GENES, CHROMOSOMES & CANCER 38:207–214 (2003)

RESEARCH ARTICLE

Expression Profiling Targeting Chromosomes for Tumor Classification and Prediction of Clinical Behavior

Yong-Jie Lu,[1,2] Daniel Williamson,[1] Rubin Wang,[1] Brenda Summersgill,[1] Sandrine Rodriguez,[1] Simon Rogers,[3] Kathy Pritchard-Jones,[2] Colin Campbell,[3] and Janet Shipley[1]*

[1]Molecular Cytogenetics, Section of Molecular Carcinogenesis, Institute of Cancer Research, Sutton, Surrey, United Kingdom
[2]Section of Paediatric Oncology, Institute of Cancer Research & Royal Marsden NHS, Sutton, Surrey, United Kingdom
[3]Department of Engineering Mathematics, University of Bristol, Bristol, United Kingdom

Tumors are associated with altered or deregulated gene products that affect critical cellular functions. Here we assess the use of a global expression profiling technique that identifies chromosome regions corresponding to differential gene expression, termed comparative expressed sequence hybridization (CESH). CESH analysis was performed on a total of 104 tumors with a diagnosis of rhabdomyosarcoma, leiomyosarcoma, prostate cancer, and favorable-histology Wilms tumors. Through the use of the chromosome regions identified as variables, support vector machine analysis was applied to assess classification potential, and feature selection (recursive feature elimination) was used to identify the best discriminatory regions. We demonstrate that the CESH profiles have characteristic patterns in tumor groups and were also able to distinguish subgroups of rhabdomyosarcoma. The overall CESH profiles in favorable-histology Wilms tumors were found to correlate with subsequent clinical behavior. Classification by use of CESH profiles was shown to be similar in performance to previous microarray expression studies and highlighted regions for further investigation. We conclude that analysis of chromosomal expression profiles can group, subgroup, and even predict clinical behavior of tumors to a level of performance similar to that of microarray analysis. CESH is independent of selecting sequences for interrogation and is a simple, rapid, and widely accessible approach to identify clinically useful differential expression.

INTRODUCTION

Tumor classification and prediction of clinical behavior are essential processes in the clinical management of patients, but for many tumors these pose considerable challenges. Tumors are associated with altered or deregulated gene products that affect critical functions such as cell division and differentiation. This may determine the morphological features and biological behavior of malignancies. Aberrant gene expression has been identified in many tumors over the last several decades. Recently, arrays of sequences representing many genes have permitted the simultaneous analysis of differential expression of genes to be investigated and the patterns used to classify tumors and predict clinical behavior (Golub et al., 1999; Schummer et al., 1999; Alizadeh et al., 2000; Notterman et al., 2001; Ramaswamy et al., 2001; Pomeroy et al., 2002; Shipp et al., 2002; van't Veer et al., 2002). The amount and complexity of the data generated by microarray approaches, although considerable, are limited to the sequences represented. There are practical problems to consider with the use of microarrays as well as recognized difficulties in reliably interpreting thousands of noisy data points to deduce significant biological information (Knight, 2001; Novak et al., 2002; Simons et al., 2003; Slonim, 2003). Other approaches to expression profiling include serial analysis of gene expression and massive parallel signature sequencing, which provides high resolution and quantitative data but requires high-through-put sequencing and is currently time consuming and expensive to apply to many samples (Velculescu et al., 1995; Brenner et al., 2000). Therefore, alternative and complementary approaches to expression profiling for tumor classification would be useful.

We previously developed and validated a rapid expression profiling technique targeting chromosomes, termed comparative expressed sequence

Y.-J. Lu and D. Williamson contributed equally to this work.

Supported by: Cancer Research United Kingdom; National Cancer Institute (Bethesda, MD); Freemasons' Grand Charity.

*Correspondence to: Dr. Janet Shipley, Molecular Cytogenetics, Institute of Cancer Research, Male Urological Cancer Research Centre, 15 Cotswold Road, Sutton, Surrey, SM2 5NG, United Kingdom. E-mail: janet.shipley@icr.ac.uk

Received 16 June 2003; Accepted 10 July 2003
DOI 10.1002/gcc.10276

hybridization (CESH), which identifies chromosome regions corresponding to differential gene expression (Lu et al., 2001). Differentially labeled probes derived from a test and a control sample are co-hybridized to normal metaphase chromosomes. The ratio between the fluorescence intensities at a chromosome location indicates the regions of the genome containing differentially expressed genes in a manner analogous to the way comparative genomic hybridization (CGH) detects genomic imbalances (Kallioniemi et al., 1992). The amount and complexity of the data generated by CESH is less than that generated by microarray approaches and is independent of selecting sequences for interrogation. CESH requires only nanogram quantities of RNA and so is readily applicable to small biopsy samples, as demonstrated in this study. Here we explore the ability of CESH profiles to classify tumors into different histological and prognostic groups and compare the level of performance with well documented and previously assessed microarray studies.

