A Gene Expression Signature Associated with Metastatic Outcome in Human Leiomyosarcomas

Yin-Fai Lee,† Megan John,† Alison Falconer,† Sandra Edwards,† Jeremy Clark,† Penny Flohr,† Toby Roe,† Rubin Wang,† Janet Shipley,† Robert J. Grimer,‡ D. Chas Mangham,‡ J. Meirion Thomas,‡ Cyril Fisher,‡ Ian Judson,† and Colin S. Cooper†

†The Male Urological Cancer Research Centre, Institute of Cancer Research, Surrey; ‡Department of Musculoskeletal Pathology, The Royal Orthopedic Hospital National Health Service Trust, Birmingham; and †Department of Histopathology, The Royal Marsden National Health Service Trust, London, United Kingdom

Abstract

Metastasis is a major factor associated with poor prognosis in cancer, but little is known of its molecular mechanisms. Although the clinical behavior of soft tissue sarcomas is highly variable, few reliable determinants of outcome have been identified. New markers that predict clinical outcome, in particular the ability of primary tumors to develop metastatic tumors, are urgently needed. Here, we have chosen leiomyosarcoma as a model for examining the relationship between gene expression profile and the development of metastasis in soft tissue sarcomas. Using cDNA microarray, we have identified a gene expression signature associated with metastasis in sarcoma that allowed prediction of the future development of metastases of primary tumors (Kaplan-Meier analysis P = 0.001). Our finding may aid the tailoring of therapy for individual sarcoma patients, where the aggressiveness of treatment is affected by the predicted outcome of disease.

Introduction

Soft tissue tumors are a heterogeneous group of mesenchymal tumors that arise as soft tissue masses and that frequently exhibit the differentiated features of adult soft tissue (1). Major histologic categories of malignant soft tissue sarcomas include leiomyosarcoma (smooth muscle), rhabdomyosarcoma (striated muscle), liposarcoma (fatty tissue), synovial sarcoma, and malignant fibrous histiocytoma. The disease accounts for ~1% of all cancers and is associated with a substantial mortality rate of ~50%, which is related in part to its propensity for metastasis (1, 2). The clinical behavior of soft tissue sarcomas is highly variable, but few reliable determinants of outcome have been identified (2, 3). New markers that predict clinical outcome, in particular the propensity of primary tumors to develop metastatic tumors, are urgently needed and would be of great clinical use, allowing for more selective treatment strategies. In this study, we have chosen leiomyosarcoma as a model to assess the relationship between gene expression profiles determined on cDNA microarray and the clinical outcome of metastasis in soft tissue sarcomas.

Materials and Methods

Microarray Procedures. Sarcoma tissue samples were collected from patients undergoing surgery at the Royal Marsden National Health Service Trust, London and the Royal Orthopedic Hospital National Health Service Trust, Birmingham, United Kingdom. Diagnoses were performed by pathologists with conventional criteria, immunohistochemistry, and electron microscopy. This study was done with approval from our local ethics committee. Tumor and control RNA preparation, cDNA microarray slide preparation, RNA labeling, and microarray hybridization were performed as in Lee et al. (4). Hybridized microarray slides were scanned in a GenePix 4000B scanner (Axon Instruments, Foster City, CA). Slides were scanned at photomultiplier tube voltage levels that provided a Cy5:Cy3 hybridization ratio across the slide of ~1. We used the GenePix Pro 3.0.6 software (Axon Instruments) to determine ratios of fluorescent intensities (Cy5:Cy3) for individual cDNA after subtraction of background. We had previously established the reliability of these microarray procedures; 12 of 12 genes showing alterations in expression in microarray exhibited similar alterations when examined by Northern blot analyses (5).

Analysis of Microarray Data. The scanned image was analyzed with the GenePix Pro 3.0.6 software (Axon Instruments). Fluorescent signals for both channels of the spots were determined. A local background in each channel was also determined for each spot, which is the median fluorescence of pixels in a halo surrounding the same array spot. Spots or areas of array with defects were flagged bad and excluded from subsequent analysis. To enhance the reliability of the expression data, another round of quality filtering was done. Spots with fluorescent spot intensity in each channel, which were >1.4 times the local background (medians) of that channel, were considered well measured (6), and the data were additionally filtered to include only these spots. The median background intensity was subtracted from the median spot intensity to generate the background-corrected signal intensity for use in additional analysis.

