ORIGINAL ARTICLE

Discovery of molecular subtypes in leiomyosarcoma through integrative molecular profiling

AH Beck[1], C-H Lee[2], DM Witten[3], BC Gleason[4], B Edris[1], I Espinosa[1], S Zhu[1], R Li[1], KD Montgomery[1], RJ Marinelli[5], R Tibshirani[3], T Hastie[3], DM Jablons[6], BP Rubin[7], CD Fletcher[4], RB West[1,8] and M van de Rijn[1]

[1]Department of Pathology, Stanford University Medical Center, Stanford, CA, USA; [2]Department of Pathology, Genetic Pathology Evaluation Centre, Vancouver General Hospital and British Columbia Cancer Agency, Vancouver, British Columbia, Canada; [3]Department of Statistics, Stanford University Medical Center, Stanford, CA, USA; [4]Department of Pathology, Brigham and Women's Hospital, Boston, MA, USA; [5]Department of Biochemistry, Stanford University Medical Center, Stanford, CA, USA; [6]Department of Surgery, The University of California, San Francisco, CA, USA; [7]Department of Anatomic Pathology, Pathology and Laboratory Medicine Institute, The Cleveland Clinic and the Cleveland Clinic Lerner College of Medicine, Case Western Reserve University, Cleveland, OH, USA and [8]Department of Pathology and Laboratory Service, Palo Alto Veterans Affairs Health Care System, Palo Alto, CA, USA

Leiomyosarcoma (LMS) is a soft tissue tumor with a significant degree of morphologic and molecular heterogeneity. We used integrative molecular profiling to discover and characterize molecular subtypes of LMS. Gene expression profiling was performed on 51 LMS samples. Unsupervised clustering showed three reproducible LMS clusters. Array comparative genomic hybridization (aCGH) was performed on 20 LMS samples and showed that the molecular subtypes defined by gene expression showed distinct genomic changes. Tumors from the 'muscle-enriched' cluster showed significantly increased copy number changes ($P = 0.04$). A majority of the muscle-enriched cases showed loss at 16q24, which contains Fanconi anemia, complementation group A, known to have an important role in DNA repair, and loss at 1p36, which contains PRDM16, of which loss promotes muscle differentiation. Immunohistochemistry (IHC) was performed on LMS tissue microarrays ($n = 377$) for five markers with high levels of messenger RNA in the muscle-enriched cluster (ACTG2, CASQ2, SLMAP, CFL2 and MYLK) and showed significantly correlated expression of the five proteins (all pairwise $P < 0.005$). Expression of the five markers was associated with improved disease-specific survival in a multivariate Cox regression analysis ($P < 0.04$). In this analysis that combined gene expression profiling, aCGH and IHC, we characterized distinct molecular LMS subtypes, provided insight into their pathogenesis, and identified prognostic biomarkers.
Oncogene advance online publication, 9 November 2009; doi:10.1038/onc.2009.381

Introduction

Cancer can be broadly divided into three main classes: leukemias/lymphomas (derived from cells of the hematopoetic system), carcinomas (derived from epithelial cells) and sarcomas (derived from mesenchymal tissues, including bone, muscle and cartilage) (Abbas et al., 2005). Within each of these broad categories, tumors have traditionally been further subdivided into specific diagnostic subtypes based primarily on their clinical and histopathological features (Lakhani and Ashworth, 2001; Ackerman and Rosai, 2004). In the past decade, gene expression profiling has been used to discover novel cancer subtypes in a variety of hematological (Alizadeh et al., 2000; Bullinger et al., 2004) and epithelial (Sorlie et al., 2001; Lapointe et al., 2004) malignancies. It is hoped that the further subclassification of cancer based on molecular features will facilitate the identification of prognostic and predictive biomarkers (Sidransky, 2002; Beck et al., 2009b), the development of therapies targeted at oncogenic pathways altered in particular subtypes (Potti et al., 2006), and ultimately, the application of a more personalized form of medicine to improve the lives of cancer patients (Sotiriou and Piccart, 2007).

Soft tissue sarcomas account for approximately 1% of all malignances diagnosed annually, and there are ~100 recognized sarcoma diagnostic subtypes (Weiss and Goldblum, 2008). Sarcomas can be subdivided in two groups, one wherein each tumor type is characterized by a unique simple recurrent genetic abnormality such as a chromosomal translocation and the other in which highly complex genetic abnormalities are present (Helman and Meltzer, 2003). Leiomyosarcomas (LMS)

Correspondence: Professor M van de Rijn, Department of Pathology, Stanford University Medical Center, 300 Pasteur Drive, L235, Stanford, CA 94305, USA.
E-mail: mrijn@stanford.edu