MATERIALS AND METHODS

Samples

Rhabdomyosarcoma samples from 39 patients and 6 derived cell lines were collected for this study. Among these, 26 cases were of alveolar histology and 19 cases were described as embryonal rhabdomyosarcomas. Nine cases of the rhabdomyosarcomas were small biopsy specimens confirmed to contain tumor cells. For the alveolar subtype of rhabdomyosarcoma, the *PAX/FOXO1A(FKHR)* fusion gene status was detected by reverse transcription polymerase chain reaction as previously described (Anderson et al., 2001). Primary tumor samples from 20 leiomyosarcoma and 21 prostate cancers were also used together with the rhabdomyosarcoma samples against a normal muscle reference sample for CESH analysis. Eighteen samples of favorable-histology Wilms tumors were taken, 14 at diagnosis including 8 from non-relapse cases. Normal lymphocytes were used as control for CESH analysis of the Wilms tumors.

Comparative Expressed Sequence Hybridization

CESH analysis was performed in the same manner as originally described (Lu et al., 2001). Briefly, total RNA was extracted and treated with DNase I (Ambion, Austin, TX) before reverse transcription by use of random hexamers and Superscript II (Invitrogen, Paisley, UK). The resulting cDNA was amplified and labeled with either FluoroRed or FluoroGreen dTTP (Amersham, Buckinghamshire, UK) by use of degenerate oligonucleotide primed-PCR. Differentially labeled test and control probes were co-hybridized to normal blood metaphase cells for 48 hr. The ratio of fluorescence intensity between test and control along the length of each metaphase chromosome was analyzed by use of standard comparative genomic hybridization analysis software (Digital Scientific, Cambridge, UK) after image capture with a cooled charge-coupled device camera attached to a fluorescence microscope. In self:self hybridizations, the average fluorescence intensity ratios and SD did not exceed 1.0 ± 0.2 along chromosome arms. An intensity ratio outside these limits at a particular chromosomal location in five good-quality metaphases was scored as a region harboring differentially expressed genes.

Microarray Analyses

Microarray analysis was carried out through the use of a previously described methodology (Lu et al., 2001; Clark et al., 2002). Clones collected for the 2p24 region plus a 5,265-clone Geneset (Institute of Cancer Research and Cancer Research, UK) were gridded onto microscopic slides, hybridized, and analyzed by use of the software Genepix (Axon Instruments, Foster City, CA).

Analysis of Predictive Accuracy by Use of Machine Learning Algorithms

The potential of the profiles produced by the CESH technique for classifying tumor groups, tumor subgroups, and tumors with different clinical characteristics was evaluated with a support vector machine (SVM); an efficient machine learning classifier. The SVM classifier is trained on a set of samples belonging to known classes (in our case, known groups of tumors) and test samples presented to evaluate the predictive ability. SVMs separate the data by maximizing the margin or closest distance between the datapoints belonging to two classes and the separating hyperplane. This geometric problem can be reduced to finding the solution of a constrained quadratic programming problem. This task is convex, thus giving a unique solution guaranteeing good generalization on binary classification problems. In our case, we have a multi-class classification problem and so we used a directed acyclic graph (DAG) approach in which multi-class classification is reduced to a series of binary classification tasks (Platt et al., 2000). To apply an SVM, each distinguishable region based on the chromosomal bands was scored (~320) as

normal, underexpressed, or overexpressed such that these became nominal variables. Regions were excluded from the analysis if they were altered in every sample and deduced to be attributed to the choice of control tissue through comparisons previously made to other normal tissues (e.g., 2q23–32 and Xp11–12 [Lu et al., 2001]).

To make maximum use of the data, we performed leave-one-out (LOO) estimations of prediction accuracy. Thus, the learning machine is trained on $(n - 1)$ samples, then asked to predict the class of the sample left out, with rotation of this excluded sample so that each sample is used once as the test example. LOO cross-validation has an expected low bias but may have high variance in the bias-variance tradeoff (Hastie et al., 2001). However, for a comparison of learning curves, which is our main objective here, we view LOO cross-validation test errors as sufficient. In determining LOO test errors the feature selection was performed only on the $(n - 1)$ training examples and did not implicitly include the test sample.

Recursive Feature Elimination

In this approach to feature selection (by use of an SVM), the least-effective features for class distinction were progressively removed through an iterative process. Thus, starting with all of the features, the feature with least influence in the classification function (Guyon et al., 2002) is identified and removed with each iteration. Features surviving longest in this process are the best discriminators for the SVM classifier to use. The utility of recursive feature elimination was twofold: first to identify regions that contain differentially expressed genes, which may be significant discriminators and worthy of further investigation, and second to create a "streamlined" classifier that can be tested on further samples.

