Bioinformatics Strategies for Translating Genome-Wide Expression Analyses into Clinically Useful Cancer Markers

DANIEL R. RHODES[a] AND ARUL M. CHINNAIYAN[a,b,c]

[a]*Department of Pathology,* [b]*Department of Urology, and* [c]*Comprehensive Cancer Center, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA*

ABSTRACT: The DNA microarray has revolutionized cancer research. Now, scientists can obtain a genome-wide perspective of cancer gene expression. One potential application of this technology is the discovery of novel cancer biomarkers for more accurate diagnosis and prognosis, and potentially for the earlier detection of disease or the monitoring of treatment effectiveness. Because microarray experiments generate a tremendous amount of data and because the number of laboratories generating microarray data is rapidly growing, new bioinformatics strategies that promote the maximum utilization of such data are necessary. Here, we describe a method to validate multiple microarray data sets, a Web-based cancer microarray database for biomarker discovery, and methods for integrating gene ontology annotations with microarray data to improve candidate biomarker selection.

KEYWORDS: cancer; biomarkers; bioinformatics; meta-analysis; microarray

INTRODUCTION

Advances in DNA microarray technologies have led to an explosion of cancer gene expression profiling studies revealing a number of potential cancer markers: both tissue markers that may aid in more accurate diagnosis and prognosis, and potential serum markers that may aid in the early detection of cancer and in monitoring the effectiveness of therapy. The task of translating genome-wide expression data into clinically useful biomarkers poses many challenges, one being the selection of the most promising potential markers for future studies and another being the careful validation of markers on a large cohort of clinical samples. Because this validation process can be labor- and resource-intensive, the task of selecting candidate markers becomes very important. Furthermore, because a relatively small number of laboratories are applying DNA microarray technology and generating genome-wide expression data, the task of making gene expression data and analysis methods available to the cancer research community is equally important. With the rising flood of cancer

Address for correspondence: Arul M. Chinnaiyan, M.D., Ph.D., Assistant Professor of Pathology and Urology, Director of the Pathology Microarray Lab, Director of the Tissue/Informatics Core, University of Michigan Medical School, Department of Pathology, 1301 Catherine Street, MS1 Room 4237, Ann Arbor, MI 48109-0602. Voice: 734-936-1887; fax: 734-763-6476.
arul@umich.edu

Ann. N.Y. Acad. Sci. 1020: 32–40 (2004).
doi: 10.1196/annals.1310.005

gene expression profiling data in the public domain, it is up to those in the field of bioinformatics to provide methods to evaluate, integrate, and make available genome-wide expression data. In this report, we will discuss bioinformatics strategies that we and others are employing to improve candidate marker selection and, ultimately, the translation of genome-wide expression analyses into clinically useful cancer markers (FIG. 1).