belong to the latter group and are malignant neoplasms of smooth muscle, which most frequently occur in the uterus or retroperitoneum but can occur throughout the body (Fletcher *et al.*, 2002). They account for ∼24% of soft tissue sarcomas, making LMS the most common soft tissue sarcoma subtype (Toro *et al.*, 2006). Although significant advances have been made in the molecular understanding of sarcoma subtypes with simple recurrent genetic abnormalities (Helman and Meltzer, 2003), the molecular pathogenesis and heterogeneity of LMS are poorly understood. Currently, the diagnosis of LMS is made based on the demonstration of smooth muscle differentiation in a histologically malignant neoplasm (Fletcher *et al.*, 2002). Clinical management typically consists of surgery with adjuvant doxorubicin-based chemotherapy with consideration for the addition of ifosfamide and radiotherapy in selected cases (Borden *et al.*, 2003). Doxorubicin-based chemotherapy has shown a marginal association with improved overall survival (Cochrane Database Syst Rev, 2000) and the addition of ifosfamide has been shown to strengthen the association with improved survival (Pervaiz *et al.*, 2008). The overall prognosis for soft tissue sarcomas is poor, with reported rates of 12-year disease-specific survival (DSS) of 64% (Kattan *et al.*, 2002). There are currently no effective targeted therapies available for LMS that are directed at molecular aberrations in specific LMS subtypes. It has been shown that gene expression signatures may be able to predict metastasis in LMS (Lee *et al.*, 2004). Our laboratory has shown earlier that macrophage infiltration correlates with poor outcome in LMS (Lee *et al.*, 2008; Espinosa *et al.*, 2009). There are currently no molecular biomarkers used in the routine prognostication or determination of treatments in LMS.

In this study, we performed global gene expression analysis of 51 LMS samples to identify three distinct LMS subtypes. We then performed array comparative genomic hybridization (aCGH) to characterize the genomic changes seen in two of the LMS subtypes. On the basis of our gene expression findings, we identified biomarkers for the most distinct LMS subtype and evaluated their pattern of expression and association with clinicopathologic variables by performing immunohistochemistry (IHC) on tissue microarrays (TMAs) containing 377 LMS samples. In this integrative analysis, we characterized distinct molecular LMS subtypes, provided insight into their pathogenesis, and identified prognostic biomarkers.
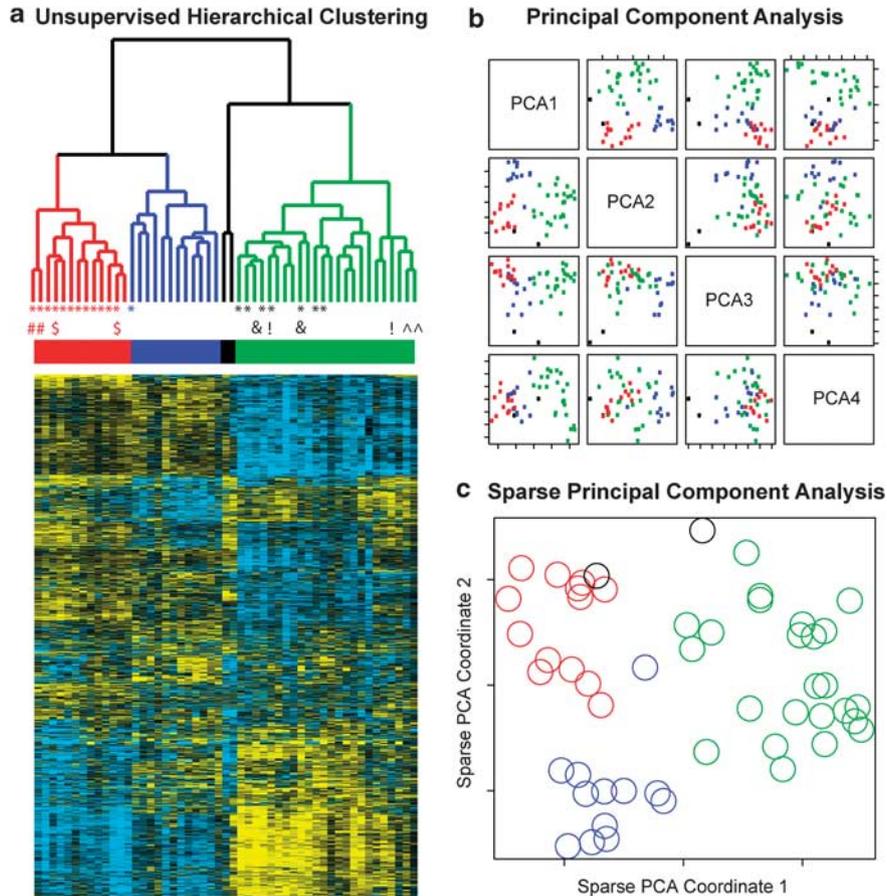
Results

Gene expression profiling

Unsupervised hierarchical clustering was performed on median centered gene expression data using the 3038 gene spots that had a s.d. of at least 1 across the 51 LMS samples and passed quality-based filtering criteria (described in Supporting Information Materials and Methods). This analysis showed three predominant clusters of LMS samples (Figure 1a). Cluster 1 contained 13 samples derived from 11 patients, cluster 2 contained 12 samples derived from 12 patients and cluster 3 contained 26 samples derived from 23 patients (Supplementary Information Table 1). For the five patients with paired primary, recurrent, and/or metastatic samples in the analysis, the matched pairs clustered into the same group (Figure 1a), suggesting that the molecular subtype is preserved during metastasis, as has been observed in other malignancies (Bernards and Weinberg, 2002; Ramaswamy *et al.*, 2003; Weigelt and van't Veer 2004; Weigelt *et al.*, 2005). To assess the similarity of each LMS sample to its cluster's centroid, we defined a centroid for the three predominant LMS clusters and assessed the correlation of each LMS sample with each LMS group centroid. In all, 48 out of 49 samples showed the highest correlation with the sample's LMS group centroid, which provides further support of the clusters' robustness (Supplementary Figure S1). Neither of the two 'outlier' cases from unsupervised hierarchical cluster analysis showed strong correlation with any of the LMS group centroids (correlation <0.35).