We used the GeneSpring software (Silicon Genetics, Redwood City, CA) to carry out additional microarray analyses. BRB-ArrayTools (Biometric Research Branch, National Cancer Institute, Bethesda, MD) was used for class comparison analyses (univariate F test/two-sample t test). Fluorescent intensity ratios of Cy5:Cy3 for individual spots of the filtered data were determined by dividing the background-corrected intensity for the Cy5 by that of the Cy3 channel. These ratios were then normalized by making the median of all measurements in each sample to be 1. Genes that have expression data in less than half of the samples were filtered out before the class comparison analysis. The samples were log2-transformed, and we compared the gene expressions of the 20 primary tumors (P) and 7 metastatic tumors (M) to find genes differentially expressed between the two classes using supervised class comparison analysis with a univariate F test/two-sample t test with randomized variance model and multivariate permutation tests to control the number of false discoveries (based on 1000 random permutations of the class labels of the experiments and controlling the number of false positive to be 30 genes 50% of the time, univariate P equals 0.0122).

Two-dimensional hierarchical clustering was then applied to the log-transformed data with average-linkage clustering with Pearson correlation around zero as the similarity metric for the 335 genes identified as differentially expressed between primary and metastatic sarcomas. This analysis divided thirty nonmetastatic tumors (P, PM, and LR) into two categories (groups 1 and 2). It is considered unlikely that exposure to chemotherapy would influence our analysis because only a single patient received chemotherapy that finished 5 weeks before surgery.

We refined the 335 differentially expressed gene list with two different supervised learning methods to find a reduced set of discriminating genes best...
for distinguishing the two groups (groups 1 and 2) of tumors. In one approach, to select genes for use in the classifier, all genes are examined individually and ranked on their power to discriminate the two classes (groups 1 and 2). For each gene, different cutoff points on the gene expression level for that gene are considered to predict class membership either above or below that cutoff. Genes are scored on the basis of the best prediction point for that class. The score function is the negative natural logarithm of the \( P \) value for the Fisher’s exact test of predicted versus actual class membership for group 1 versus group 2. To additionally check that the top-ranked discriminating genes could distinguish the two groups of tumors, we carried out supervised class prediction with the k-nearest-neighbor method and a leave-one-out cross-validation (to avoid overestimating the performance of the classifier) with the top-ranked discriminating genes. A selected number of top-ranked discriminating genes were used as the classifier for assigning the class membership of the left-out samples (using 1-, 3-, and 5-nearest-neighbor method with similarity measured by Euclidean distance between the samples). In a second approach, two-sample \( t \) tests were used to identify the genes that showed the most differential expression between the two prognostic groups of tumors.

**Other Statistical Analyses.** Other statistical analyses, including Fisher’s exact test and Kaplan-Meier analysis, were performed with SPSS (Chicago, IL). Fisher’s exact test was used for assessing the significance of association between categorical variables where appropriate. For the Kaplan-Meier analysis, metastasis was used as end points. Log-rank test was used to compare cluster groups.

**Results**

In this study, cDNA microarrays containing 5603 I.M.A.G.E. cDNA clones (5) were used to obtain genome-wide expression profiles for 37 leiomyosarcomas, including 20 primary tumors (P), 4 primary tumors that had concurrent metastases at the time of presentation (PM), 6 local recurrences of primary tumors (LR), and 7 metastatic sarcomas (M). In each experiment, Cy5-labeled sarcoma cDNA was cohybridized with Cy3-labeled reference cDNA from pooled human cell lines (4) that served as an internal standard for the comparison of different experiments. The array data were then filtered and normalized (4). As we were specifically interested in the expression differences that may characterize metastatic sarcomas, we compared the gene expression profiles of the 20 primary tumors (P) and 7 metastatic tumors (M) with supervised class comparison analysis with a univariate \( F \) test and multivariate permutation tests to control the number of false discoveries. The comparison gave rise to a list of 335 genes differentially expressed between the two types of tumors (Supplementary Table 1). Of these 335 genes, 228 genes had higher expression in metastatic tumors, and 107 genes had higher expression in primary tumors.