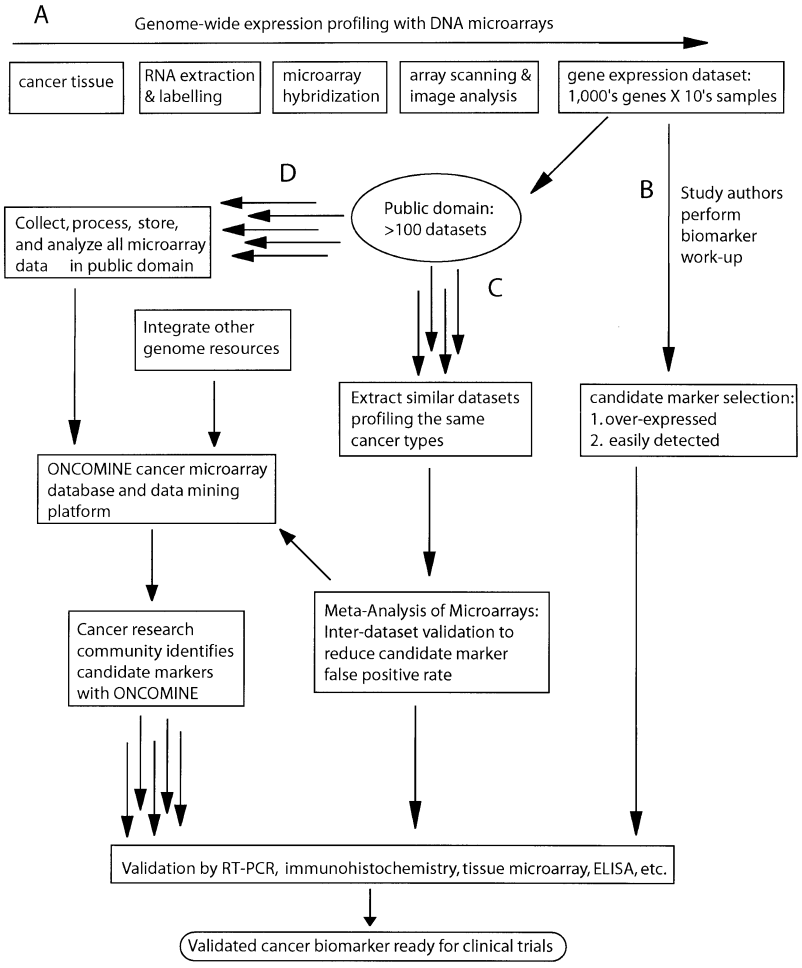

FIGURE 1. A flowchart demonstrating the application of bioinformatic strategies to the improved identification of candidate biomarkers from cancer genome-wide expression analyses. **(A)** The typical experimental procedure used to generate genome-wide expression data. **(B)** Study authors may attempt to identify and validate candidate biomarkers. **(C)** Microarray data in the public domain allows for multiple data set analysis strategies including meta-analysis of microarrays, a statistical method used to intervalidate data sets and thus reduce the candidate biomarker false-positive rate. **(D)** Furthermore, data in the public domain can be unified and made available to the cancer research community, as with the ONCOMINE cancer microarray database, so that more potential markers can be identified and validated.

MARKERS IDENTIFIED BY GENOME-WIDE EXPRESSION ANALYSIS

To date, more than 100 studies have profiled human cancer samples using DNA microarrays; however, only a fraction of these studies have demonstrated thorough validation of results and the development of clinically useful biomarkers (FIGS. 1A and 1B). Validation usually involves confirming that the protein product of a gene highly expressed in cancer is similarly overexpressed. This is necessary because most clinical tests involve measuring protein level, either by immunohistochemistry in the case of tissue biomarkers or by ELISA in the case of serum biomarkers. Furthermore, validation should be carried out on a large sample set to assure that the results can be generalized to the population of patients. This need has been addressed by the development of tissue microarrays, a technology that allows one to measure a protein's expression level in hundreds or thousands of clinical samples in a single assay.

An example of a cancer tissue marker discovered by genome-wide expression profiling and then validated by tissue microarray is alpha-methylacyl CoA racemase (AMACR), a protein specifically overexpressed in prostate cancer.[1] Multiple gene expression profiling studies,[2–5] as well as a meta-analysis,[6] found the *AMACR* gene transcript to be highly overexpressed in prostate cancer. Immunohistochemical analysis with tissue microarrays revealed that AMACR was similarly overexpressed at the protein level in 94 prostate cancer needle biopsy samples (97% sensitivity, 100% specificity). Studies are now under way evaluating the clinical utility of AMACR in uncertain diagnoses. Another example of a marker discovered in this manner is enhancer of *zeste homolog 2* (*EZH2*). *EZH2* gene transcript and protein were found to be highly expressed in metastatic prostate cancer[7] and, interestingly, were more highly expressed in tumors of patients with progressive disease. A follow-up study found that EZH2 protein level in conjunction with E-cadherin protein level significantly predicted disease recurrence following surgery in a multivariable model, which included other clinical and pathological prognostic variables.[8]

Genome-wide expression analyses have also been useful in identifying serum biomarkers that could potentially aid in the early detection of cancer. One study used microarrays to identify genes overexpressed in ovarian cancer.[9] They identified prostasin, a gene that encodes a protein thought to be secreted from cells, as one of the most highly expressed transcripts. Validation by ELISA confirmed that prostasin protein is at high levels in the serum of patients with ovarian cancer and that it may serve as a biomarker useful for the early detection of ovarian cancer (92% sensitivity, 94% specificity). Another example was the discovery of osteopontin as a potential serum biomarker for hepatocellular carcinoma.[10]

BIOINFORMATICS HURDLES SLOWING BIOMARKER DISCOVERY

While the examples highlighted in the previous section serve to illustrate the potential use of genome-wide expression data in the discovery of clinically important biomarkers, it is worth noting that these examples only represent a small minority of studies and that, in the majority of genome-wide analyses, little or no validation is performed. While the lack of validation is unfortunate, the data from these analyses are often made available to the public, so it is conceivable that those researchers interested in cancer biomarkers could further analyze published data to identify

promising candidate markers for validation. This task presents a number of challenges, many of which are being addressed by the growing field of bioinformatics.