Error Rate Estimation

For a wide variety of different algorithms (including SVMs with linear kernels), the functional relationship between test error rate $e(n)$ and training set size n is given by Zipf's Law:

$$e(n) = an^{-\alpha} + b$$

where a is the learning rate, α is the decay rate, and b is the minimal error rate achievable because of the existence of noise in the data. The dependency enables comparisons to be made for different experimental or algorithmic techniques. We have given this dependency as a function of sample size and extrapolated the error rate for a fixed size of 400 samples for each of the CESH classifications. These parameters were previously determined for published microarray experiments by use of SVMs trained with a linear kernel (Golub et al., 1999; Schummer et al., 1999; Alizadeh et al., 2000; Notterman et al., 2001; Ramaswamy et al., 2001; Pomeroy et al., 2002; Shipp et al., 2002; van't Veer et al., 2002; Mukherjee et al., 2003). We compared these data with a similar analysis of the CESH data to assess the performance of the CESH approach to classifying tumors.

RESULTS

CESH analysis was successfully applied to small biopsy samples of rhabdomyosarcoma available for this study involving as little as 20 ng of RNA. Examples of the CESH analysis indicating differential expression in rhabdomyosarcomas and leiomyosarcomas compared to normal muscle tissue are shown in Figure 1. The results of all of the CESH analysis are presented in the supplemental data (www.icr.ac.uk/home/lu/). CESH data for the three groups of tumors with clear histopathological diagnoses (45 rhabdomyosarcomas, 20 leiomyosarcomas, and 21 prostate cancers) identified tumor-specific regions differentially expressing genes relative to muscle. This included overexpression from 11q13 in 62% of prostate samples and from 2p23–24 in 62% of rhabdomyosarcomas. Other regions frequently indicated as differentially expressed in rhabdomyosarcomas are summarized in Table 1. Overexpression of genes from 4p15–16 (80%) and Xp22 (95%) and underexpression of genes from 11q14–23 (85%) were detected in leiomyosarcomas. The CESH data set for the prostate cancer, rhabdomyosarcomas, and leiomyosarcomas was subjected to a multi-class classification experiment by use of an SVM. Through use of a LOO approach, this gave the test error rate of 0.058 (5/86) (i.e., it was able to classify correctly 94.2% of test samples). By use of soft margins to compensate for the existence of noise in the data, arising from a non-zero training error, the test error rate was 0.035 (3/86; i.e., correct classification of 96.5% of test samples). This establishes CESH expression patterns as specific for each of these different types of tumor.

To assess whether the CESH profiles were distinctive for the different subtypes of rhabdomyosarcomas, data from the 16 alveolar cases with the *PAX/FOXO1A* fusion genes and the 19 embryonal rhabdomyosarcomas were analyzed. It was notable that both subgroups of rhabdomyosarcomas show

