

# Strategies for discovering novel cancer biomarkers through utilization of emerging technologies

Vathany Kulasingam and Eleftherios P Diamandis\*

## SUMMARY

The introduction of technologies such as mass spectrometry and protein and DNA arrays, combined with our understanding of the human genome, has enabled simultaneous examination of thousands of proteins and genes in single experiments, which has led to renewed interest in discovering novel biomarkers for cancer. The modern technologies are capable of performing parallel analyses as opposed to the serial analyses conducted with older methods, and they therefore provide opportunities to identify distinguishing patterns (signatures or portraits) for cancer diagnosis and classification as well as to predict response to therapies. Furthermore, these technologies provide the means by which new, single tumor markers could be discovered through use of reasonable hypotheses and novel analytical strategies. Despite the current optimism, a number of important limitations to the discovery of novel single tumor markers have been identified, including study design bias, and artefacts related to the collection and storage of samples. Despite the fact that new technologies and strategies often fail to identify well-established cancer biomarkers and show a bias toward the identification of high-abundance molecules, these technological advances have the capacity to revolutionize biomarker discovery. It is now necessary to focus on careful validation studies in order to identify the strategies and biomarkers that work and bring them to the clinic as early as possible.

**KEYWORDS** mass spectrometry, microarrays, multiparametric, proteomics, tumor markers

## REVIEW CRITERIA

The information for this Review was compiled by searching the PubMed database for articles published up until 6 August 2007. Electronic early-release publications were included. Only articles published in English were considered. The search terms included "tumor markers" in association with the following search terms: "reviews", "mass spectrometry", "protein arrays", "gene expression profiling", "proteomics", "molecular markers", "cancer biomarker guidelines", "peptidomics" and "microarrays". When possible, primary sources have been quoted.

## CME

V Kulasingam is a postdoctoral trainee in clinical biochemistry, and EP Diamandis is Professor and Head of Clinical Biochemistry, Department of Laboratory Medicine and Pathobiology, University of Toronto, University Health Network and Toronto Medical Laboratories and Mount Sinai Hospital, Toronto, ON, Canada.

## Correspondence

\*Department of Pathology and Laboratory Medicine, Mount Sinai Hospital, 600 University Avenue, Toronto, ON M5G 1X5, Canada  
ediamandis@mtsina.on.ca

Received 4 October 2007 Accepted 16 April 2008 Published online 12 August 2008

www.nature.com/clinicalpractice  
doi:10.1038/ncponc1187

## Medscape Continuing Medical Education online

Medscape, LLC is pleased to provide online continuing medical education (CME) for this journal article, allowing clinicians the opportunity to earn CME credit. Medscape, LLC is accredited by the Accreditation Council for Continuing Medical Education (ACCME) to provide CME for physicians. Medscape, LLC designates this educational activity for a maximum of 1.0 **AMA PRA Category 1 Credits™**. Physicians should only claim credit commensurate with the extent of their participation in the activity. All other clinicians completing this activity will be issued a certificate of participation. To receive credit, please go to <http://www.medscape.com/cme/ncp> and complete the post-test.

## Learning objectives

Upon completion of this activity, participants should be able to:

- 1 Identify how cancer biomarkers are best applied to clinical care.
- 2 Describe the impact of biomarkers on specific types of cancer.
- 3 Describe the process and applicability of gene expression profiling.
- 4 List potential advantages of mass spectrometry-based proteomic profiling.

## Competing interests

The authors and the Journal Editor L Hutchinson declared no competing interests. The CME questions author CP Vega declared that he has served as an advisor or consultant to Novartis, Inc.

## INTRODUCTION

Cancer continues to be a major cause of morbidity and mortality among men and women. In the US in 2006, over 1.4 million new cases of cancer were diagnosed and over half a million people died from this disease; the disease accounts for approximately 25% of all deaths in the US each year.<sup>1</sup> With increasing life expectancy, the prevalence of many cancers will probably increase. Early detection of various forms of cancer before they spread and become incurable is an important incentive for physicians and research scientists.<sup>2</sup> One of the best ways to diagnose cancer early, aid in its prognosis, or predict therapeutic response, is to use serum or tissue biomarkers.