One challenge lies in coping with multiple data sets that have profiled similar cancer samples. While it is surely best to use these multiple data sets to validate one another so that the most promising candidate biomarkers can be identified, this task is challenging because microarray data exist on a variety of scales depending on the specific technological platform utilized as well as the experimental procedure. Usually, microarray data from independent laboratories are not thought to be directly comparable. Computational/statistical methods are being developed so that independent microarray data sets can be easily compared.

Another challenge lies simply in data availability and data exchange. While many genome-wide expression data sets are made freely available upon publication, the format that the data are stored in is often heterogeneous. Recently, standards have been developed for microarray data storage and exchange, designated Minimum Information about Microarray Experiments (MIAME).[11] These standards will likely facilitate use of public repositories and common data analysis tools. Multiple repositories that implement the MIAME standards are being developed, including Gene Expression Omnibus[12] and Array Express.[13] While these efforts are already beginning to prove fruitful, a mass of invaluable data in the public domain may not be deposited in these repositories. In order for these data sets to be utilized, they must be actively collected, which can be a daunting task.

A final challenge lies in microarray data analysis. This is a complex field requiring computational and statistical expertise. Moreover, because those most likely to translate genome-wide expression data into useful markers do not always possess such expertise, it is up to those in the field of bioinformatics to provide the tools necessary to analyze and visualize the data, as well as integrate the data with gene annotation resources.

META-ANALYSIS OF MICROARRAYS

While most cancer gene expression profiling studies claim to identify large sets of potential cancer markers, it is generally thought necessary to demonstrate independent experimental validations using techniques such as reverse transcriptase polymerase chain reaction (RT-PCR), Northern blots, or tissue microarrays before a gene (or a set of genes) is considered as a valid potential marker. Validation is necessary because microarray studies are known to generate falsely positive results for a number of reasons including random chance, experimental artifacts, sampling bias, cross-hybridization, etc.; therefore, it is commonplace to use the microarray as a screening tool and then to validate a few promising candidates for future study (FIGS. 1A and 1B). While this model has proved somewhat fruitful in identifying markers, it underutilizes the original microarray data set, often overlooking many other possible markers. It is likely that the best markers may have been missed or overlooked simply because of the challenge in validating many genes.

With the increasing number of publicly available gene expression data sets, we have proposed a meta-analysis of multiple data sets that addresses similar hypotheses in order to validate and statistically assess all positive results simultaneously (FIG. 1C).[6] While validating microarray data sets against one another does not offer

the same confidence as validation by protein expression profiling with tissue microarrays, it does rid us of most of the causes of false positives and is sure to be void of artifacts of individual studies. Interstudy validation of microarray data sets poses unique challenges both statistically and computationally—for while the hypotheses in microarray studies are often similar (i.e., identify genes differentially expressed in cancer), individual investigators often use distinct protocols, microarray platforms, and analysis techniques, and additionally the raw gene expression measurements are often incomparable.

We developed a method, termed meta-analysis of microarrays, that compares statistical measures across studies instead of actual gene expression measurements—for while the actual expression measurements may have different meaning in different studies, *P* values generated for each study by a common statistical test are easily comparable. Our method begins by assigning *P* values to each gene in each study using the *t* statistic, and then the similarity of *P* values for each gene profiled is assessed using traditional meta-analysis methods combined with a simple correction for multiple hypothesis testing.