We were interested in whether the 335 genes associated with primary and metastatic distinction would be useful in classifying nonmetastatic tumors into groups that have different potential to develop metastasis and that would allow us to predict the future development of metastasis of primary tumors. We would only expect this to happen if the gene expression profile associated with metastasis is already present in some form in the bulk of cells in the primary tumors. The expression of the 335 genes was therefore used in hierarchical clustering to classify a group of 30 nonmetastatic tumors (P, PM, and LR; Fig. 1). The tumors were clustered into two distinct groups, with the expression profile of the two groups (groups 1 and 2) highly correlating with the original primary tumor versus metastatic tumor distinction, with group 1 having a more metastatic gene expression profile (Fisher’s exact test \( P < 0.001 \), Table 1). The distribution of the various tumor types (P, PM, and LR) in the two cluster groups is shown in Table 2. We predicted that the tumors with a more metastatic gene expression profile (group 1) would have a worse prognosis on disease progression to metastasis. Indeed, of the primary tumors (P) where follow-up data were available, all six primary tumors in cluster group 1 developed metastases, whereas only 3 of 11 tumors in group 2 developed metastases. Kaplan-Meier analysis on primary tumors (P) revealed a significant difference in the time to develop metastases in the two groups (log-rank test, \( P = 0.001 \); Fig. 2), with primary tumors in group 1 developing metastases much more rapidly than group 2 (mean time to develop metastases = 0.95 years in group 1 versus 5.18 years in group 2).

If membership of cluster group 1 predicts worse prognosis on disease progression to metastasis, we would also expect all other tumors in group 1 to rapidly progress to metastases. In agreement with this proposition, the remaining tumors in group 1 included all four primary tumors that had concurrent metastases at time of presentation (PM). In addition, the three local recurrent tumors (LR) present in this group from which clinical follow-up data were available had all given rise to metastatic tumors within 1 year. Hence, remarkably, all tumors (P, PM, and LR) in group 1 with clinical follow-up data were found to have or develop metastases. The observations from PM and LR tumors provide additional independent support of the finding that membership of group 1 is associated with the rapid development of metastasis. Notably, there is no statistical significant association of the cluster grouping with any of the published prognostic factors (2, 3, 7) associated with metastasis in sarcomas: (tumor size, \( \leq 5 \) or \( > 5 \) cm), Fisher’s exact test \( P = 0.602 \); site (superficial or deep), Fisher’s exact test \( P = 0.333 \), and grade (low or high), Fisher’s exact test, \( P = 1.000 \) (Supplementary Table 2); or the following demographic and clinical characteristics [age, Mann-Whitney test, \( P = 0.262 \); sex, Fisher’s exact test, \( P = 1.000 \); tumor site (retroperitoneal or nonretroperitoneal), and Fisher’s exact test, \( P = 0.722 \)]. Hence, other clinicopath-
scored on the basis of best cutoff point for discrimination by Fisher's exact test. Genes were selected with the two-sample t-test to be associated with distinction between groups 1 and 2. We hypothesized that not all of the genes in the subset of 335 genes contributed to the differences.

We were interested in identifying the genes best for distinguishing the two groups of tumors with different prognosis. We hypothesized that not all of the genes in the subset of 335 genes contributed to the distinction between groups 1 and 2, and we applied two different supervised learning methods in an attempt to identify the genes best associated with distinction between groups 1 and 2. Genes were scored on the basis of best cutoff point for discrimination by Fisher's exact test of predicted versus actual class membership. To confirm that the top-ranked discriminating genes could discriminate between the two prognostic groups of tumors, we carried out supervised class prediction with the k-nearest-neighbor method and a leave-one-out cross-validation with the top-ranked discriminating genes. A model with the top 80 discriminating genes (Supplementary Table 3 and Supplementary Fig. 1) accurately assigned the leave-out sample to the right group, with either 1- or 5-nearest-neighbor methods (30 of 30 samples in leave-one-out cross-validation test). This gene list could be additionally reduced to 20 genes with a slight loss of accuracy (29 of 30 samples in leave-one-out cross-validation test with a 1-, 3- or 5-nearest-neighbor methods). Additional reduction of the gene number to <20 resulted in additional loss of accuracy. We also additionally verified the gene selection with a second approach, where we carried out two-sample t tests to identify the genes that showed the most differential expression between the two prognostic groups. This gives rise to 99 genes with significant different expression ($P < 0.001$). In fact, we found that the genes selected with the two methods were in very good agreement; as many as 77 genes of the 80 most discriminating genes from method 1 were also present in this 99 significant gene list.