To visualize the variability within LMS by an additional unsupervised technique, we performed principal component analysis, which showed that the molecular subtypes observed in hierarchical clustering were largely captured in the first two principal components (Figure 1b). To identify a sparse set of genes composing the first two principal components, we performed sparse principal component analysis (Zou *et al.*, 2006) and identified 45 genes in component 1 and 40 genes in component 2. Plotting the LMS samples along these two sparse principal components largely recreates the clustering structure observed in the hierarchical clustering analysis (Figure 1c), and the classifications into three subtypes made using a simple nearest neighbor classifier based on these two principal components is concordant with the classifications made by unsupervised hierarchical clustering for 48 out of 49 LMS samples (Figure 1c).

Significance analysis of microarrays (Tusher *et al.*, 2001) was performed to identify sets of genes highly differentially expressed between the three LMS subtypes. The set of genes most highly and differentially expressed in LMS group I was significantly enriched for functional annotation terms relating to muscle contraction and the actin cytoskeleton (including CALD1, SLMAP, DMD, ACTG2, CASQ2,CFL2, MYLK and LPP) (Supplementary Information Table 1). As this gene set is highly enriched for genes encoding proteins involved in muscle differentiation and function, we refer to this LMS group as the 'group I/muscle-enriched' LMS molecular subtype. This gene set also showed significant enrichment for phosphoproteins, protein kinases and kinase-binding proteins (Supplementary Information Table 1). The set of genes most highly and differentially expressed in LMS group II was significantly enriched for functional annotation terms relating to protein metabolism, regulation of cell proliferation and organ development (Supplementary

a Unsupervised Hierarchical Clustering

b Principal Component Analysis

c Sparse Principal Component Analysis

Figure 1 Unsupervised clustering of 51 leiomyosarcoma (LMS) samples reveals three reproducible molecular subtypes. (**a**) Unsupervised hierarchical clustering was performed on 51 LMS samples with 3038 genes that showed at least one s.d. across the samples. The 20 samples that were also profiled for DNA copy number changes with array comparative genomic hybridization (aCGH) are indicated by an asterisk. The five paired primary-metastasis samples are indicated by a paired symbol (#, $, &, !, ^). On the sample dendrogram, the group I cases are highlighted in red, group II blue and group III green. The two cases that did not cluster into a group are indicated in black. Within the heatmap, yellow indicates relatively increased expression, black indicates median expression and blue indicates relatively decreased expression. (**b**) Principal component analysis of the 51 LMS samples with 3038 genes. Each sample is represented in the figure by a colored box. The color indicates the clustering designation made by hierarchical clustering: red = group I, blue = group II and green = group III. A majority of the variance between the three groups are captured in the first two principal components. (**c**) Sparse principal component analysis. The 51 LMS samples were plotted against the sparse principal component analysis (PCA) coordinate 1 (containing 45 genes) and sparse PCA coordinate 2 (containing 40 genes). Each sample is represented by a colored circle, and the color indicates the clustering designation made by hierarchical clustering: red = group I, blue = group II and green = group III. A majority of the variance between the three LMS molecular subtypes are explained by these two sparse principal components.

Information Table 1). The set of genes most highly and differentially expressed in LMS group III was significantly enriched for annotation terms relating to organ and system development, metal binding, extracellular proteins, proteins involved in the wound response and ribosomal proteins involved in protein synthesis (Supplementary Information Table 1). Of the three LMS subtypes, the group III gene set contained the maximum number of genes (23) and the highest proportion of genes from the CSF1 response gene expression signature (Beck et al., 2009a). In a previous study, the CSF1 response gene expression signature was shown to be present in a subset of LMS, and the expression of four CSF1 response signature-associated proteins was associated with poor prognosis in LMS (Espinosa et al., 2009).

To compare the pattern of gene expression seen in these LMS subtypes with other sarcomas, we performed unsupervised clustering of a large and diverse set of soft tissue tumors (STTs, n = 291, spanning 25 diagnostic subtypes) analysed in our laboratory with the 51 LMS samples using the same gene list that was used above to cluster only the LMS samples. In this analysis, samples from the group I/muscle-enriched cluster continue to cluster together, while cases from the other LMS clusters were interspersed in the dendrogram with other STT types (Supplementary Figure S2a). We performed sparse principal component analysis on the same set of LMS and STTs and similarly found that the group I/muscle-enriched LMS samples formed the most distinct cluster of LMS cases, whereas the other subtypes of LMS were intermixed with other STTs (Supplementary

Figure S2b). These results suggest that of the three LMS subtypes, group I/muscle-enriched shows the most distinct and specific gene expression profile.

To evaluate the reproducibility of the LMS molecular subtypes in an independent data set, we searched the gene expression omnibus for publically available sarcoma gene expression data sets and identified one data set containing > 15 LMS samples (Baird *et al.*, 2005 (GSE2553) contains 17 LMS samples). We used the clusterRepro algorithm (Kapp and Tibshirani, 2007) by training it on our LMS data set and testing it on the Baird set of 17 LMS samples. In this analysis, 12 of the 17 Baird LMS samples were classified as group I/muscle-enriched LMS, and the group I LMS achieved statistically significant cluster reproducibility (IGP = 1, $P = 0.03$). Only 1 of the 17 cases was classified as group II and 4 of the 17 cases were classified as group III. Neither LMS group II nor III achieved statistically significant reproducibility on this data set (IGP < 0.5, $P > 0.5$). These findings suggest that the group I/muscle-enriched LMS subtype is significantly reproducible in the Baird data set, and most LMS cases from this data set are best characterized as group I/muscle-enriched LMS (Supplementary Figure S3). Although the Baird data set provides no definite support for group II or III LMS, this may be partially explained by the data set's relatively small number of LMS samples and the relatively low level of variability observed in the LMS samples in this data set (Baird *et al.*, 2005).

