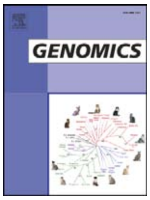
Genomics 93 (2009) 13–16

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno



“Omics” data and levels of evidence for biomarker discovery

Debashis Ghosh [a,b,*], Laila M. Poisson [c]

[a] *Department of Statistics, College of Medicine, Pennsylvania State University, University Park, PA 16802, USA*
[b] *Department of Public Health Sciences, College of Medicine, Pennsylvania State University, University Park, PA 16802, USA*
[c] *Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA*

ARTICLE INFO

Article history:
Received 7 January 2008
Accepted 2 July 2008
Available online 14 September 2008

Keywords:
Biomarker
Cancer
Gene Expression
High-dimensional data
Proteomics

ABSTRACT

With the development of new technologies for assaying biological activity on a global basis in experimental samples, various new “-omics” signatures have been developed to predict disease progression. Such signatures hold the potential to alter the nature of clinical management of human disease. In this article, we describe some necessary statistical considerations needed to take these signatures from the discovery phase to a clinically useful assay. Much of the work discussed is in the area of cancer.

Introduction

The explosion of high-throughput technologies available for generating large-scale molecular-level measurements in human populations has led to an increased interest in the discovery and validation of molecular biomarkers in medical research. Uses of biomarkers in medical decision making is quite varied and includes such key features as surrogate endpoints [1], proxies for exposure [2], early detection of disease [3], and identification of predictive and prognostic factors in disease management [4].

A biomarker is formally defined as “a biological characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention” [5]. In translating this definition into the context of “omics” data (e.g., transcriptomic, proteomic, genomic) it is difficult to identify what is meant by a “biological characteristic.” Often when omics data are evaluated for features associated with the medical condition of interest multiple molecular features emerge. Combined, these features may have biomarker potential and thus the biological characteristic of interest is in fact a set of features. This subtle change from considering a single molecular biomarker to considering a biomarker profile has motivated discussion on their proper reporting [6] and governmental regulation of clinical use [7]. Ultimately, a biomarker profile should undergo the same scrutiny required of single molecular markers.

In the following we explore current trends in biomarker research in the context of omics data. Examples will focus primarily on cancer, given our expertise, though we acknowledge that omics-based biomarker research is utilized in other areas [8].

The bioinformatics approach to signature discovery and its reporting

Studies reporting new genomic signatures are being published at an astonishing rate in the scientific and medical literature. Before we can assess the utility of these signatures as biomarkers we must consider the complexity of the data analyses used to derive them. Many of these studies use the same general statistical paradigm, data preprocessing within and between samples to reduce experimental noise, followed by statistical assessment of association of the molecular measures with disease. Associations may be assessed by supervised methods, in which clinical outcomes are related directly to genomic features through statistical tests and models [9], or unsupervised methods, in which measured elements are clustered independent of clinical outcome and then assessed for clinical trends [10]. Associations found are nominated as candidate molecules for further study or experimentation. The actual statistical test used will depend on the nature of the data generated. For example, quantitative measures of gene expression from microarrays will require a treatment different from that of the allele calls that arise using high-throughput single-nucleotide polymorphism arrays.

Although this general bioinformatic paradigm is consistent across omics studies, the actual experimental design and analysis may vary

[*] Corresponding author.
E-mail address: ghoshd@psu.edu (D. Ghosh).

0888-7543/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2008.07.006

widely within and between omics areas. With this experimental variation in mind McShane and coauthors [6] recently proposed a set of criteria that should be used when reporting the results of biomarker studies. Known as REMARK (Reporting Recommendations for Tumor Marker Prognostic Studies), this document provides a detailed description as to the minimum amount of information that should be given in the reporting of results from tumor biomarker studies. The REMARK guidelines can be found at the URL http://cancerdiagnosis.nci.nih.gov/assessment/progress/remark.htm. Specifically, they list 20 items that investigators should attempt to report in any tumor biomarker study. Included are items such as a description of the patient population, a report of why some data are missing from the study, and a statistical report of how cut-points are determined for categorization of biomarkers as well as the model building procedures used for univariate and multivariate analyses.