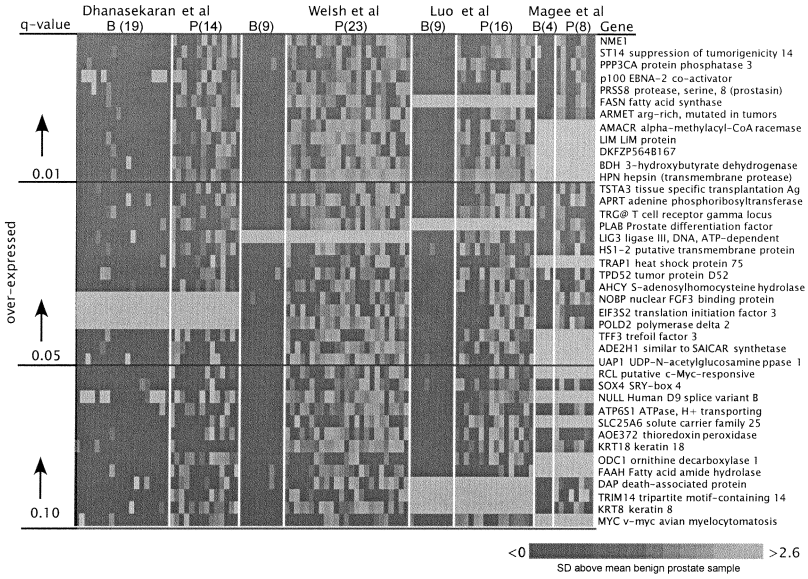

FIGURE 2. Meta-analysis of four prostate cancer gene expression data sets present in the public domain identifies *in silico* validated genes that are overexpressed in prostate cancer (P) relative to benign prostate (B). Each column represents an individual sample (number of samples is in parentheses) and each row represents a specific gene. Within each study, the data were normalized so that the mean expression level of the genes in the benign prostate specimens equaled zero and the standard deviation equaled one. Forty genes with the lowest meta-*q* value for overexpression are shown. Gray intensity level indicates degree of overexpression, while black indicates equal or lower expression than the mean benign sample (see scale).

The model was first implemented on four publicly available prostate cancer gene expression data sets generated by independent laboratories.[2–4,14] All four studies made comparisons between the gene expression profiles of clinically localized prostate cancer and benign prostate tissue, with the goal of identifying genes differentially expressed between the two sample groups. Two of the groups used spotted cDNA technology,[2,3] while two groups used commercial oligonucleotide-based technology.[4,14] As anticipated, a large group of genes, many more so than would be expected by chance, were significantly differentially expressed in multiple independent data sets, suggesting that they are truly differentially expressed, thus increasing the likelihood that they could serve as potential cancer markers. We found 50 genes to be reliably overexpressed and 103 genes to be reliably underexpressed in prostate cancer at a q value (i.e., meta-analysis measure of significance) of 0.10. FIGURE 2 displays the top 40 overexpressed genes. Several of the genes validated or commented on in the individual studies scored high in this analysis, including *hepsin*, *myc*, and *fatty acid synthase*, and *AMACR*, but importantly many genes that had not yet been validated scored equally high. In summary, our method for the meta-analysis of microarrays provided a statistical framework for interstudy validation, suggesting a new approach for dealing with multiple analogous microarray data sets.

We have recently extended our approach to a large compendium of public cancer microarray data sets, further defining cancer-type specific meta-profiles as well as generating meta-profiles common to multiple cancer types.[15]

THE AVAILABILITY AND INTEGRATION OF GENE EXPRESSION PROFILING DATA

In the previous section, we discussed the issues of microarray result reliability and statistical approaches for the validation of multiple analogous data sets. In this section, we will address the importance of data availability and the need for bioinformatics tools to make cancer gene expression data available and easily interpretable by the cancer research community. This issue is critical because only a fraction of laboratories are applying DNA microarrays to cancer genome-wide expression profiling, but the results from these experiments could be potentially useful to a large number of cancer researchers, both basic science and clinical. For most published microarray studies, which may comprise thousands of gene measurements across tens or hundreds of cancer specimens, the authors have usually presented one interpretation of their data and have reported on only a subset of genes that demonstrate their particular hypothesis. Furthermore, the focus is not always on developing novel cancer markers, so often times there is no validation or follow-up studies. This may be due to the fact that the researchers involved in cancer genomics are often interested in global patterns of gene expression and are not necessarily interested in translating gene expression profiles into novel cancer markers. For those interested in developing cancer markers, the complete microarray data sets are sometimes made available as supplementary data; however, even if that is the case, the data sets often sit as cryptic text files, stored and processed in an unsystematic manner, and thus difficult to interpret unless one has a fair amount of computational expertise. While the aforementioned standards and repositories have begun to ameliorate this problem, cancer microarray data will be most useful to clinical cancer researchers only when it is unified, logically analyzed, and made easily accessible.