Cancer biomarkers can be DNA, mRNA, proteins, metabolites, or processes such as apoptosis, angiogenesis or proliferation.<sup>3</sup> The markers

are produced either by the tumor itself or by other tissues, in response to the presence of cancer or other associated conditions, such as inflammation. Such biomarkers can be found in a variety of fluids, tissues and cell lines. Tumor markers can be used for screening the general population, for differential diagnosis in symptomatic patients, and for clinical staging of cancer. Additionally, these markers can be used to estimate tumor volume, to evaluate response to treatment, to assess disease recurrence through monitoring, or as prognostic indicators of disease progression (Box 1). Given the low prevalence of cancer in any given population, no marker has yet been discovered that meets all of these criteria.

A number of different types and forms of tumor markers exist. These markers include hormones, as well as different functional subgroups of proteins such as enzymes, glycoproteins, oncofetal antigens and receptors. Furthermore, other changes in tumors, such as genetic mutations, amplifications or translocations, and changes in microarray-generated profiles (genetic signatures), are also forms of tumor markers. Regardless of the type of tumor marker or profile, the use of a tumor marker must be associated with proven improvements in patient outcomes, such as increased survival or enhanced quality of life, in order to be substantiated.<sup>3</sup> An ideal tumor marker should be able to be measured easily, reliably and cost-effectively by use of an assay with high analytical sensitivity and specificity (Box 2). A caveat concerning currently used tumor markers is that, generally, they suffer from low diagnostic specificity and sensitivity (Table 1). Only a few markers have entered routine use, and then only for a limited number of cancer types and clinical settings. In the majority of cases, the current markers are used in conjunction with imaging, biopsy and associated clinicopathological information before a clinical decision is made.

The first cancer marker ever reported was the light chain of immunoglobulin in the urine, as identified in 75% of patients with myeloma in an 1848 study.<sup>4</sup> The test for this marker is still employed by clinicians today, but with use of modern quantification techniques. From 1930 to 1960, scientists identified numerous hormones, enzymes and other proteins, the concentration of which was altered in biological fluids from patients with cancer. The modern era of monitoring malignant disease, however, began in the 1960s with the discovery of alfa-fetoprotein<sup>5</sup> and

### Box 1 Definitions and specifications of biomarkers.

#### Diagnostic (screening) biomarker

A marker that is used to detect and identify a given type of cancer in an individual. These markers are expected to have high specificity and sensitivity; for example, the presence of Bence-Jones protein in urine remains one of the strongest diagnostic indicators of multiple myeloma.

#### Prognostic biomarker

This type of marker is used once the disease status has been established. These biomarkers are expected to predict the probable course of the disease including its recurrence, and they therefore have an important influence on the aggressiveness of therapy. For example, in testicular teratoma, human chorionic gonadotropin and alfa-fetoprotein levels can discriminate two groups with different survival rates.

#### Stratification (predictive) biomarker

This type of marker serves to predict the response to a drug before treatment is started. This marker classifies individuals as likely responders or nonresponders to a particular treatment. These biomarkers mainly arise from array-type experiments that make it possible to predict clinical outcome from the molecular characteristics of a patient's tumor.

#### Specificity

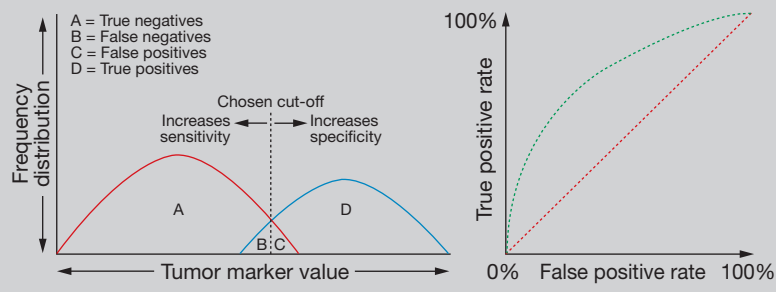
The proportion of control (normal) individuals who test negative for the biomarker.