Information on patient age, tumor grade, tumor histological subtype, primary tumor site, tumor status (primary, recurrence, metastasis) and patient treatment were available for the LMS cases in the gene expression analysis (Supplementary Information Table 1). There was no significant difference in patient age or tumor grade between the three molecular subtypes ($P > 0.15$). Group I LMS tended to be conventional LMS histological subtype (8/11, 73%), group II LMS were relatively evenly distributed between conventional (5/12, 42%) and pleomorphic/mixed (7/12, 58%) and group III LMS tended to be pleomorphic/mixed histological subtype (16/21, 76%) ($P = 0.03$). Group I LMS (10/11, 91% extra-uterine) and group II LMS (9/12, 75% extra-uterine) tended to have an extra-uterine primary tumor site, while group III LMS was relatively evenly distributed between uterine and extra-uterine primary sites (9/19, 47% extra-uterine) ($P = 0.04$). In all, 9 out of 13 group I samples and 10 out of 12 group II samples analysed by gene expression profiling were from the primary tumor site, while only 6 out of 23 group III tumors were from the primary site ($P = 0.002$) and the remainder came from a recurrence (7/23) or a metastasis (10/23). (Although there are a total of 24 group III samples, the site for one of the samples (STT4401) was not known to be primary or metastasis/recurrence, and this sample was excluded from this analysis). Information on patient treatment with radiotherapy and/or chemotherapy was available for 45 of the group I, II and III samples. There was no significant difference in proportion of treatment with radiotherapy and/or chemotherapy in the molecular subtypes: 11 out of 13

group I samples, 11 out of 12 group II samples and 14 out of 20 group III samples were resected from patients with no history of previous treatment with radiotherapy or chemotherapy ($P = 0.3$). Three patients contributed tumor samples removed pre- and post-treatment with chemotherapy and/or radiotherapy, and in all three cases the pre- and post-treatment samples clustered into the same LMS molecular subtype.

Array comparative genomic hybridization
Array comparative genomic hybridization was performed to characterize genomic changes in 20 LMS samples and 4 normal smooth muscle samples: 12 of the LMS samples were from group I/muscle-enriched (derived from 10 patients), 7 from group III (derived from six patients) and 1 from group II. To identify regions of genomic gain and loss, the fused lasso technique was performed (Tibshirani and Wang, 2008) with an FDR threshold of 10% for calling regions of gain/loss (Supplementary Information Table 1).

The group I/muscle-enriched and group III LMS samples showed distinct patterns of genomic gain and loss. The group I samples showed significantly increased genomic gains and losses compared with the group III samples (mean proportion of genome involved by gain/loss = 17% in group I samples vs 2% in group III samples, $P = 0.04$) (Figure 2). Analyzing the seven group III samples together showed no statistically significant shared regions of gain or loss (all consensus FDR > 5%). In contrast, analyzing the 12 group I/muscle-enriched samples together showed 691 spots that each had a consensus FDR < 5% (Supplementary Information Table 1). Taken together, these findings suggest distinct pathways of oncogenesis in the two LMS subtypes, with decreased genomic stability in group I/muscle-enriched LMS.

Interestingly, loss of a 291-kb region on 16q24 was seen in 7 of the 12 group I/muscle-enriched samples (all of which showed at least 7% genomic changes) and none of the group II (0/1) or group III (0/7) samples. This genomic region contains several cancer-associated genes, including the Fanconi anemia, complementation group A (*FANCA*) gene, which is a core Fanconi anemia protein that functions as a signal transducer and DNA-processing molecule in a DNA-damage repair network (Wang, 2007). In acute myeloid leukemia, it has been shown that acquired *FANCA* dysfunction promotes cytogenetic instability and clonal progression (Lensch *et al.*, 2003). The loss of *FANCA* may be an important event that is specific for the molecular pathogenesis of the group I/muscle-enriched LMS subtype and suggests an etiology for the increased genomic complexity observed in this LMS subtype. In addition, this region contains CBFA2T3, which is known to be involved in a translocation with *RUNX1* (*AML1*) in a subset of therapy related acute myeloid leukemia (Ottone *et al.*, 2009) and was identified as a putative breast tumor suppressor gene (Kochetkova *et al.*, 2002). The shared deleted region on 16q24 spans a total of 36 genes and includes *CDK10*, *TCF25*, *FOXF1* and *IRF8* in addition