To this end, we initiated an effort to systematically curate, analyze, and make available all public cancer microarray data via a Web-based database and data-mining platform, designated "ONCOMINE" (http://www.oncomine.org/)[16] (FIG. 1D). Our effort also includes centralizing gene annotation data from various genome resources to facilitate rapid interpretation of a gene's potential role in cancer. Furthermore, we have integrated microarray data analysis with other resources including gene ontology annotations and a therapeutic target database so that clinically interesting subsets of genes can be focused on. Currently, the ONCOMINE database houses 65 independent data sets comprising nearly 50 million gene expression measurements from more than 4700 microarray experiments. More than 100 differential expression analyses define the genes most over- and underexpressed in nearly every major cancer type as well as a number of clinical and pathology-based cancer subtypes. It is our hope that, by making these data easily accessible to the cancer research community, potential

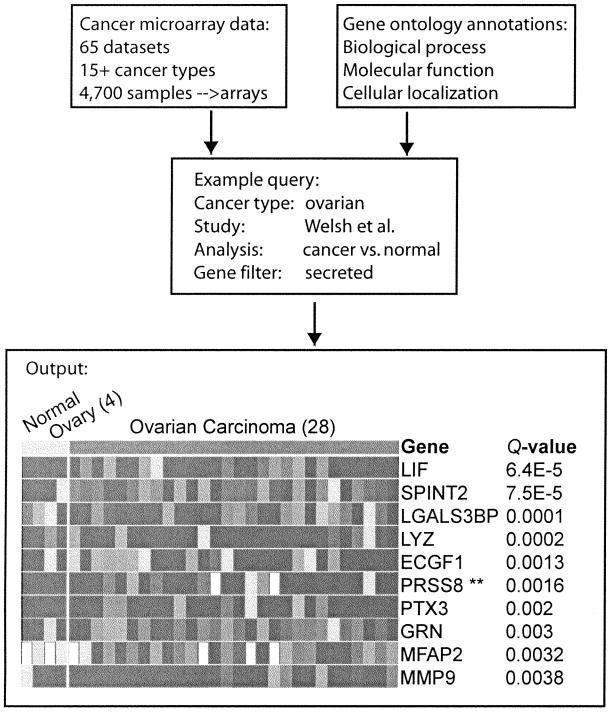

FIGURE 3. The ONCOMINE cancer microarray database used to identify potential serum biomarkers for ovarian cancer. By selecting an ovarian cancer data set generated by Welsh *et al*.,[5] specifying the differential expression analysis that identified genes overexpressed in ovarian cancer relative to normal ovary tissue, and then applying the "secreted" gene filter, which was derived from gene ontology cellular localization annotations, ONCOMINE provides a gene expression heatmap representation of the 10 genes that encode secreted proteins that are most highly expressed in ovarian cancer.

cancer markers will be easily identified, promoting an increase in validation studies and ultimately an increase in clinically useful markers.

Genes are usually considered as potential markers if they are differentially overexpressed in a particular cancer and their molecular function or localization suggests that they might be amenable to detection in serum or tissue. To provide a platform for the discovery of potential markers that are overexpressed in cancer, ONCOMINE is integrated with gene ontology annotations from the Gene Ontology Consortium.[17] Now, rather than investigating the function and localization of the genes most overexpressed in cancer to assess their potential as candidate markers, users can begin with only those genes whose function or localization suggests that they might be useful markers. For example, genes that encode proteins that are secreted from cancer cells would likely serve as candidate serum biomarkers. These include genes with a cellular localization annotation of extracellular space, extracellular matrix, or extracellular. With this ontology filter in place, for each cancer type in the ONCOMINE database, users can quickly identify genes that encode secreted products that are specifically overexpressed in a particular cancer. FIGURE 3 shows an analysis session as an example from ONCOMINE that highlights the genes that encode secreted proteins that are overexpressed in ovarian cancer relative to normal ovarian tissue.[5] Interestingly, the fifth most overexpressed gene, prostasin, was recently found to be a novel serum biomarker for ovarian cancer.[9] Perhaps the other genes that are more significantly overexpressed may serve as even better biomarkers. Further analysis with ONCOMINE revealed that prostasin is similarly overexpressed in prostate cancer, suggesting a broadened role for this marker.