#### Sensitivity

The proportion of individuals with confirmed disease who test positive for the biomarker.

#### Receiver operating characteristic (ROC) curve

A graphical representation of the relationship between sensitivity and specificity. This curve is used to evaluate the efficacy of a tumor marker at various cut-off points. An ideal graph is the one giving the maximum area under the curve (AUC). In the given example, the red curve represents a useless test (AUC = 0.5). The green curve represents a useful (AUC < 1.00) but not perfect (AUC = 1.00) test.



carcinoembryonic antigen (CEA),<sup>6</sup> which was facilitated by the introduction of immunological techniques such as the radioimmunoassay. In the 1980s, the era of hybridoma technology enabled development of the ovarian epithelial cancer marker carbohydrate antigen (CA) 125.<sup>7</sup> In 1980, prostate-specific antigen (PSA [KLK3]), considered one of the best cancer markers, was discovered.<sup>8</sup> Table 2 summarizes some currently used markers and their clinical utility.

Every era of biomarker discovery seems to be associated closely with the emergence of a new and powerful analytical technology. The past

**Box 2** Factors that are ideal for a serological tumor marker.

- Produced by the tumor cells and enters the circulation
- Present at low levels in the serum of healthy individuals and those with benign disease but increases substantially in cancer (preferably in one cancer type only)
- Easily quantifiable with an inexpensive assay
- Present in detectable (or higher than normal) quantities at early or preclinical stages
- Quantitative levels of the tumor marker reflect the tumor burden
- High diagnostic sensitivity (few false negatives) and specificity (few false positives)

decade has witnessed an impressive growth in the field of large-scale and high-throughput biology, which has contributed to an era of new technology development. The completion of a number of genome-sequencing projects, the discovery of oncogenes and tumor-suppressor genes, and recent advances in genomic and proteomic technologies, together with powerful bioinformatics tools, will have a direct and major impact on the way the search for cancer biomarkers is conducted. Early discoveries of cancer biomarkers were based mainly on empirical observations, such as the overexpression of CEA. The modern technologies are capable of performing parallel rather than serial analyses, and they can help to identify distinguishing patterns and multiple markers rather than just a single marker; such strategies represent a central component and a paradigm shift in the search for novel biomarkers (Box 3).

These breakthroughs have paved the way for countless new avenues for biomarker identification. Very few serum tumor markers, however, have been introduced to the clinic over the past 15 years.<sup>9</sup> In this Review, we highlight some of the mechanisms behind biomarker elevation in biological fluids, and outline strategies for novel marker identification. These strategies should facilitate the delivery of potential candidate molecules for cancer diagnosis and prognosis and for prediction of therapy. These projected discoveries may be instrumental in substantially reducing the burden of cancer by providing prevention, individualized therapies, and improved monitoring following treatment.

## MECHANISMS OF BIOMARKER ELEVATION IN BIOLOGICAL FLUIDS

Five major mechanisms exist by which molecules can be elevated in biological fluids during cancer initiation and progression. Such molecules could serve as effective cancer biomarkers. The mechanisms involved are outlined below; some of the different human body fluids that could be used as a source of biomarkers for specific types of cancers are shown in Table 3.

### Gene overexpression

The protein encoded by a gene can be expressed in increased quantities as a result of increases in gene or chromosome copy number (i.e. gene amplification) or through increased transcriptional activity. The latter process could be the result of imbalances between gene repressors and gene activators. Epigenetic changes, such as DNA methylation, are also known to affect gene expression. On a larger scale, chromosomal translocations can result in gene regulation by promoters that are sometimes enhanced by steroid hormones;<sup>10</sup> transposons can serve a similar role.