Leiomyosarcoma Regions of Genomic Gain and Loss, n=20 (FDR = 0.10)
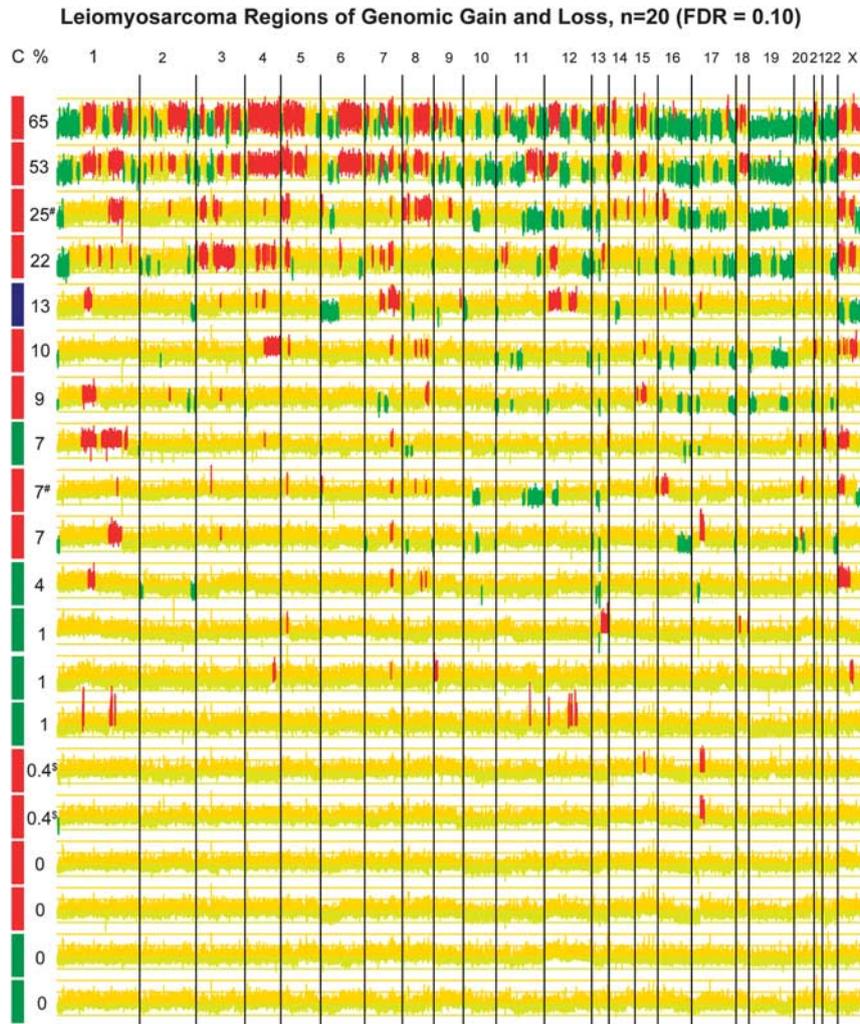
Figure 2 Array comparative genomic hybridization (aCGH) of 20 leiomyosarcoma (LMS) samples. The 20 samples are arranged along the *y* axis and ordered according to amount of DNA copy number changes. Chromosomal locations are indicated along the *x* axis. Copy number changes were called using the cghFLasso algorithm with an overall false discovery rate of 0.10. Regions of genomic gain are indicated in red and loss in green. The proportion of genome showing gain or loss is indicated to the left of each row. The gene expression defined molecular subtype is indicated on the color bar on the left: red = group I, blue = group II and green = group III. The group I cases show significantly increased regions of genomic gain/loss compared with the group III cases ($P = 0.04$).

to *FANCA*, *CBFA2T3* and others. The full list of genes can be found in the Supplemental Workbook.

The most commonly shared region of gain or loss in the group I/muscle-enriched cases was a 2.5-Mb region on 1p36.32, which spans PRDM16, TNFRSF14, C1orf93 and MMEL1. This region was lost in 8 of the 12 group I/muscle-enriched samples (consensus FDR = 0.01). This change was specific to this LMS subtype and there was no loss at 1p36.32 observed in the one group II sample or in the seven group III samples. The *PRDM16* gene has recently been shown to control a brown fat/skeletal muscle switch. Loss of PRDM16 from brown fat precursors promotes skeletal muscle differentiation and leads to elevated expression of muscle specific genes (Seale *et al.*, 2008). Group I LMS showed high expression of a diverse set of muscle-associated genes, including genes expressed in smooth, cardiac and skeletal muscle (including *ACTG2*, *MYLK*

and *PDLIM5*), genes expressed primarily in cardiac and skeletal muscle (including *CFL2*, *SLMAP*), and genes expressed primarily in cardiac muscle (*CASQ2*). The loss of *PRDM16* gene in most group I cases suggests a potential etiology of the 'muscle-enriched' pattern of gene expression observed in group I tumors, which includes both genes expressed in smooth and skeletal/cardiac muscle, suggesting either that PRDM16 may have a role in expression of genes involved in skeletal, smooth and cardiac muscle or a separate transcriptional regulatory factor may account for the increased expression of smooth muscle and cardiac muscle associated genes in group I/muscle-enriched LMS. *MYOCD* amplification has recently been shown to have an important role in LMS pathogenesis (Perot *et al.*, 2009). In our study, a region on 17p11 containing MYOCD (as well as MAP2K4) was amplified in three group I/muscle-enriched tumors and the one group II sample.

Other cancer-associated genes that showed copy number gains in at least six of the group I samples included: TCF12 (15q21), ABL2 (1q24) and the MET oncogene (7q31). MET overexpression has been reported earlier in a variety of sarcomas, including alveolar soft part sarcoma (Jun et al., 2009), osteosarcoma, chondrosarcomas and LMS (Rong et al., 1993). Copy number gain suggests a possible mechanism of MET overexpression in LMS. Additional cancer-associated genes that were lost in at least six group I samples include the alveolar soft part sarcoma chromosome region, candidate 1 (ASPSCR; 17q25), BCL3 (19q13), ERCC2 (19q13), FSTL3 (19p13), RB1 (13q14), STK11 (19p13) and TCF3 (19p13) (An expanded table of cancer-associated genes with the copy number changes observed in our study is provided in the Supplemental Workbook).