The REMARK guidelines are applicable to all biomarker studies and are not specific to those arising from omics data. The use of high-throughput data additionally provides unique challenges that need careful attention. Consider, for instance, a gene expression study in which the features are the expression levels of the individual spots on a microarray so that thousands of features will be tested for association with disease. Two cautionary flags should be immediately raised.

First, if each feature is tested independently, then the major issue of multiple comparisons arises [11], namely that the number of features is so large that one would expect to find many spurious associations by chance. Practically this means that many of the significant associations in the data are likely spurious. The study of the effects of multiple comparisons on discovery error (false positive and false negative rates) in omics data analysis has produced a growing body of literature. Much work has been done to adjust *p* values or *p*-value thresholds to control the false discovery rate while maintaining power [12,13]. Though the issue of multiple comparisons is fairly obvious when hundreds or thousands of tests are run, it should be considered even beyond the univariate analysis phase of analysis. Ideally preplanned analyses addressing a priori hypotheses should be used to control discovery error rates. Given that most omics studies are exploratory, spurious findings are best sifted out by follow-up studies and independent validation.

The second cautionary flag should be raised for the so-called "large *P*, small *N*" [14] problem. When there are more parameters ("*P*") than subjects ("*N*") many standard statistical models and tests are no longer applicable. In particular, it will be possible for many different signatures to perform equally well based on classification performance. Such an example was seen in the work of Fan and colleagues [15], in which several breast cancer genomic signatures were found to have strong concordance in classification of patients despite having no overlap. Additionally, when deriving a biomarker profile, a set of *N* subjects can be perfectly classified into clinically meaningful categories using only *N* features. Here again we see that it is of utmost importance that biomarker profiles be validated appropriately.

Given the complexity of the data analysis from which biomarkers are derived, either singly or as profiles, it is clear that their utility will be determined through additional studies. In fact, the amount of evidence supporting a biomarker's validity is minimal at the time of first discovery. This will be discussed further in the section entitled Two paradigms for assessing the strength of evidence. Through "transparent and complete reporting" [6] of biomarker discovery studies, as promoted by the REMARK guidelines, independent groups, whether labs or regulatory agencies, can evaluate their potential use in clinical settings.

Predictive and prognostic biomarkers and their potential clinical utility

In the area of disease management, two types of biomarker signatures are of general interest: predictive and prognostic. Clinically, a prognostic biomarker is one that can separate a diseased population into groups of similar prognosis, while a predictive factor is one that can identify a subpopulation of patients that is more likely to benefit from a certain treatment. These two concepts are related, since a prognostic marker may also be predictive, though their distinction is important for experimental design, analysis, reporting, and validation. In statistical jargon, prognostic factors involve finding important main effects, while predictive factors are those that represent an interaction between the treatment and the factor. While there are similarities in the statistical techniques used to determine if biomarkers are prognostic or predictive, in practice they will be used quite differently. An important point is that if the standard treatment for a disease changes, a predictive biomarker would have to be revalidated in the context of the new treatment.

For the sake of illustration, consider the breast cancer genomic studies conducted by van't Veer et al. [16] and Paik et al. [17]. In the gene expression study by van't Veer et al. [16], a series of 98 breast cancers was profiled using a 25,000-gene microarray, which was then used to develop a 70-gene signature that is prognostic for recurrence of aggressive breast cancer. In the Paik et al. [17] study, investigators started with a set of known cancer-related genes and then used a reverse transcriptase–polymerase chain reaction assay to quantify gene expression. From that initial set of genes, they selected a set of 21 genes that associated significantly with likelihood of distal recurrence. Using this 21-gene signature they were able to predict patient recurrence risk as falling into one of three categories (low, medium, and high). Since the population of patients in the study had no positive lymph nodes and had ER-positive breast cancer, this 21-gene signature was hailed as a breakthrough in the management of this particular patient population.