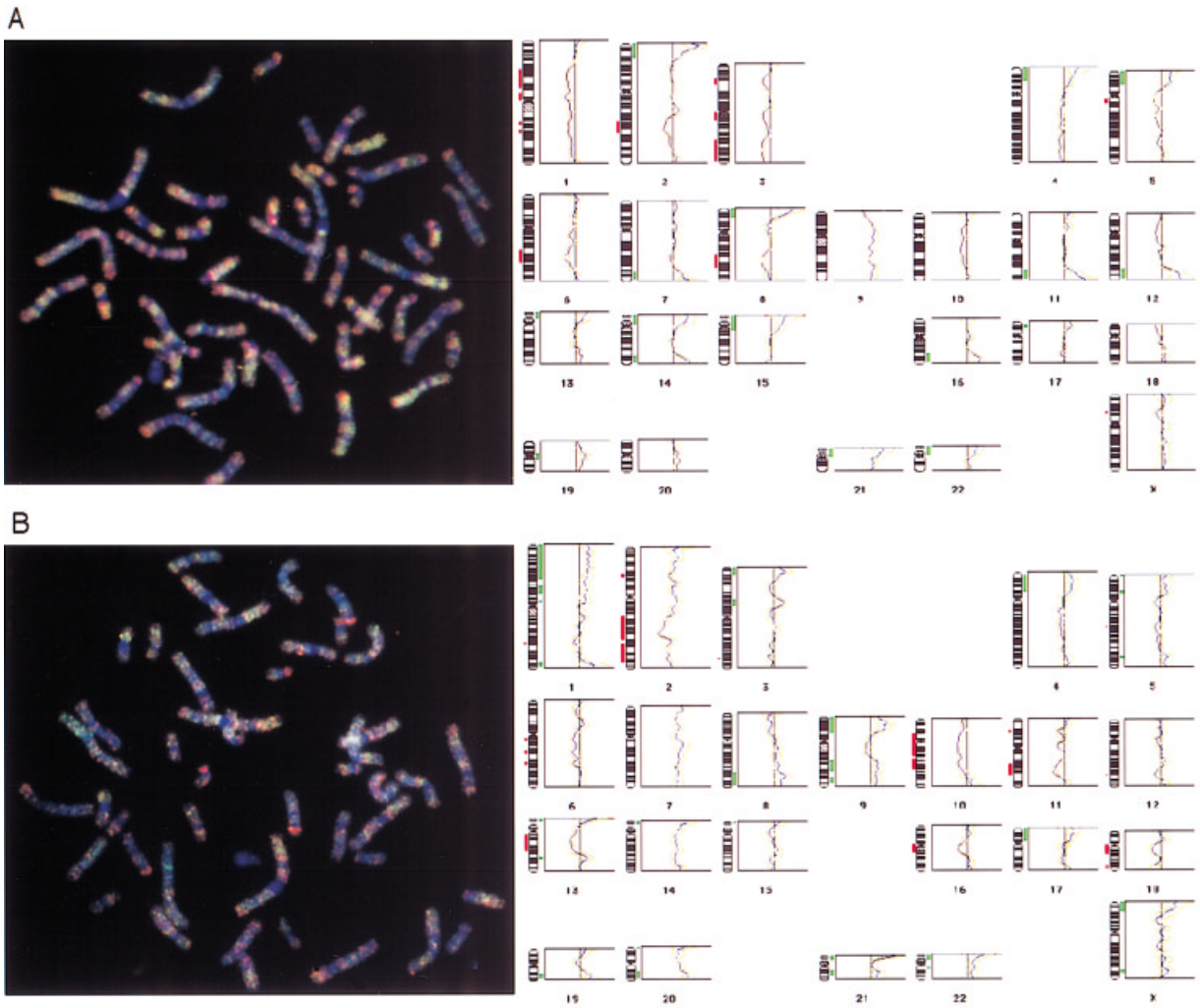

Figure 1. An example of expression profiling by use of CESH analysis. **A:** A rhabdomyosarcoma; **B:** a leiomyosarcoma. A representative image of each case is shown on the left. The average profile along the chromosomes from six cells from each case is presented on the right. Red bars on the left side of the chromosomes indicate regions with relative underexpression of genes, and green bars on the right side of the chromosomes show regions of relative gene overexpression.

common regions of frequent differential expression. In addition, the frequency of involvement of other regions was associated with one or other of the subtypes (Table 1). The data were investigated further by use of a binary classification SVM, which gave a LOO error rate of 0.17 (6/35; i.e., the SVM correctly classified new test samples 83% of the time). Perceptron neural networks were also applied and demonstrated a similar learning efficiency (data not shown). To determine which differentially expressed regions were most useful in distinguishing alveolar from embryonal, rhabdomyosarcoma (RMS) we used a recursive feature elimination method. Through the use of recursive feature elimination, we found that a feature size of 43 was able to achieve a LOO error of 4/35 (i.e., correctly classifying new test samples 88.6% of the time with zero training error). Distinguishing regions were identified that lie within the regions indicated in Table 1. In addition to these chromosome bands, the SVM also selected overexpression from other regions such as 5q35, 13q32, and Xq13 as being useful in classifying cases, although the frequency of involvement of these was lower. Microarray analysis of clones collected for the 2p23–25 region plus a 5,265 clone Geneset was carried out for five cases of rhabdomyosarcomas to identify some of the genes in the regions indicated by CESH. The genes found here and in the literature are shown in Table 1.

TABLE 1. Chromosome Regions of Frequent Differential Gene Expression Identified by CESH Analysis of Rhabdomyosarcomas Compared to Muscle Tissue and Candidate Genes for Involvement*

Chromosome region	Frequency of changes RMS (n = 35)	Alveolar (n = 16)	Embryonal (n = 19)	Genes altered in rhabdomyosarcomas
Regions common to subtypes				
2q14 ↑	51%	50%	53%	AMPHL
8p23 ↑	46%	56%	32%	
19q13 ↑	46%	50%	42%	**ZFP137,** TNNT1
3p22–24 ↓	46%	50%	42%	
6q21–23 ↓	80%	69%	89%	**Phospholamban,** connective tissue growth factor
10q21–22 ↓	51%	44%	58%	
Discriminatory regions				
2p23–25 ↑ (alveolar)[a]	66%	94%	42%	**POML, RHOE, DDEF2, LOC165323**, PIG3, MYCN, DDX1, SRC-1 (NCOA1), IMAGE 2008980, 1471296, 2072624, 2091812
5p14–15 ↑ (alveolar)[a]	26%	50%	5%	**TRIO, BASP1, IMAGE 1031125**, 246377,
12q24 ↑ (alveolar)[a]	31%	50%	16%	**HM74, P2RX4**, POLE
16q22–24 ↑ (alveolar)[a]	23%	44%	5%	**WWOX, LOC197256**
20q12–13 ↑ (alveolar)[a]	40%	56%	26%	**AIB2**, MYBL2, HSRANSEB, TFAP2C
8p12–q21 ↑ (embryonal)[a]	29%	0%	53%	Slug
8q23–24 ↑ (embryonal)[a]	43%	19%	63%	**MYC, PLEC1, FAK, Lipoprotein lipase, ANNEXIN XIII, LY6E, LOC14A1, IMAGE 1031991**, TAF2, NOV

