Review

Genomic approaches to research in lung cancer

Edward Gabrielson

The Johns Hopkins University School of Medicine, Baltimore, USA

Received: 21 April 2000
Revisions requested: 11 May 2000
Revisions received: 1 June 2000
Accepted: 1 June 2000
Published: 23 June 2000

Respir Res 2000, **1**:36–39

Abstract

The medical research community is experiencing a marked increase in the amount of information available on genomic sequences and genes expressed by humans and other organisms. This information offers great opportunities for improving our understanding of complex diseases such as lung cancer. In particular, we should expect to witness a rapid increase in the rate of discovery of genes involved in lung cancer pathogenesis and we should be able to develop reliable molecular criteria for classifying lung cancers and predicting biological properties of individual tumors. Achieving these goals will require collaboration by scientists with specialized expertise in medicine, molecular biology, and decision-based statistical analysis.

Keywords: cDNA arrays, genomics, lung cancer

Introduction

Genomics – the discipline that characterizes the structural and functional anatomy of the genome – has attracted continuously increased interest and investment over the past decade. The complete sequencing of the human genome is expected within a few years; together with the identification of expressed sequences and polymorphic sequences, a vast information infrastructure will be available to medical researchers throughout the world.

The rationale for this ambitious project is now well known to the medical community. The discovery of genes involved in the pathogenesis of human diseases will, it is hoped, lead to new targets for diagnosis and treatment of those diseases. Knowing the polymorphisms that make each of us unique individuals could be the key in the future to predicting individual risks for developing disease and individual responses to pharmacological agents. However, the full impact of genomics on medical research is still unknown. Acknowledging the limitations of predicting the role of genomics far in the future, this paper limits its discussion to current applications of genomics to research on lung cancer.

Genomics and gene discovery

A new era of gene discovery began in 1991, when Craig Venter and colleagues reported the sequencing of randomly selected human brain cDNA clones [1]. In this pilot project, 337 of the sequences, termed expressed

CGH = comparative genomic hybridization; DLBCL = diffuse large B-cell lymphoma; EST = expressed sequence tag; SAGE = serial analysis of gene expression.

sequence tags (ESTs), were found to represent new human genes. Over the following years, this strategy has been used by many laboratories to sequence several hundred thousand ESTs, representing most of the genes discovered so far.

However, the era for the actual discovery of new genes has a finite life. Programs such as the Cancer Genome Anatomy Project [2], sponsored by the biotechnology industry and by government, have probably sequenced ESTs for nearly all of the estimated 100 000–120 000 total human genes, and much of this information is currently in public-domain databases. The sequencing of the whole genome and the prediction of coding regions from genomic sequences will probably complete the 'discovery' of gene sequences within a few years.

However, only approximately 5000 genes have known functions or even names, so there will remain much to be learned about how the genes function in health and disease. Thus, cancer researchers will have great opportunities in the next few years to mine genomics databases and identify candidates for genes important in cancer pathogenesis. A good example of such mining of genome databases is the discovery of a new tumor suppressor gene for lung and colon cancers on human chromosome 11 [3]. After localizing a region of frequent chromosomal deletions in lung cancer to a locus at 11q22–24, Wang and colleagues searched genome databases for genes previously mapped to that locus. One gene, PPP2R1B, was found to have somatic mutations in lung cancers and the altered gene products were then found to have functional consequences that would be expected to contribute to the malignant phenotype.

The discovery of PPP2R1B as a gene for a lung cancer tumor suppressor serves as an example of how genomics databases will probably make traditional positional cloning unnecessary. Numerous other chromosomal aberrations and loci of chromosomal deletions have already been defined for lung cancer and, with the increasing availability of gene maps, the next few years are likely to see an acceleration in our recognition of new genes involved in lung cancer pathogenesis.

Functional genomics and lung cancer

Some of the most exciting applications of genomics to cancer research come from measurements of the gene expression of cancer cells with the use of high-throughput technologies. SAGE (serial analysis of gene expression), oligonucleotide arrays, and cDNA arrays are new tools that allow investigators to measure the expression of thousands of genes in a single experiment. Experiments using such approaches are leading to additional discoveries of genes involved in cancer pathogenesis and providing new strategies for classifying cancers.

With SAGE, short tags of mRNA (eg nine bases) are cut from defined positions, serially linked together, and sequenced to provide a quantitative measurement of genes expressed in a sample [4]. For lung cancer, Jen and colleagues at Johns Hopkins sequenced and analyzed over 226 000 tags from two primary lung cancers and two bronchial epithelial cell cultures, finding 175 transcripts to be significantly underexpressed and 142 transcripts to be significantly overexpressed in the cancers in comparison with normal cells [5]. A few of the genes are being studied further for their role in lung cancer pathogenesis, but clearly this single experiment has opened the door to a large number of future studies.

