changes in type 2 diabetes in mouse, rat, and human. *Physiol Genomics*, 29:99–108. PMID:17190852

111. Ringeissen S, Connor SC, Brown HR *et al.* (2003). Potential urinary and plasma biomarkers of peroxisome proliferation in the rat: identification of N-methylnicotinamide and N-methyl-4-pyridone-3-carboxamide by 1H nuclear magnetic resonance and high performance liquid chromatography. *Biomarkers*, 8:240–271.doi:10.1080/1354750031000149124 PMID:12944176

112. Connor SC, Hodson MP, Ringeissen S *et al.* (2004). Development of a multivariate statistical model to predict peroxisome proliferation in the rat, based on urinary 1H-NMR spectral patterns. *Biomarkers*, 9:364–385.doi:10.1080/13547500400006005 PMID:15764299

113. Chalmers RA, Jones MG, Goodwin CS, Amjad S (2006). CFSUM1 and CFSUM2 in urine from patients with chronic fatigue syndrome are methodological artefacts. *Clin Chim Acta,* 364:148–158.doi:10.1016/j.cccn.2005.05.036 PMID:16095585

114. Jones MG, Cooper E, Amjad S *et al.* (2005). Urinary and plasma organic acids and amino acids in chronic fatigue syndrome. *Clin Chim Acta,* 361:150–158.doi:10.1016/j.cccn.2005.05.023 PMID:15992788

115. Brindle JT, Antti H, Holmes E *et al.* (2002). Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using 1H-NMR-based metabonomics. *Nat Med,* 8:1439–1445.doi:10.1038/nm802 PMID:12447357

116. Kirschenlohr HL, Griffin JL, Clarke SC *et al.* (2006). Proton NMR analysis of plasma is a weak predictor of coronary artery disease. *Nat Med,* 12:705–710.doi:10.1038/nm1432 PMID:16732278

117. Dunn WB (2008). Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes. *Phys Biol,* 5:011001.doi:10.1088/1478-3975/5/1/011001 PMID:18367780

118. Luscombe NM, Greenbaum D, Gerstein M (2001). What is bioinformatics? A proposed definition and overview of the field. *Methods Inf Med,* 40:346–358. PMID:11552348

119. Mount DW, Pandey R (2005). Using bioinformatics and genome analysis for new therapeutic interventions. *Mol Cancer Ther,* 4:1636–1643.doi:10.1158/1535-7163.MCT-05-0150 PMID:16227414

120. Li J, Zhang Z, Rosenzweig J *et al.* (2002). Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clin Chem,* 48:1296–1304. PMID:12142387

121. Azuaje F. Bioinformatics and biomarker discovery: "omic" data analysis for personalized medicine. Hoboken (NJ): Wiley-Blackwell; 2010.

122. Ginsburg GS, Haga SB (2006). Translating genomic biomarkers into clinically useful diagnostics. *Expert Rev Mol Diagn,* 6:179–191.doi:10.1586/14737159.6.2.179 PMID:16512778

123. Rhodes DR, Yu J, Shanker K *et al.* (2004). ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia*, 6:1–6. PMID:15068665

124. Baggerly KA, Coombes KR, Morris JS (2005). Bias, randomization, and ovarian proteomic data: a reply to "producers and consumers.". Cancer Inform, 1:9–14.

125. Liu J, Zheng S, Yu JK *et al.* (2005). Serum protein fingerprinting coupled with artificial neural network distinguishes glioma from healthy population or brain benign tumor. *J Zhejiang Univ Sci B,* 6:4–10.doi:10.1631/jzus.2005.B0004 PMID:15593384

126. Yang YH, Dudoit S, Luu P *et al.* (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res,* 30:e15.doi:10.1093/nar/30.4.e15 PMID:11842121

127. Irizarry RA, Hobbs B, Collin F *et al.* (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4:249–264.doi:10.1093/biostatistics/4.2.249 PMID:12925520

128. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19:185–193. doi:10.1093/bioinformatics/19.2.185 PMID:12538238

129. Chou JW, Paules RS, Bushel PR (2005). Systematic variation normalization in microarray data to get gene expression comparison unbiased. *J Bioinform Comput Biol,* 3:225–241.doi:10.1142/S0219720005001028 PMID:15852502

130. Petricoin EF 3rd, Ardekani AM, Hitt BA *et al.* (2002). Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, 359:572–577.doi:10.1016/S0140-6736(02)07746-2 PMID:11867112

131. Tan CS, Ploner A, Quandt A *et al.* (2006). Finding regions of significance in SELDI measurements for identifying protein biomarkers. *Bioinformatics*, 22:1515–1523. doi:10.1093/bioinformatics/btl106 PMID:16567365