* ↑ relative overexpression of genes in the chromosome region. ↓ relative underexpression of genes in the chromosome region.
[a]Regions that are more frequent in one or other of the subtypes and that are retained by recursive feature elimination in greater than 16 times out of 35 iterations at a feature size of 43. Genes in bold letters are those identified by our microarray analysis, and others are from the literature (Khan et al., 1998, 1999, 2001; Lu et al., 2001; Manara et al., 2002).

Binary classification by SVM analysis of CESH data for Wilms tumors (supplemental data on www.icr.ac.uk/home/lu/) gave a LOO error of 2/18, correctly classifying whether tumors would relapse 88.9% of the time. To determine which differentially expressed regions are most useful in distinguishing relapse from non-relapse, we again used a recursive feature elimination method (Guyon et al., 2002). At a feature size of 22, a LOO rate of 1/18 with zero training error (94.4% new test samples predicted correctly) was achieved. The effect of this feature selection is illustrated in Figure 2. Those features that survive longest and most frequently are likely to be significant discriminators and include overexpression from regions 1q41–43, 1q23, 18q21, 2p23, and 19q13.4, and underexpression from 1p22, 6p12, and 8q22.

The classification performance of our CESH analysis was compared to a previous analysis of high-profile microarray experiments that used a recently developed technique for determining the estimated dependency of test error on sample size (Mukherjee et al., 2003). Table 2 shows this error rate for the CESH analysis and the previously published error rates of the microarray studies.

DISCUSSION

Here, we demonstrate the ability of CESH profiling to distinguish tumor groups, subgroups, and tumors with different biological behaviors with a level of performance similar to that of microarray approaches. Generally, the CESH profiles appeared more consistent between samples and to involve smaller regions than the corresponding DNA profiles reported in previous CGH studies of the rhabdomyosarcomas and leiomyosarcomas (Fig. 1, Table 1, and supplementary data on www.icr.ac.uk/home/lu/) (Weber-Hall et al., 1996; Gordon et al., 2000; Wang et al., 2001). The expression patterns at the chromosomal level were shown to be distinctive for rhabdomyosarcoma, leiomyosarcoma, and prostate cancers—tumors with clearly different pathologies. We also investigated the rhabdomyosarcomas in more detail and found that the CESH patterns were distinctive in different subtypes. Rhabdomyosarcomas are a heterogeneous group of malignant tumors that resemble developing skeletal muscle and are mainly found in children. There are two main subtypes, known as embryonal and alveolar. The alveolar subtype is frequently associated with

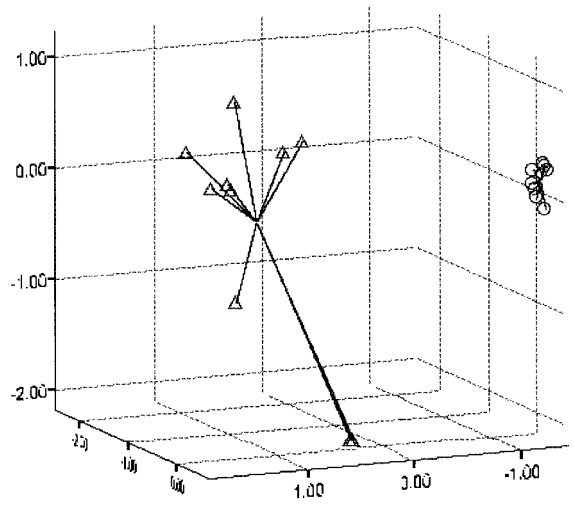

Figure 2. Multidimensional scaling plot representing the Wilms tumor expression data after RFE at a feature size of 22 (error = 1/18). The model was constructed from a squared Euclidean distance proximity measurement calculated from features retained in more than 10 out of 18 feature elimination iterations. ○ represents non-relapse cases and △ represents relapse cases. The tumors form two separate groups reflecting the selection of features that distinguish those tumors that subsequently relapsed from those that did not.

chromosome translocations and corresponding fusion gene products involving the *PAX3* or *PAX7* and *FOXO1A* genes (Anderson et al., 1999; Barr, 2001). To assess whether the CESH profiles were distinctive for these subtypes, data from the 16 alveolar cases with the *PAX/FOXO1A* fusion genes and the 19 embryonal rhabdomyosarcomas were analyzed. It was notable that both subgroups of rhabdomyosarcomas showed common regions of frequent differential expression. This suggests that the expression of at least some genes is common to the different subtypes. In addition, the frequency of involvement of other regions was associated with one or other of the subtypes (Table 1). Therefore, although the CESH profiles reflect underlying similarities between the two main subtypes of rhabdomyosarcomas, it is possible in the majority of cases to distinguish these subtypes on the basis of their chromosomal expression profiles.