Wang and colleagues at Corixa recently reported an alternative approach to the discovery of lung cancer genes [6]. This group first used subtractive hybridization to isolate cDNA clones highly expressed in squamous cell lung cancers. Analyzing the expression of a larger number of specimens by using cDNA arrays representing the selected clones, genes overexpressed in squamous cell cancers – and potential therapeutic targets – were found.

Both of these projects used 'open' gene discovery strategies, in which all genes present could potentially be detected. Most gene arrays, in contrast, are 'closed' systems for measuring gene expression because the measurements are limited to the genes represented on the arrays. However, arrays are being constructed that represent increasingly large numbers of genes, and the efficiency and relatively low cost of arrays, particularly cDNA arrays [7], will probably lead to the increasing use of this technology for gene discovery.

Genomics and classification of lung cancer

Developing a well-defined taxonomy for cancer is important, both for clinical management of the disease and for cancer research. Because of implications for treatment and prognosis, few would question the significance of differentiating lymphoma from carcinoma, or small cell carcinoma from non-small cell carcinoma in the evaluation of a lung tumor. However, our inability to subclassify lung cancer further is not frequently recognized as a critical deficiency. Most clinical lung cancer research recognizes the standard morphological classification of lung cancers, which is unable to provide critical information on the aggressiveness of a particular cancer or how the cancer will respond to radiation therapy or chemotherapy.

There has been a long-standing hope that molecular markers will help to predict important clinical features of cancers. In most molecular studies, candidate markers were tested individually for their association with outcome. More recently, genomics and technologies such as gene arrays have offered the opportunity to test large numbers of genes as potential predictive markers. Because the number of variables measured invariably exceeds the number of samples in such studies, some skeptics have argued that any association between a single marker and an outcome would be meaningless by traditional statistical criteria. To address this problem, statisticians have quickly begun to apply decision-based analysis strategies to sort through gene expression data. For example, self-organizing maps, originally designed for functions such as voice recognition, and hierarchical clustering methods, long used in biological classifications, have been designed to focus on overall patterns of gene expression rather than on individual genes.

The most significant advances in the use of genomics to classify cancers so far have been made with leukemia and lymphoma. A study of myelogenous and lymphocytic leukemias at Harvard with gene arrays demonstrated the ability to find distinguishing gene expression profiles of previously defined classes (class distinction) as well as to rediscover the two classes by gene expression profiles alone (class discovery) [8]. In another study, a collaborative group from Stanford and the NCI found two molecularly distinct forms of diffuse large B-cell lymphoma (DLBCL) on the basis of gene expression profiles, and showed the two categories to have significantly different prognoses [9]. Interestingly, differentiation genes were major distinguishing features between the two subclasses of DLBCL, suggesting that the two types of DLBCL arise from different progenitor cells. These different patterns of differentiation can be distinguished by gene expression profiles but not by morphology.

Will gene expression profiles also help to distinguish between different phenotypes of lung cancer? The problem of classifying lung cancer might be more difficult than that for subclassifying hematopoietic neoplasms, because leukemia and lymphoma already have strong pre-existing classification frameworks, permitting the study of focused problems in taxonomy. In addition, tissue samples of leukemia and lymphoma typically have a great predominance of neoplastic cells, whereas tissue samples of lung cancer often have more lymphocytes and stromal cells than cancer cells. Thus, the analysis of lung cancer specimens for gene expression will not be straightforward.

To address this problem of impure lung cancer tissue samples, it will most probably be necessary first to purify the cancer cells from the heterogeneous mix. Laser capture microdissection is one technique that can be applied to the purification of the neoplastic cells [10], and recently a simple technique for scraping nearly pure clusters of neoplastic cells from tumor tissues was developed by the Gazdar laboratory at the University of Texas [11]. The development and utilization of such tissue-processing methods will be essential for the successful execution of molecular phenotyping projects.

Another key element for lung cancer classification projects will be the development of reliable and reproducible technologies for measuring the genes that are most important to lung cancer. One approach is to develop custom 'pneumochip' arrays that represent the genes expressed in respiratory epithelium and lung cancer. Already, collaborative efforts to develop such arrays are under way in major lung cancer research programs.