A recent study by Welsh *et al.*[18] demonstrated a similar approach to serum marker discovery from genome-wide expression data. They used the gene ontology–based approach also used in ONCOMINE, as well as a sequence-based approach to define additional genes that encode secreted products. By integrating the secreted annotations defined by these two approaches with a large multicancer-type microarray data set, they were able to define 74 potential serum cancer markers. A number of the identified markers have been shown previously to be elevated in the serum of cancer patients, including kallikreins in ovarian carcinomas, gastrin releasing peptide in lung carcinomas, and alpha-fetoprotein in liver carcinomas. Other new markers were also identified and validated, including MIC-1, which by ELISA was found to be at high levels in a number of cancer types.

CONCLUSIONS

In summary, the field of bioinformatics is playing a critical role in beginning the translation of cancer gene expression profiling into useful cancer markers. Our approach for meta-analysis allows for the integration and validation of multiple microarray data sets so that the most promising candidate markers can be identified and followed up on. Furthermore, the development of ONCOMINE and other cancer microarray databases should promote the maximum utilization of cancer microarray data by the research community. Finally, integration of gene expression profiling data with other genome resources such as gene ontology annotations provides a powerful platform for marker discovery, as evidenced by the recent work of Welsh *et al.*[18]

REFERENCES

1. RUBIN, M.A. *et al.* 2002. Alpha-methylacyl coenzyme A racemase as a tissue biomarker for prostate cancer. JAMA **287:** 1662–1670.
2. DHANASEKARAN, S.M. *et al.* 2001. Delineation of prognostic biomarkers in prostate cancer. Nature **412:** 822–826.
3. LUO, J. *et al.* 2001. Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. Cancer Res. **61:** 4683–4688.
4. MAGEE, J.A. *et al.* 2001. Expression profiling reveals hepsin overexpression in prostate cancer. Cancer Res. **61:** 5692–5696.
5. WELSH, J.B. *et al.* 2001. Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. Proc. Natl. Acad. Sci. USA **98:** 1176–1181.
6. RHODES, D.R. *et al.* 2002. Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. Cancer Res. **62:** 4427–4433.
7. VARAMBALLY, S. *et al.* 2002. The polycomb group protein EZH2 is involved in progression of prostate cancer. Nature **419:** 624–629.
8. RHODES, D.R. *et al.* 2003. Multiplex biomarker approach for determining risk of prostate-specific antigen-defined recurrence of prostate cancer. J. Natl. Cancer Inst. **95:** 661–668.
9. MOK, S.C. *et al.* 2001. Prostasin, a potential serum marker for ovarian cancer: identification through microarray technology. J. Natl. Cancer Inst. **93:** 1458–1464.
10. YE, Q.H. *et al.* 2003. Predicting hepatitis B virus–positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning. Nat. Med. **9:** 416–423.
11. BRAZMA, A. *et al.* 2001. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat. Genet. **29:** 365–371.
12. EDGAR, R., M. DOMRACHEV & A.E. LASH. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. **30:** 207–210.
13. BRAZMA, A. *et al.* 2003. Array Express—a public repository for microarray gene expression data at the EBI. Nucleic Acids Res. **31:** 68–71.
14. WELSH, J.B. *et al.* 2001. Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. Cancer Res. **61:** 5974–5978.
15. RHODES, D.R. *et al.* 2004. Large-scale meta-analysis of cancer microarray data identifies common gene expression profiles of neoplastic transformation and progression. Under review.
16. RHODES, D.R. *et al.* 2004. ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia **6:** 1–6.
17. ASHBURNER, M. *et al.* 2000. Gene ontology: tool for the unification of biology—The Gene Ontology Consortium. Nat. Genet. **25:** 25–29.
18. WELSH, J.B. *et al.* 2003. Large-scale delineation of secreted protein biomarkers over-expressed in cancer tissue and serum. Proc. Natl. Acad. Sci. USA **100:** 3410–3415.