Differentially expressed genes from the chromosomal regions indicated were identified through the limited microarray analysis carried out here and other microarray analyses reported in the literature (Table 1) (Khan et al., 1998, 2001; Lu et al., 2001; Manara et al., 2002). These may be genes involved in the development of rhabdomyosarcoma subtypes. Region-specific microarray analysis represents one approach to identifying genes that may be involved in the development of tumors or that are of use in discriminating between groups of tumors. The CESH data highlight chromosomal regions for further investigation that were not apparent in the microarray analysis, such as 8p23 and 10q21–22. This may be attributed to poor gene coverage on the microarrays for these regions. In addition, the microarray data show that more than one gene from a region is likely to be involved (Table 1). This is consistent with other microarray data (not shown) and recent literature that used serial analysis of gene expression data and may reflect a more general phenomenon of region-specific gene expression (Caron et al., 2001). This may result from co-regulation of functionally related genes or more general regulatory mechanisms, such as those involving chromatin conformation and epigenetic changes (Roy et al., 2002).

The most striking finding is that analysis of CESH data has the potential to indicate tumor behavior. This was revealed by comparing the relapse and the non-relapse Wilms cases. Wilms tumors are the most common renal neoplasm of childhood, and relapsed tumors, although rare, respond poorly to intensive second-line therapy. We previously showed an association between subsequent relapse and gain of genomic material from 1q and overexpression of genes from this region irrespective of genomic gain (Hing et al., 2001; Lu et al., 2002). Here, we demonstrate that the expression pattern as a whole was indicative of outcome.

To put the classification performance of our CESH analysis into context, the performance was compared to previous analyses of high-profile microarray experiments that used a recently developed technique for determining the estimated dependency of test error on sample size (Golub et al., 1999; Schummer et al., 1999; Alizadeh et al., 2000; Notterman et al., 2001; Ramaswamy et al., 2001; Pomeroy et al., 2002; Shipp et al., 2002; van't Veer et al., 2002; Mukherjee et al., 2003). Although the CESH comparisons here used different tumor samples and the published studies address a variety of classification issues, the error rates were similar. This demonstrates that CESH data are comparable in test performance to those of microarray studies.

Approaches to expression profiling include serial analysis of gene expression, massive parallel signature sequencing, and currently the most widely used approaches of microarray-based analyses (Velculescu et al., 1995; Golub et al., 1999; Schummer et al., 1999; Alizadeh et al., 2000; Brenner et al., 2000; Notterman et al., 2001; Ramaswamy et al., 2001; Pomeroy et al., 2002; Shipp et al., 2002; van't Veer et al., 2002). The first two of these approaches

TABLE 2. Comparison of CESH Classification Performance With Microarray Classification Performance

Learning curve $e(n) = an^{-\alpha} + b$	Error rate at n = 400	Analysis	Study and reference
$e(n) = 1.42n^{-0.52} + 0.0098$	7.278×10^{-2}*	Microarray	Variety tumors/normal (Ramaswamy et al., 2001)
$e(n) = 0.7706n^{-0.63} + 0.009$	2.668×10^{-2}*	Microarray	Leukemia AML/ALL (Golub et al., 1999)
$e(n) = 0.7362n^{-0.6864}$	1.205×10^{-2}*	Microarray	Ovarian tumor/normal (Schummer et al., 1999)
$e(n) = 0.57n^{-0.7073} + 0.0006$	8.831×10^{-3}*	Microarray	Lymphoma follicular/B-cell (Alizadeh et al., 2000)
$e(n) = 1.115n^{-0.3295} + 0.006$	1.608×10^{-1}*	Microarray	Medulloblastoma treatment outcome poor/successful (Pomeroy et al., 2002)
$e(n) = 0.9431n^{-0.2957} + 0.01$	1.704×10^{-1}*	Microarray	Lymphoma treatment outcome poor/succesful (Shipp et al., 2002)
$e(n) = 0.4852n^{-0.0733} + 0.01$	3.227×10^{-1}*	Microarray	Breast cancer treatment outcome metastatic/disease= free (van't Veer et al., 2002)
$e(n) = 0.4798n^{-0.2797}$	8.980×10^{-2}*	Microarray	Colon tumor/normal (Notterman et al., 2001)
$e(n) = 1.172n^{-0.575}$	3.739×10^{-2}	CESH	Rhabdomyosarcoma, leiomyosarcoma, prostate cancer (this study)
$e(n) = 0.655n^{-0.321}$	9.572×10^{-2}	CESH	Alveolar vs. embryonal rhabdomyosarcoma (this study)
$e(n) = 0.747n^{-0.544}$	2.869×10^{-2}	CESH	Wilms, treatment outcome (this study)