Both the van't Veer and the Paik gene signatures were discovered to be prognostic signatures but are currently being used in clinical trials to test their utility as predictive markers for determining treatment in early stage breast cancer populations. Coordinated by the European Organisation for Research and Treatment of Cancer, the van't Veer signature is being tested in a phase III clinical trial entitled MINDACT (**M**icroarray **in** **N**ode-Negative **D**isease May **A**void **C**hemo**t**herapy) [18]. As described in their title, they hope that their 70-gene signature will identify women who will benefit from chemotherapy among those with lymph-node-negative disease. Sponsored by the National Cancer Institute the 21-gene signature of Paik et al. is being tested in the Trial Assigning Individualized Options for Treatment (Rx), or TAILORx [19]. Again, they are hoping that their gene signature will help women with early stage breast cancer to make decisions about their treatment, specifically chemotherapy.

Although the above-mentioned gene signatures are promising, for most studies there are issues in the practical use of biomarker profiles as either prognostic or predictive factors. First, in many of the studies in which candidate biomarker profiles are generated, samples are collected as convenience samples rather than through a randomized protocol. This is often because of limited sample availability for retrospective study but may result in signatures that simply recapitulate standard prognostic schemes or may be subject to complicated confounding patterns [20]. We therefore suggest that omics biomarker discovery include models that adjust for other known predictive or prognostic factors as much as possible. For instance, clinical parameters can be included in univariate ANOVA models run in place of *t* tests for per-feature association. Additionally, candidate biomarkers could be tested with other known prognostic factors in survival analysis models. In this way we can be assured that the candidate biomarker will provide information above and beyond that which is already known by standard prognostic factors. Ransohoff [21] argues that hidden and hardwired biases in the experiment will always exist (e.g., sample handling/preparation/processing). Consequently, one should proceed very cautiously when finding associations in the study.

A second issue is the notion of validating on an independent dataset. What is typically done is to find a predictive gene expression signature on a training dataset and then test the classifier on a test dataset. In initial gene expression experiments, both the test and the training data were used from the same study. More recently, people have started using data from one study as the training dataset and those from a second study as a testing dataset. One example of this is given in [22], in which the investigators were interested in developing a gene expression profile that predicted survival in lung cancer patients. After finding a set of genes that were associated with survival, the investigators then validated their signature based on a separate collection of lung adenocarcinomas collected at a different institution [23].

Most major journals now require that high-throughput data be made publicly available upon publication of the article. Therefore it should become increasingly feasible to use publicly available omics data, generated by other investigators studying a related problem, as a method of validating biomarker profiles. Results that are reproducible across multiple studies show stronger evidence of a true association and thus have more potential for clinical utility. This is an area that our group has been fairly active in from both a methodological and a scientific viewpoint [24,25].

Two paradigms for assessing the strength of evidence

For biomarkers undergoing validation studies, let us consider the strength of evidence provided by various study designs. There are two such paradigms that provide a standard by which to gauge the strength of evidence supporting a biomarker's utility. Ultimately, before its use in a clinical setting, a biomarker must be validated using a more classical clinical trials-based paradigm. The five phases of biomarker development set forth by researchers of the Early Detection Research Network [26] mimic this clinical-trials process. Alternatively, the Levels of Evidence from the Oxford Centre for Evidence Based Medicine [27] provides a ranking of experimental types according to the generalizability of their results.

The paradigm set forth by investigators in the Early Detection Research Network, funded by the National Cancer Institute [26], describes a series of five phases for developing biomarkers (see Table 1), focusing mainly on biomarkers for screening. The five phases begin with exploratory research (phase 1) and progress through development of a clinically useful assay (phase 2), timing of optimal screening period in the disease progression (phase 3), study of biomarker characteristics (phase 4), and assessment of reduced disease burden in the population (phase 5).