132. Shi L, Campbell G, Jones WD *et al.*; MAQC Consortium (2010). The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat Biotechnol,* 28:827–838.doi:10.1038/nbt.1665 PMID:20676074

133. Perou CM, Sørlie T, Eisen MB *et al.* (2000). Molecular portraits of human breast tumours. *Nature*, 406:747–752.doi:10.1038/35021093 PMID:10963602

134. Li L, Umbach DM, Terry P, Taylor JA (2004). Application of the GA/KNN method to SELDI proteomics data. *Bioinformatics*, 20:1638–1640.doi:10.1093/bioinformatics/bth098 PMID:14962943

135. Li L, Weinberg CR, Darden TA, Pedersen LG (2001). Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics*, 17:1131–1142.doi:10.1093/bioinformatics/17.12.1131 PMID:11751221

136. Peng Y (2006). A novel ensemble machine learning for robust microarray data classification. *Comput Biol Med,* 36:553–573. doi:10.1016/j.compbiomed.2005.04.001 PMID:15978569

137. Fushiki T, Fujisawa H, Eguchi S (2006). Identification of biomarkers from mass spectrometry data using a "common" peak approach. *BMC Bioinformatics,* 7:358.doi:10.1186/1471-2105-7-358 PMID:16869977

138. Qiu P, Wang ZJ, Liu KJ *et al.* (2007). Dependence network modeling for biomarker identification. *Bioinformatics*, 23:198–206. doi:10.1093/bioinformatics/btl553 PMID:17077095

139. Chu W, Ghahramani Z, Falciani F, Wild DL (2005). Biomarker discovery in microarray gene expression data with Gaussian processes. *Bioinformatics*, 21:3385–3393. doi:10.1093/bioinformatics/bti526 PMID:15937031

140. Tan Y, Shi L, Hussain SM *et al.* (2006). Integrating time-course microarray gene expression profiles with cytotoxicity for identification of biomarkers in primary rat hepatocytes exposed to cadmium. *Bioinformatics,* 22:77–87.doi:10.1093/bioinformatics/bti737 PMID:16249259

141. Zuber V, Strimmer K (2009). Gene ranking and biomarker discovery under correlation. *Bioinformatics,* 25:2700–2707.doi:10.1093/bioinformatics/btp460 PMID:19648135

142. Joyce AR, Palsson BO (2006). The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol,* 7:198–210. doi:10.1038/nrm1857 PMID:16496022

143. Waters M, Stasiewicz S, Merrick BA *et al.* (2008). CEBS–Chemical Effects in Biological Systems: a public data repository integrating study design and toxicity data with microarray and proteomics data. *Nucleic Acids Res,* 36 Database;D892–D900.doi:10.1093/nar/gkm755 PMID:17962311

144. Muthusamy B, Hanumanthu G, Suresh S *et al.* (2005). Plasma Proteome Database as a resource for proteomics research. *Proteomics*, 5:3531–3536.doi:10.1002/pmic.200401335 PMID:16041672

145. Amin RP, Vickers AE, Sistare F *et al.* (2004). Identification of putative gene based markers of renal toxicity. *Environ Health Perspect,* 112:465–479.doi:10.1289/ehp.6683 PMID:15033597

146. Searfoss GH, Jordan WH, Calligaro DO *et al.* (2003). Adipsin, a biomarker of gastrointestinal toxicity mediated by a functional gamma-secretase inhibitor. *J Biol Chem,* 278:46107–46116.doi:10.1074/jbc.M307757200 PMID:12949072

147. McDonough JL, Arrell DK, Van Eyk JE (1999). Troponin I degradation and covalent complex formation accompanies myocardial ischemia/reperfusion injury. *Circ Res,* 84:9–20. PMID:9915770

148. Oh P, Li Y, Yu J *et al.* (2004). Subtractive proteomic mapping of the endothelial surface in lung and solid tumours for tissue-specific therapy. *Nature,* 429:629–635.doi:10.1038/nature02580 PMID:15190345

149. Sawada H, Takami K, Asahi S (2005). A toxicogenomic approach to drug-induced phospholipidosis: analysis of its induction mechanism and establishment of a novel in vitro screening system. *Toxicol Sci,* 83:282–292.doi:10.1093/toxsci/kfh264 PMID:15342952

150. Tohgi H, Abe T, Yamazaki K *et al.* (1999). Remarkable increase in cerebrospinal fluid 3-nitrotyrosine in patients with sporadic amyotrophic lateral sclerosis. *Ann Neurol,* 46:129–131.doi:10.1002/1531-8249(199907)46:1<129::AID-ANA21>3.0.CO;2-Y PMID:10401792