*Estimate taken from Mukherjee et al. (2002).

are technically demanding, require a lot of sequencing power, and, although they provide high resolution, quantitative data are time-consuming and expensive to apply to many samples. Microarray analyses have successfully addressed a number of classification issues, but it is recognized that there are methodological and statistical challenges and problems associated with the analysis of thousands of variable datapoints (Knight, 2001; Novak et al., 2002; Simons et al., 2003; Slonim, 2003). In addition, a study of matched mRNA that used two different microarray techniques showed that there was generally a poor correlation between the data from the different platforms, suggesting that probe-specific factors influence measurements (Kuo et al., 2002). The CESH technique has some advantages as an alternative or complement to microarray analysis. The use of chromosomes as a target involves no pre-selection of gene sequences for interrogation, as highlighted in a commentary on the technique (Chin, 2001). Although the data are less complex than those from microarrays, we have demonstrated here the remarkable ability of expression profiling at the chromosomal level to distinguish tumor groups, subgroups, and tumors with different clinical behaviors to a level of performance similar to that of recent high-profile microarray studies. Analysis is relatively rapid, making it possible to screen large numbers of samples, and it is particularly applicable to analysis of small amounts of tissue. The key regions identified by CESH could be the focus of further investigation, including analysis of samples identified with these regions involved. This could include the use of custom-designed microarrays representing all known genes for a chromosomal region and/or investigations of specific candidate genes and their products to identify differentially expressed genes that may be useful disease markers or novel targets for therapy. Existing fluorescence in situ hybridization expertise and equipment are appropriate to perform CESH analysis, and it is therefore readily accessible to many laboratories.

In conclusion, the analysis of CESH data described here represents a useful approach to addressing problems in classifying tumors that may be difficult by use of standard methods of diagnosis. It may also be able to predict behavior in other tumor types and lead to better clinical management of patients.

ACKNOWLEDGMENTS

We thank Cyril Fisher, Julia Bridge, Gyula Kovacs, Hartwig Huland, Neils Atkin, Colin Cooper, Sandra Edwards, Paul Grundy, Sandra Hing, Richard Williams, Iona Jeffrey, Anna Kelsey, and Steven Variend, who were involved in sample collection or preparation for this study. We are grateful to the United Kingdom Children's Cancer Study Group Tumor Bank and the National Wilms' Tumor Study Group and the Cooperative Tissue Network (Columbus, OH). We are also indebted to Colin Cooper, Barry Gusterson, Denise Sheer, and Mike Stratton for critical review of the manuscript and helpful comments.

REFERENCES

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. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403: 503–511.

Anderson J, Gordon A, Pritchard-Jones K, Shipley J. 1999. Genes, chromosomes, and rhabdomyosarcoma. Genes Chromosomes Cancer 26:275–285.

Anderson J, Gordon T, McManus A, Mapp T, Gould S, Kelsey A, McDowell H, Pinkerton R, Shipley J, Pritchard-Jones K. 2001. Detection of the PAX3-FKHR fusion gene in paediatric rhabdomyosarcoma: a reproducible predictor of outcome? Br J Cancer 85:831–835.

Barr FG. 2001. Gene fusions involving PAX and FOX family members in alveolar rhabdomyosarcoma. Oncogene 20:5736–5746.

Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, Roth R, George D, Eletr S, Albrecht G, Vermaas E, Williams SR, Moon K, Burcham T, Pallas M, DuBridge RB, Kirchner J, Fearon K, Mao J, Corcoran K. 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol 18:630–634.

Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, Hermus MC, van Asperen R, Boon K, Voute PA, Heisterkamp S, van Kampen A, Versteeg R. 2001. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. Science 291:1289–1292.

Chin G. 2001. Potential for tumor classification. Science 293:1015.

Clark J, Edwards S, John M, Flohr P, Gordon T, Maillard K, Giddings I, Brown C, Bagherzadeh A, Campbell C, Shipley J, Wooster R, Cooper CS. 2002. Identification of amplified and expressed genes in breast cancer by comparative hybridization onto microarrays of randomly selected cDNA clones. Genes Chromosomes Cancer 34:104–114.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286:531–537.

Gordon AT, Brinkschmidt C, Anderson J, Coleman N, Dockhorn-Dworniczak B, Pritchard-Jones K, Shipley J. 2000. A novel and consistent amplicon at 13q31 associated with alveolar rhabdomyosarcoma. Genes Chromosomes Cancer 28:220–226.

Guyon I, Weston J, Barnhill S, Vapnik V. 2002. Gene selection for cancer classification using support vector machine. Machine Learn 46:389–422.

Hastie T, Tibshirani R, Friedman J. 2001. The elements of statistical learning (Springer Series in Statistics). New York: Springer-Verlag, p. 193–222.