An example of a putative biomarker is the protein human epididymal secretory protein 4 (HE4, also known as WFDC2), which is overexpressed in ovarian carcinoma. When complementary DNA microarrays were used to identify overexpressed genes in ovarian carcinoma, 101 transcripts were shown to be overexpressed in ovarian cancers compared with normal tissues.<sup>11,12</sup> Real-time polymerase-chain reaction (PCR) assay of an independent set of benign and malignant tissues confirmed that 12 of these transcripts were overexpressed in ovarian cancers. The transcripts *HE4* and *MSLN* seemed to be the most differentially expressed between the tumor and normal tissues. Quantification of HE4 levels in serum revealed that this protein could be a potential biomarker for ovarian cancer,<sup>13</sup> although clinical evaluation is pending. Gene and protein expression of HE4 in a large series of normal and malignant adult tissues, however, showed that HE4 is present in pulmonary, endometrial and breast adenocarcinomas, in addition to staining positively in ovarian carcinoma.<sup>14</sup>

### Increased protein secretion and shedding

Given that 20–25% of all proteins are secreted, aberrant secretion or shedding of membrane-bound proteins with an extracellular domain (ECD) is another means by which molecules can be elevated in biological fluids. Alterations in

**Table 1** Current applications of tumor markers and their limitations.<sup>a</sup>

Application	Current usefulness	Comments
Population screening	Limited	A screening test should have very high sensitivity and exceptional specificity, to avoid too many false positives in populations with a low cancer prevalence. The test must demonstrate a benefit in terms of clinical outcome. Current biomarkers suffer from too low diagnostic sensitivity and specificity to serve as screening markers. Except for PSA, current tumor markers are more frequently elevated at late stages of disease.
Diagnosis	Limited	Current biomarkers suffer from too low diagnostic sensitivity and specificity to serve as diagnostic markers.
Prognosis	Limited	Most cancer markers have some prognostic value. Specific therapeutic interventions cannot be determined because the accuracy of prediction of current markers is rather poor.
Prediction of therapeutic response	High	Very few markers have predictive power (exceptions include steroid hormone receptors and HER2 amplification for breast cancer), but the information they provide aids therapy selection.
Tumor staging	Limited	Besides AFP and HCG, the accuracy of the markers in determining tumor stage is poor.
Detecting early tumor recurrence	Controversial	Lead time is short and does not considerably affect outcome. Clinical relapses could occur without biomarker elevation. Biomarker elevation can be nonspecific
Monitoring effectiveness of cancer therapy	High	Current biomarkers provide information on therapeutic response (effective or noneffective) that is readily interpretable and more economical than imaging modalities.

<sup>a</sup>Table modified with permission from Diamandis EP *et al.* (2002) Tumor markers: past, present, and future. In: Diamandis EP *et al.*, eds. *Tumor Markers: Physiology, Pathobiology, Technology, and Clinical Applications*. Washington DC: AACCC Press.<sup>80</sup> Abbreviations: AFP, alfa-fetoprotein; HCG, human chorionic gonadotropin; PSA, prostate-specific antigen.

the signal peptide of proteins as a result of single nucleotide polymorphisms can result in atypical secretion patterns.<sup>15</sup> Moreover, elevation of molecules in biological fluids can result from a change in the polarity of cancer cells, which can lead to the release of cancer-associated glycoproteins into the circulation. Increased expression of proteases that cleave the ECD portion of membrane proteins could also cause increased circulating levels.

Many proteins are secreted into the circulation; one example is alfa-fetoprotein, which is rapidly released from both normal and cancer cells.<sup>16</sup> A classic example of shedding of membrane proteins into fluids (and thus serving as a cancer biomarker) is HER2 (also known as ERBB2). HER2 is a cell membrane surface-bound tyrosine kinase that is involved in cell growth and differentiation.<sup>17</sup> Overexpression of this protein is associated with high risks of relapse and death from breast and ovarian cancers, and HER2 is the target of the therapeutic monoclonal antibody trastuzumab (Herceptin®; Genentech, San Francisco, CA).<sup>18</sup> The HER2 protein consists of a cysteine-rich extracellular ligand-binding domain, a short transmembrane domain, and a cytoplasmic protein tyrosine kinase domain.

The ECD of HER2 can be released by proteolytic cleavage from the full-length receptor protein and can be detected in serum. High levels of HER2 in serum correlate with poor prognosis in patients with breast cancer.<sup>19</sup> In 2000, the FDA approved the serum HER2 test, which is the first blood test for measuring circulating levels of HER2 to have been approved for the follow-up and monitoring of patients with metastatic breast cancer.