There were no recurrent genomic changes seen in > 3 of the 7 group III samples, and all of the changes seen in multiple group III samples were also observed in multiple group I/muscle-enriched samples. The only change involving a known gene that was shared in at least 3 of the 7 group III cases was gain at 7q31.2, which includes the CAV-1 gene. This region was also gained in 7 of the 12 group I/muscle-enriched cases and the one group II case. Caveolin-1 is known to be expressed on smooth muscle and has been shown to activate the Akt pathway in an in vitro prostate cancer model (Li et al., 2003). The Akt pathway has been shown to have an important role in LMS (Hernando et al., 2007).

To determine whether the gene expression defined molecular subtypes could be accurately predicted based solely on aCGH changes, we used three classification techniques: prediction analysis of microarrays, prediction analysis of microarrays-FusedLasso (PAM-FL) and K nearest neighbor. The aCGH-based K nearest neighbor classifier obtained a cross validation error rate of 4 out of 19 (21%, corresponding to a permutation-based P-value of 0.05), the prediction analysis of microarrays classifier obtained a cross validation error rate of 2 out of 19 (11%) and the PAM-FL obtained a cross validation error rate of 1 out of 19 (5%). These findings show that the gene expression defined groups I and III can be predicted with significant accuracy by aCGH changes alone. The PAM-FL centroid, which summarizes genomic changes in group I relative to group III is presented as Supplementary Figure S4.

Tissue microarray analysis

On the basis of the findings from gene expression profiling (which showed that group I/muscle-enriched LMS represented the most distinct molecular subtype) and aCGH (which showed that group I/muscle-enriched LMS had the most recurrent regions of genomic gain and loss), we chose to focus our TMA analysis on evaluating the protein expression of genes with high levels of messenger RNA expression in the group I/muscle-enriched LMS molecular subtype.

The protein expression of five markers highly and differentially expressed in group I/muscle-enriched LMS was examined on LMS TMAs. For the 377 LMS cases represented on the TMAs, there were evaluable results for all five stains for 275 cases. For the purposes of clinicopathologic analysis, we determined the sum total of positive staining markers for each case. Clinicopathologic data (including FNLCC histologic grade, mitotic count, necrosis and presence of CSF1-response protein expression signature (Beck et al., 2009a; Espinosa et al., 2009)), and DSS were available for 124 of the 275 cases. Information on anatomic site (gynecological vs non-gynecological) was available for 273 samples.

The stains showed significant correlation with each other (all pairwise Spearman's rho $P < 0.005$) with a minimum correlation of 0.170 between ACTG2 and CASQ2 and a maximum correlation of 0.658 between ACTG2 and SLMAP (Figure 3). In all, 19% (51/275) of the cases showed coordinate expression of all five evaluable markers, similar to the 25% of cases present in the muscle-enriched cluster by gene expression arrays.

The number of positive markers showed no association with site (mean positive markers = 3.1 in both gynecological and non-gynecological LMS), no significant association with grade (mean positive markers = 3.1 in grade 1, 3.3 in grade 2 and 2.9 in grade 3; $P = 0.56$), no significant association with mitotic figure count ($P = 0.335$), and there was a trend for a negative association with the presence of necrosis ($P = 0.07$). In a multivariate model incorporating the CSF1-response protein expression signature, the number of positive group I/muscle-enriched markers, grade, tumor site (uterine vs extra-uterine), necrosis and mitotic figures, the CSF1 response signature (summarizing the expression of four CSF1 response signature-associated proteins) and the number of positive group I/muscle-enriched markers were the only two significant predictors of survival, with the CSF1-response protein expression signature showing an association with poor outcome whereas expression of the group I/muscle-enriched markers was associated with a more favorable outcome (Table 1).

The findings from our TMA analysis show that the group I/muscle-enriched markers show correlated protein expression, and the expression of group I/muscle-enriched markers is associated with improved DSS independent of grade, mitotic figures, necrosis, site and the CSF1 response signature.

Discussion

Leiomyosarcoma is an aggressive malignant neoplasm, and its molecular pathogenesis is poorly understood. Treatment options are limited, and there is a major clinical interest in gaining a better understanding of LMS pathogenesis to facilitate the development of targeted therapies.