Table 1
The five phases of biomarker development for use as a screening tool

Phase	Goals/aims	Experimentation	Sample details
1	Exploratory; nominate and rank candidate biomarker profiles	Preclinical study comparing diseased and nondiseased subjects	Be aware of bias from convenience sampling
2	Develop an assay with clinically reproducible results	Test a noninvasive clinical assay developed from a candidate biomarker profile	Sample population should represent the target population
3	Optimize time interval in which to screen for biomarker	Screen for biomarker in longitudinal study	Collect longitudinal samples from target population
4	Determine operating characteristics of the biomarker as a screening device	Prospective study design testing the screening ability of the biomarker	Sample population should be the target population
5	Assess whether screening reduces the burden of disease	Prospective study assessing survival in the screened vs the unscreened population	Sample population should be the target population

Table 2
The Levels of Evidence

Level	Experimental evidence
1	(a) Data are from a randomized clinical trial with sufficiently high follow-up or (b) Data are from a meta-analysis of several randomized clinical trials in which there is strong evidence of homogeneity across studies
2	(a) Data are from cohort studies (either prospectively collected with poor follow-up or retrospectively collected in untreated subjects) or (b) Data are from a meta-analysis of such studies in which there is strong evidence of homogeneity across studies
3	(a) Data are from ecologic studies or (b) Data are from well-designed case–control studies
4	Data are from poor-quality cohort, case–control, or case–series studies
5	Data are from biologically oriented findings or "wet-lab" experiments

While the five phases are needed for a screening biomarker, a three-phase period of development, which modifies the original framework, has been suggested for biomarkers that will serve as prognostic or diagnostic markers (S. Srivastava, personal communication). What is key to note here is that many of omics data that are used for the generation of candidate biomarker profiles come from data collected in phase 1 of this paradigm. Thus, there is a long road to travel from the reporting of these initial findings, which have appeared in high-profile journals such as *Science* or *Nature*, to the development of an assay that will be clinically useful.

The Levels of Evidence [27] paradigm originated from the literature on clinical trials. They use several criteria for ranking the validity of evidence about preventive interventions. They were originally applied to making recommendations about antithrombotic medications but have been used more generally in evidence-based medicine. The scoring of levels of evidence is from 1 to 5, 1 being the strongest level of evidence and 5 the weakest (see Table 2).

If we were to apply the criteria to the studies generating the genomic profiles, most would rate as evidence of level 3 or weaker (levels 3, 4, and 5). There are several reasons for this, primarily resulting from poor clinical annotation and convenience sampling. Additionally, omics studies are often underpowered for finding biomarker profiles associated with clinical parameters owing to limited sample size. These limitations again highlight the need for extensive experimentation before an omics-based signature can become a practically useful clinical assay. Again, examples of studies that would rate favorably in terms of levels of evidence are the MINDACT and TAILORx trials that are currently in progress.

Conclusion

We approach the future of bioinformatics in clinical studies designed to identify new biomarkers with both enthusiasm and caution. The enthusiasm is due to the explosion in new technologies and assays that will allow for potentially high-throughput measurement of different types of biochemical activity. While microarrays for gene expression are currently the most common, new microarray-based technologies for assessing copy number variation and single-nucleotide polymorphisms are starting to gain more widespread use in clinical research as well. Also, there is a lot of interest in using proteomics and metabolomics to define signatures for predicting disease progression. It is highly likely that there are emerging and new technologies that will also gain popularity in the future.

The caution arises from the fact that analysis and interpretation of the high-dimensional data are subject to many potential pitfalls. Statistically, the issues that are problematic have been well documented in the literature; we have touched on them here. More fundamental are issues of sample collection, study design, and phenotypical heterogeneity. These pitfalls limit the strength of evidence demonstrated by most bioinformatic findings currently.

While attempting to address these issues better, it will also be important to view signature development as being no different from any other biomarker. Thus, multiple hurdles will need to be cleared until the signature will be ready for primetime.

References

[1] T. Burzykowski, G. Molenberghs, M. Buyse, The Evaluation of Surrogate Endpoints, Springer-Verlag, Berlin/New York, 2005.

[2] M.J. Slotnick, J.O. Nriagu, Validity of human nails as a biomarker of arsenic and selenium exposure: a review, Environ. Res. 102 (2006) 125–139.

[3] R. Etzioni, N. Urban, S. Ramsey, M. McIntosh, S. Schwartz, B. Reid, J. Radich, G. Anderson, L. Hartwell, The case for early detection, Nat. Rev. Cancer 3 (2003) 243–252.

[4] D.F. Hayes, Prognostic and predictive factors revisited, Breast 14 (2005) 493–499.