Hing S, Lu YJ, Summersgill B, King-Underwood L, Nicholson J, Grundy P, Grundy R, Gessler M, Shipley J, Pritchard-Jones K. 2001. Gain of 1q is associated with adverse outcome in favorable histology Wilms' tumors. Am J Pathol 158:393–398.

Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. 1992. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 258: 818–821.

Khan J, Simon R, Bittner M, Chen Y, Leighton SB, Pohida T, Smith PD, Jiang Y, Gooden GC, Trent JM, Meltzer PS. 1998. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. Cancer Res 58:5009–5013.

Khan J, Bittner ML, Saal LH, Teichmann U, Azorsa DO, Gooden GC, Pavan WJ, Trent JM, Meltzer PS. 1999. cDNA microarrays detect activation of a myogenic transcription program by the PAX3-FKHR fusion oncogene. Proc Natl Acad Sci USA 96:13264–13269.

Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS. 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 7:673–679.

Knight J. 2001. When the chips are down. Nature 410:860–861.

Kuo WP, Jenssen TK, Butte AJ, Ohno-Machado, Kohane IS. 2002. Analysis of matched mRNA measurements from two different microarray technologies. Bioinformatics 18:405–412.

Lu YJ, Williamson D, Clark J, Wang R, Tiffin N, Skelton L, Gordon T, Williams R, Allan B, Jackman A, Cooper C, Pritchard-Jones K, Shipley J. 2001. Comparative expressed sequence hybridization to chromosomes for tumor classification and identification of genomic regions of differential gene expression. Proc Natl Acad Sci USA 98:9197–9202.

Lu YJ, Hing S, Williams R, Pinkerton R, Shipley J, Pritchard-Jones K. 2002. Chromosome 1q expression profiling and relapse in Wilms' tumour. Lancet 360:385–386.

Manara MC, Perbal B, Benini S, Strammiello R, Cerisano V, Perdichizzi S, Serra M, Astolfi A, Bertoni F, Alami J, Yeger H, Picci P, Scotlandi K. 2002. The expression of ccn3(nov) gene in musculoskeletal tumors. Am J Pathol 160:849–859.

Mukherjee S, Tamayo P, Rogers S, Rifkin R, Engle A, Campbell C, Golub TR, Mesirov JP. 2003. Estimating dataset size requirements for classifying DNA microarray data. J Comput Biol 10: 119–142. (available at http://www.ai.mit.edu/people/sayan/new_pub.html).

Notterman DA, Alon U, Sierk AJ, Levine AJ. 2001. Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. Cancer Res 61:3124–3130.

Novak JP, Sladek R, Hudson TJ. 2002. Characterization of variability in large-scale gene expression data: implications for study design. Genomics 79:104–113.

Platt J, Cristianini N, Shawe-Taylor J. 2000. Large margin DAGS for multiclass classification. In: Solla SA, Leen TK, Muller KR, editors. Advances in neural information processing systems. Cambridge, MA: MIT Press, p. 547–553.

Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, Kim JY, Goumnerova LC, Black PM, Lau C, Allen JC, Zagzag D, Olson JM, Curran T, Wetmore C, Biegel JA, Poggio T, Mukherjee S, Rifkin R, Califano A, Stolovitzky G, Louis DN, Mesirov JP, Lander ES, Golub TR. 2002. Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 415:436–442.

Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Gerald W, Loda M, Lander ES, Golub TR. 2001. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci USA 98:15149–15154.

Roy PJ, Stuart JM, Lund J, Kim SK. 2002. Chromosomal clustering of muscle-expressed genes in Caenorhabditis elegans. Nature 418:975–979.

Schummer M, Ng WV, Bumgarner RE, Nelson PS, Schummer B, Bednarski DW, Hassell L, Baldwin RL, Karlan BY, Hood L. 1999. Comparative hybridization of an array of 21,500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas. Gene 238:375–385.

Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, Ray TS, Koval MA, Last KW, Norton A, Lister TA, Mesirov J, Neuberg DS, Lander ES, Aster JC, Golub TR. 2002. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med 8:68–74.

Simons R, Radmacher MD, Dobbin K, McShane LM. 2003. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 95:14–18.

Slonim D. 2003. From patterns to pathways: gene expression data analysis comes of age. Nat Genet 32:502–508.

van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. 2002. Expression profiling predicts outcome in breast cancer. Nature 415:530–536.

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. 1995. Serial analysis of gene expression. Science 270:484–487.

Wang R, Lu YJ, Fisher C, Bridge JA, Shipley J. 2001. Characterization of chromosome aberrations associated with soft-tissue leiomyosarcomas by twenty-four-color karyotyping and comparative genomic hybridization analysis. Genes Chromosomes Cancer 31:54–64.

Weber-Hall S, Anderson J, McManus A, Abe S, Nojima T, Pinkerton R, Pritchard-Jones K, Shipley J. 1996. Gains, losses, and amplification of genomic material in rhabdomyosarcoma analyzed by comparative genomic hybridization. Cancer Res 56:3220–3224.