Several previous studies have performed gene expression profiling on relatively small numbers ($n = 3$–13) of LMS samples (Nielsen et al., 2002; Shmulevich et al., 2002; Ren et al., 2003; Segal et al., 2003; Skubitz and
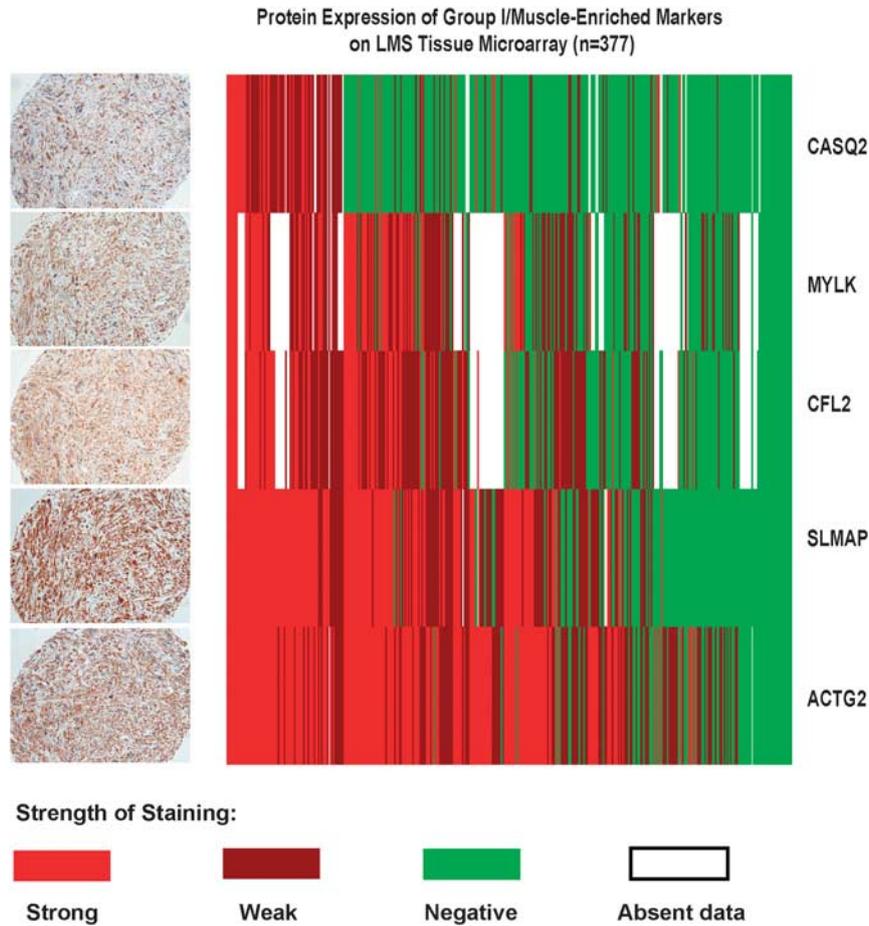
Figure 3 Protein expression of group I markers on leiomyosarcoma (LMS) tissue microarray (TMA). We performed immuno-histochemistry (IHC) for five markers that showed high levels of expression in group I LMS in the gene expression analysis (CASQ2, MYLK, CFL2, SLMAP and ACTG2). The LMS TMAs contained a total of 377 samples that were scored as strong positive (bright red in the heatmap), weak positive (dull red) or negative (green). The antibodies are listed along the *y* axis and the 377 samples along the *x* axis. Missing data are indicated by white in the heatmap. Pictures of an LMS sample showing strong expression of all five stains is shown to the left of the heatmap (magnification = 200 ×). The five stains showed significantly correlated expression (all pairwise Spearman's rho $P < 0.005$, with a minimum correlation of 0.17 between ACTG2 and CASQ2 and a maximum correlation of 0.66 between ACTG2 and SLMAP).

Table 1 Multivariate Cox regression analysis of disease-specific survival from LMS tissue microarray ($n = 124$)

	Wald	HR (95% CI)	P-value
CSF1-response protein signature (Espinosa *et al*., 2009)	13.0	4.7 (2.0–10.9)	<0.001
Number of positive group I/muscle-enriched markers	4.5	0.77 (0.6–0.98)	0.035
Site	1.8	0.35 (0.1–1.6)	0.178
Mitotic figures	0.8	1.3 (0.7–2.4)	0.371
Grade	0.32	1.3 (0.6–2.5)	0.570
Necrosis	0.01	1.0 (0.5–2.2)	0.910

Abbreviations: CI, confidence interval; HR, hazard ratio; LMS, leiomyosarcoma.
In the multivariate model, only the CSF1-response protein expression signature and the expression of group-I/muscle-associated proteins showed a significant association with survival. In the table, the first column lists each variable included in the model, the second column lists the variable's Wald test statistic, the third column lists the variable's HR and 95% CI and the fourth column lists the variable's Wald test *P*-value.

Skubitz, 2003; Quade *et al*., 2004; Baird *et al*., 2005; Henderson *et al*., 2005; Nakayama *et al*., 2007). Owing to the small number of cases in each study it is difficult to draw conclusions on the heterogeneity within LMS based on these data. Francis *et al*. (2007) performed gene expression profiling on 177 STTs, including 40 LMS samples. They identified a distinct cluster of 11 LMSs that clustered together, while the remaining 29 LMS samples showed more heterogeneous patterns of gene expression. The distinct cluster of 11 LMS cases

from this data set were reported to show high expression levels of a group of muscle-associated genes, many of which were also identified as highly expressed in group I/muscle-enriched LMS in our study (including CALD1, SLMAP, ACTG2, CFL2, MYLK, ACTA2, MBNL1, TPM1, PPP1R12A, DTNA, FZD6, PPP1R12A, CLIC4, CDC42EP3, BARD1, TPM1, RAB27A, MAP1B and EDIL). We find a similar muscle-enriched LMS cluster in our data set and in the Baird data set (Baird *et al.*, 2005). Our findings and those from the literature suggest that multiple molecular subtypes of LMS exist and that the 'muscle-enriched' subtype has been reproducibly identified in at least two of the largest data sets.