#### Angiogenesis, invasion and destruction of tissue architecture

Tissue invasion by the tumor might permit direct release of molecules into the interstitial fluid and subsequent delivery by the lymphatics into the blood. For epithelial cancer types, the proteins must break through the basement membrane of the invading tumor before they appear in the blood. For example, PSA is abundantly expressed by prostatic columnar epithelial cells and secreted into the glandular lumen, comprising a major component of seminal plasma (0.5–3.0 g/l) upon ejaculation. In healthy men, low levels of PSA enter the circulation by diffusing through a number of anatomic barriers, including the basement membrane, the stromal layer, and the walls

**Table 2** Cancer biomarkers that are currently in clinical use.

Tumor marker	Cancer type	Year of discovery and reference	Application based on ASCO and/or NACB recommendations	Reference
Alfa-fetoprotein	Germ-cell hepatoma	1963 <sup>5</sup>	Diagnosis Differential diagnosis of NSGCT Staging Detecting recurrence Monitoring therapy	80
Calcitonin	Medullary thyroid carcinoma	1970s <sup>81</sup>	Diagnosis Monitoring therapy	82
CA125	Ovarian	1981 <sup>7</sup>	Prognosis Detecting recurrence Monitoring therapy	80
CA 15-3	Breast	1984–5 <sup>83,84</sup>	Monitoring therapy	77
CA 19-9	Pancreatic	1979 <sup>85</sup>	Monitoring therapy	86
Carcinoembryonic antigen	Colon	1965 <sup>86</sup>	Monitoring therapy Prognosis Detecting recurrence Screening for hepatic metastases	77,80
ER and PgR	Breast	1970s <sup>87</sup>	Select patients for endocrine therapy	77
HER2	Breast	1985–6 <sup>88,89</sup>	Select patients for trastuzumab therapy	77
Human chorionic gonadotropin- $\beta$	Testicular	1938 <sup>90</sup>	Diagnosis Staging Detecting recurrence Monitoring therapy	80
Lactate dehydrogenase	Germ cell	1954 <sup>91</sup>	Diagnosis Prognosis Detecting recurrence Monitoring therapy	80
Prostate-specific antigen	Prostate	1979 <sup>92</sup>	Screening (with DRE) Diagnosis (with DRE)	80
Thyroglobulin	Thyroid	1956 <sup>93</sup>	Monitoring	82

Abbreviations: DRE, digital rectal examination; ER, estrogen receptor; NACB, National Academy of Clinical Biochemistry; NSGCT, nonseminomatous germ cell tumor; PgR, progesterone receptor.

**Box 3** Why the recent optimism for biomarker discovery?

The emergences of new technologies and new resources have created optimistic views that many more biomarkers will be discovered and validated. New technologies and resources include the following:

- Completion of the Human Genome Project
- Advanced bioinformatics
- Array analysis (e.g. DNA, RNA, protein)
- Mass-spectrometry-based profiling and identification
- Laser-capture microdissection
- Databases of single nucleotide polymorphisms
- Comparative genomic hybridization
- High-throughput sequencing

of blood and lymphatic capillaries. This process gives rise to a normal serum PSA level in the range 0.5–2.0 ng/ml.

Prostatic carcinomas most often arise in the glandular epithelium of the prostate periphery. Although *PSA (KLK3)* gene transcription is down-regulated in prostate cancer, PSA protein levels in the circulation of patients with prostate cancer increase through disruption of the anatomic barriers between the glandular lumen and capillaries. Concomitant to early-stage prostate cancer is the loss of basal cells, disruption of cell attachment, degradation of the basement membrane, initiation of lymphangiogenesis<sup>20</sup> and loss of the polarized structure and luminal secretion by tumor cells. Consequently, PSA levels in the serum can rise to 4–10 ng/ml. Late-stage prostate cancer is characterized by invasion of tumor cells into the stromal layers and the circulation, and by

total loss of glandular organization. This situation enables considerable amounts of PSA to leak into the bloodstream, resulting in typical levels ranging from 10 ng/ml to 1,000 ng/ml (Figure 1).