[5] Biomarkers Working Group, Biomarkers and surrogate endpoints: preferred definitions and conceptual framework, Clin. Pharmacol. Ther. 69 (2001) 89–95.

[6] L.M. McShane, D.G. Altman, W. Sauerbrei, et al., Reporting recommendations for tumor marker prognostic studies, J. Clin. Oncol. 23 (2005) 9067–9072.

[7] Draft Guidance for Industry, Clinical Laboratories, and FDA Staff – In Vitro Diagnostic Multivariate Index Assays. U.S. Food and Drug Administration. Web site: http://www.fda.gov/cdrh/ovid/guidance/1610.html. August 25, 2008.

[8] A. Kriete, B.A. Sokhansanj, D.L. Coppock, G.B. West, Systems approaches to the networks of aging, Ageing Res. Rev. 5 (2006) 434–448.

[9] T. Hastie, R. Tibshirani, J.H. Friedman, The Elements of Statistical Learning, Springer, New York, 2001.

[10] A.K. Jain, R.C. Dubes, Algorithms for Clustering Data, Prentice Hall, New York, 1988.

[11] R.G. Miller, Simultaneous Statistical Inference, (2nd ed.)Springer-Verlag, New York, 1981.

[12] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc., B. 57 (1995) 289–300.

[13] G.R. Abecasis, D. Ghosh, T.E. Nichols, Linkage disequilibrium: ancient history drives the new genetics, Hum. Hered. 59 (2005) 118–124.

[14] M. West, Bayesian factor regression models in the "large p, small n" paradigm, Bayesian Stat. 7 (2003) 723–732.

[15] C. Fan, D.S. Oh, L. Wessels, B. Weigelt, D.S.A. Nuyten, A.B. Nobel, L.J. van't Veer, C.M. Perou, Concordance among gene-expression-based predictors for breast cancer, N. Engl. J. Med. 355 (2006) 560–569.

[16] L.J. van't Veer, H. Dai, M.J. van de Vijver, et al., Gene expression profiling predicts clinical outcome of breast cancer, Nature 415 (2002) 530–536.

[17] S. Paik, S. Shak, G. Tang, et al., A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer, N. Engl. J. Med. 351 (2004) 2817–2826.

[18] MINDACT (Microarray In Node negative Disease may Avoid ChemoTherapy) A prospective, randomised study comparing the 70-gene expression signature with common clinical-pathological criteria in selecting patients for adjuvant chemotherapy in node-negative breast cancer. (EORTC Protocol 10041 – BIG 3-04). Web site: http://www.eortc.be/services/unit/mindact/documents/MINDACT_trial_outline.pdf. Date: August 25, 2008.

[19] National Cancer Institute : The TAILORx Breast Cancer Trial. Web site: http://www.cancer.gov/clinicaltrials/digestpage/TAILORx. Date: August 25, 2008.

[20] D. Ghosh, A.M. Chinnaiyan, Covariate adjustment in the analysis of microarray data from clinical studies, Funct. Integr. Genomics 5 (2005) 18–27.

[21] D.F. Ransohoff, Bias as a threat to the validity of cancer molecular-marker research, Nat. Rev. Cancer 5 (2005) 142–149.

[22] D.G. Beer, S.L. Kardia, C.C. Huang, et al., Gene-expression profiles predict survival of patients with lung adenocarcinoma, Nat. Med. 8 (2002) 816–824.

[23] A. Bhattacharjee, W.G. Richards, J. Staunton, et al., Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses, Proc. Natl. Acad. Sci. U. S. A. 98 (2001) 13790–13795.

[24] D. Rhodes, T.R. Barrette, M.A. Rubin, et al., Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer, Cancer Res. 62 (2002) 4427–4433.

[25] R. Shen, D. Ghosh, A.M. Chinnaiyan, Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data, BMC Genomics 5 (2005) 94.

[26] M.S. Pepe, R. Etzioni, Z. Feng, et al., Phases of biomarker development for early detection of cancer, J. Natl. Cancer Inst. 93 (2001) 1054–1061.

[27] Levels of evidence, BJU Int. 101 (2008) 150, doi:10.1111/j.1464-410x.2007.07384.x.