High-throughput analysis of genomic alterations in cancer

Although gene expression patterns are closely linked to a cell's function, it is ultimately genetic alterations that are responsible for the cancer phenotype. Although high-throughput technology for the analysis of mutations in genomic DNA is not as developed as the technology for the analysis of gene expression, some advances have been made in this area. In particular, oligonucleotide arrays can represent a series of different sequences for each gene, including wild-type and single-base mismatch sequences. This technology was found to be reasonably effective for detecting p53 mutations in lung cancers [12], but these arrays can only detect mutations in the specific genes and of the specific mutated sequences represented.

cDNA arrays have also been used as a substrate for comparative genomic hybridization (CGH) in pilot experiments [13], and this could obviously be applied to the study of lung cancer. With conventional CGH, differentially labeled test and reference genomic DNAs are co-hybridized to normal metaphase chromosomes, and fluorescence ratios along the lengths of these chromosomes are used to estimate variations in DNA copy number. Notably, CGH with arrays will probably have significantly higher resolution than CGH on metaphase chromosomes, and could detect changes in chromosomal copy numbers occurring at the single-gene level. Obviously, CGH by arrays could be combined with gene expression levels by arrays to find genes that have both an amplified gene copy number and overexpression.

The future of genomics in lung cancer research

In the next few years, a considerable effort will be made to classify lung cancers, to discover genes involved in lung cancer pathogenesis, and to study the biochemistry and function of those genes in lung cancers. The major tools for tissue processing, array production, and decision-based statistical analysis strategies all seem to be in place for these efforts. But clinical scientists, pathologists, molecular biologists, statisticians, and informatics specialists will all need to pull together to make lung cancer genomics programs successful. Collaborative efforts will probably be increasingly important in cancer research because sophisticated tools require specialized expertise.

Predictions beyond these obvious research targets are not so easy. Promising new high-throughput proteomics technologies might not only measure protein levels but might also recognize post-translational modifications of proteins. Integrating these measurements with those of gene expression could add a whole new dimension to our understanding of cancer. Furthermore, we must remember that our current focus of genomics on gene expression virtually ignores the 95% of the genome that does not encode proteins or regulatory information. Although the function of this vast amount of our genome is unknown, it is often thought to be involved in stabilizing chomosomes and thus should be considered a likely target for cancer-related aberrations.

References

1. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, Kerlavage AR, McCombie WR, Venter JC: **Complementary DNA sequencing: expressed sequence tags and human genome project**. *Science* 1991, **252**: 1651-1656
2. Strausberg RL, Buetow KH, Emmert-Buck MR, Klausner RD: **The cancer genome anatomy project: building an annotated gene index.** *Trends Genet* 2000, **16**:103–106.
3. Wang SS, Esplin ED, Li JL, Huang L, Gazdar A, Minna J, Evans GA: **Alterations of the PPP2R1B gene in human lung and colon cancer.** *Science* 1998, **282**:284–287.
4. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: **Serial analysis of gene expression.** *Science* 1995, **270**:484–487.
5. Hibi K, Liu Q, Beaudry GA, Madden SL, Westra WH, Wehage SL, Yang SC, Heitmiller RF, Bertelsen AH, Sidransky D, Jen J: **Serial analysis of gene expression in non-small cell lung cancer.** *Cancer Res* 1998, **58**:5690–5694.
6. Wang T, Hopkins D, Schmidt C, Silva S, Houghton R, Takita H, Repasky E, Reed SG: **Identification of genes differentially over-expressed in lung squamous cell carcinoma using combination of cDNA subtraction and microarray analysis.** *Oncogene* 2000, **19**:1519–1528.
7. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray.** *Science* 1995, **270**: 467–470.
8. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286**:531–537.
9. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling**. *Nature* 2000, **403**: 503-511.
10. Emmert-Buck MR, Bonner RF, Smith PD, Chuaqui RF, Zhuang Z, Goldstein SR, Weiss RA, Liotta LA: **Laser capture microdissection.** *Science* 1996, **274**:998–1001.
11. Maitra A, Wistuba II, Virmani AK, Sakaguchi M, Park I, Stucky A, Milchgrub S, Gibbons D, Minna JD, Gazdar AF: **Enrichment of epithelial cells for molecular studies.** *Nat Med* 1999, **5**: 459–463.
12. Ahrendt SA, Halachmi S, Chow JT, Wu L, Halachmi N, Yang SC, Wehage S, Jen J, Sidransky D: **Rapid p53 sequence analysis in primary lung cancer using an oligonucleotide probe array.** *Proc Natl Acad Sci USA* 1999, **96:**7382–7387.
13. Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D, Brown PO: **Genome-wide analysis of DNA copy-number changes using cDNA microarrays.** *Nat Genet* 1999, **23**:41–46.

Author's affiliation: Departments of Pathology and Oncology, The Johns Hopkins University School of Medicine, Baltimore, Maryland, USA

Correspondence: Departments of Pathology and Oncology, The Johns Hopkins University School of Medicine, 4940 Eastern Avenue, Baltimore, Maryland 21224, USA. Tel: +1 410 550 3668; fax: +1 410 550 0075; email: egabriel@jhmi.edu