### STRATEGIES FOR DISCOVERY OF CANCER BIOMARKERS

Genomic and proteomic technologies have greatly increased the number of potential DNA, RNA and protein biomarkers under investigation. A paradigm shift has recently been realized, whereby single-biomarker analysis is being replaced by multiparametric analysis of genes or proteins. This development has triggered the question of whether cancer has a unique fingerprint (i.e. genomic, proteomic, or metabolomic). We outline a number of strategies for cancer biomarker discovery that utilize emerging technologies, and we discuss their merits and limitations (Figure 2).

#### Gene-expression profiling

Genomic microarrays represent a highly powerful technology for gene-expression studies. Microarray experiments are usually performed with DNA or RNA isolated from tissues, which are labeled with a detectable marker and allowed to hybridize to arrays comprised of gene-specific probes that represent thousands of individual genes.<sup>21</sup> The greater the degree of hybridization, the more intense the signal, thus implying a higher relative level of expression. The massive amount of data per experiment means that the molecular markers and their expression patterns need to be analyzed by elaborate computational tools, which add an additional layer of statistical complexity. Two basic forms of analysis are unsupervised and supervised hierarchical clustering algorithms;<sup>22</sup> the latter tools identify gene-expression patterns that discriminate tumors on the basis of predefined clinical information.<sup>23</sup> A third method, quantitative real-time PCR, is generally considered the 'gold standard' against which other methods are validated.

The cancer subclassification hypothesis states that gene-expression patterns identified with DNA microarrays can predict the clinical behavior of tumors.<sup>24</sup> The proof-of-principle for the cancer subclassification hypothesis has been provided for various malignancies, such as leukemias, breast cancers and many other tumor types.<sup>25–31</sup> For example, results from gene-array technologies have enabled breast cancers to be classified into prognostic categories dependent on the expression of certain genes. The 70-gene-panel microarray study of survival prediction led to the development

**Table 3** Human biological fluids: a source for biomarker discovery.

Human biological fluid	Cancer type
Plasma	Broad spectrum of diseases
Serum	Broad spectrum of diseases
Cerebrospinal fluid	Brain
Nipple aspirate fluid	Breast
Breast cyst fluid	Breast
Ductal lavage	Breast
Cervicovaginal fluid	Cervical and endometrial
Stool	Colorectal
Pleural effusion	Lung
Bronchoalveolar lavage	Lung
Saliva	Oral
Ascites fluid	Ovarian
Pancreatic juice	Pancreatic
Seminal plasma	Prostate and testicular
Urine	Urological

of MammaPrint® (Agendia, Amsterdam, The Netherlands),<sup>32</sup> which in February 2007 became the first multigene panel test to be approved by the FDA for predicting breast cancer relapse. Another genomic microarray, Oncotype DX® (Genomic Health, Redwood City, CA), based on quantitative real-time PCR, has been commercially available for the same use since 2004 since the validation of its gene signature for predicting the recurrence of tamoxifen-treated, node-negative breast cancer.<sup>33</sup> For the validation, clinical trials initiated by the National Surgical Adjuvant Breast and Bowel Project (NSABP) in the 1980s were retrospectively analyzed, covering a median follow-up of 14 years. Oncotype DX® and MammaPrint® use different analytical platforms and, despite their similar clinical indication, they have only a single gene overlap in their panels. Nevertheless, over the past decade, a tremendous growth in the application of gene-expression profiling has been witnessed. This growth has contributed to the cancer subclassification theory,<sup>24</sup> insights into cancer pathogenesis, and the discovery of a large number of diagnostic markers.<sup>34</sup>

In 2005, Michiels *et al.* performed a meta-analysis of seven of the most prominent studies on cancer prognosis that used microarray-based expression profiling.<sup>35</sup> Surprisingly, in five of these studies the original data could not be reproduced.<sup>36</sup> The analysis of the other two studies provided much weaker prognostic information