All except one of the 80 discriminating genes (MVP) had higher expression in group 1 than group 2 tumors. The 80 most discriminating genes included genes encoding proteins involved in biological processes associated with tumor development and invasion such as controlling cell growth and transition through the cell cycle (BMP2, PDAP1, CDC27, and CDK2AP1), signal transduction (IFNAR2, RIT1, GPSM1, GRB7, MAPKAPK2, and PAK2), apoptosis (BCL2A1), and nucleotide metabolism (GMPS).

Discussion

Genome-wide analysis of gene expression with microarray technology has become an important aid in the molecular diagnosis and classification of human malignancies, including soft tissue sarcomas (4, 6, 8–11). It has also been shown that gene expression analysis can be used for predicting clinical outcome of some cancers (12–15). Recently, Ramaswamy et al. (13) identified a 17-gene expression signature associated with metastasis and poor clinical outcome in a range of tumor types. It is interesting to note that although the 17-gene expression signature was of prognostic value in various adenocarcinomas, including lung adenocarcinoma, breast adenocarcinoma, prostate adenocarcinoma, and also in medulloblastomas, this gene signature did not predict outcome in diffuse large B-cell lymphomas, which they suggested to be related to the mesodermal origin of lymphomas. We found that 12 genes in the 17-gene signature were present in our array, and 11 of the 12 genes passed our filtering criteria. However, only one (HLA-DPB1) of them is included in the 335 genes showing differential expression in the primary and metastatic tumors, and none were included in our list of 80 predictor genes. Clustering analysis with these 11 genes on our sarcoma data set did not give clusters with significant difference in time to develop metastasis (Kaplan-Meier analysis log-rank test, $P = 0.90$), and class prediction with these genes did not work well. Because soft tissue sarcomas are also of mesodermal origin, these observations were consistent with the notion that the mechanism of metastasis in mesodermal tumors may be distinct from tumors of other embryonic origins.

Both van’t Veer et al. (12) and Ramaswamy et al. (13) have provided microarray data showing that an expression signature determining the propensity to metastasize can be detected in the primary epithelial tumor. However, there are several interpretations of these observations. Ramaswamy et al. (13) favor the view that the propensity to metastasize is a bulk property of the primary tumor, thus challenging the notion that metastases arise from rare cells within a primary tumor that have the ability to metastasize; and Bernards and Weinberg (16) have suggested that it is the combination of early oncogenic alterations within a tumor rather than some late alterations specific for metastasis that determine metastatic potential. Fidler and Kripke (17) have argued that these microarray-based observations simply reflect the heterogeneity of the primary tumor and do not prove that all of the genetic changes required for metastasis are present in individual tumor cells; hence, it may still be only the rare cells that have completed all of the steps in the metastatic process that give rise to a metastasis. It has also been suggested that host genetic background is a major determinant both of metastatic ability and of the metastasis-related expression profile within the primary tumor (18). Our microarray data capture the average gene expression of the tumor mass under study and do not theoretically exclude any of the above interpretations. Nonetheless, our microarray data do provide impor-

![Fig. 2. Kaplan-Meier curves (metastasis-free interval) of patients with primary tumors by cluster group. Log-rank test, $P = 0.001$.](image-url)
tant information on tumor prognosis and suggest that it is also possible to predict metastasis with the average gene expression profile found in the bulk of the primary mesenchymal tumor.

A number of the 80 discriminating genes or their closely related counterparts have been previously reported to be associated with metastasis, e.g., bone morphogenetic protein 2 (BMP2) had higher expression in a highly metastatic breast cancer cell line (19), and growth factor receptor-bound protein 7 (GRB7) signal transduction protein has been reported to contribute to the metastatic potential of cancer cells (20). In conclusion, we have identified a gene expression signature in leiomyosarcomas that is predictive of metastatic outcome for tumors at time of presentation. These findings could have important applications in the clinic, where the choice of how aggressively the patient should be treated is affected by the predicted outcome of disease and demonstrated the importance of genomics research in medicine.

Acknowledgments

Dr. Ian Giddings is thanked for his expert assistance with preparation of the supplementary information and minimum information about microarray experiment submission.

References

A Gene Expression Signature Associated with Metastatic Outcome in Human Leiomyosarcomas

Yin-Fai Lee, Megan John, Alison Falconer, et al.

Cancer Res 2004;64:7201-7204.