Several previous reports have looked at comparative genomic hybridization changes in LMS. Meza-Zepeda *et al.* (2006) performed aCGH on 12 LMS samples and 7 gastrointestinal stromal tumors and observed that LMS showed more genomic losses than gains with the most frequent minimal regions of loss at 10q21.3 and 13q14.2–q14.3, each detected in 9 of the 12 LMS samples in their study. In our study, we identified loss at 10q21.3 in 5 of the 12 group I/muscle-enriched samples and in none of the group II or III samples. We identified loss at 13q14.2 in 6 of the 12 group I/muscle-enriched samples, 0 of the 1 group II samples and 2 of the 7 group III samples. The common region of 13q14.2 that was lost in all eight samples includes the *RB1* gene, a well-characterized tumor-suppressor whose loss has been shown to contribute to sarcomagenesis (Landis-Piwowar *et al.*, 2008). Meza-Zepeda *et al.* (2006) also noted loss at 16q21.2–q22.1 in 6 of the 12 samples and 1p36.32–p36.21 in 4 of the 12 samples, which are both changes we find in our study, specifically in group I/muscle-enriched LMS. Larramendy *et al.* (2008) evaluated 102 malignant fibrous histiocytomas and 82 LMS cases by conventional comparative genomic hybridization and identified 11 regions with significantly increased losses in LMS compared with malignant fibrous histiocytomas, including 1p36.1 ∼ pter (10% of LMS vs 1% of malignant fibrous histiocytomas), and 16qter (34% of LMS vs 3% of malignant fibrous histiocytomas), both of which were identified as lost in most group I/muscle-enriched LMS cases in our analysis. It is to be noted that the 1p36 region contains *PRDM16* and the 16q24.3 region contains *FANCA*. To our knowledge, our study is the first to integrate aCGH data with gene expression analysis.

Prognosis in LMS is currently predicted using a combination of traditional clinicopathologic features (Kattan *et al.*, 2002). There are currently no molecular biomarkers used in prognostication in LMS in clinical practice. Gene expression microarrays have been used to identify signatures to predict metastasis in LMS (Lee *et al.*, 2004). Our group has previously identified macrophage infiltration (Lee *et al.*, 2008) and the CSF1 response signature (Espinosa *et al.*, 2009) as predictors of poor prognosis in LMS. In this study, we have identified protein markers from the group I/muscle-enriched LMS subtype and showed that their expression correlates with improved DSS. These findings suggest that despite showing increased genomic

complexity, group I/muscle-enriched LMS may be intrinsically less aggressive and more differentiated than other LMS subtypes. In a multivariate model, incorporating traditional clinicopathologic features (size, grade, necrosis and site) as well as the CSF1 response signature and the group I/muscle-enriched markers, we find that only the CSF1 response signature and the number of positive muscle-enriched markers emerged as significant predictors of survival, with the CSF1 response signature correlating with poor prognosis and the expression of group I/muscle-enriched markers correlating with improved prognosis. These prognostic biomarkers, which can be measured with IHC on paraffin-embedded formalin fixed tissue, may prove useful for the clinical management of LMS. Ultimately, we hope that the characterization of distinct molecular subtypes in LMS will lead not only to the identification of clinically useful prognostic markers, but also to the development of treatments to target-specific molecular aberrations observed in the subtypes.

Materials and methods

The 51 tumor samples were obtained from 46 patients (five patients each contributed two samples). Clinicopathologic features of these tumors are provided in Supplementary Information Table 1. The studies were performed using the approval by the Institutional Review Board at Stanford University Medical Center.

Briefly, gene expression profiling was performed on 51 LMS samples using 44K spotted complementary DNA microarrays. To identify molecular subtypes, unsupervised hierarchical clustering and principal component analysis were performed. To assess the reproducibility of the clusters in an independent data set, the clusterRepro algorithm (Kapp and Tibshirani, 2007) was used with the LMS samples from the GSE2553 data set (Baird *et al.*, 2005) used as the testing data set. For 20 of the LMS samples containing gene expression data, aCGH was performed on 44K Agilent arrays. The fused lasso algorithm was applied to identify regions of copy number gain and loss (Tibshirani and Wang, 2008). To determine whether the gene expression defined molecular subtypes could be accurately predicted based solely on aCGH changes, three classification techniques were used: prediction analysis of microarrays (Tibshirani *et al.*, 2002), PAM-FL (See Supporting Information Materials and Methods for explanation of PAM-FL), and K nearest neighbor. IHC was performed on LMS TMAs using antibodies for ACTG2, CASQ2, SLMAP, CFL2, and MYLK, and the stains were scored by two surgical pathologists (AHB and RBW). Additional information on the methods for gene expression profiling, aCGH, IHC and statistical analysis are provided in the Supporting Information Materials and methods. All IHC images used in this study are accessible from the accompanying website: http://tma.stanford.edu/tma_portal/LMS_IMP. In addition, gene expression and aCGH data have been deposited in the Gene Expression Omnibus (Edgar *et al.*, 2002) with accession number GSE17555.

Conflict of interest

The authors declare no conflict of interest.

References

Abbas AK, Cotran RS, Fausto N, Kumar V, Perkins JA, Robbins SL. (2005). *Robbins and Cotran Pathologic Basis of Disease [print/digital]*, Vol xv, 7th edn. Elsevier/Saunders: Philadelphia, 1525, pp.

Ackerman LV, Rosai J. (2004). *Rosai and Ackerman's Surgical Pathology [print]*, 9th edn. Mosby: Edinburgh; New York.

Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**: 503–511.

Baird K, Davis S, Antonescu CR, Harper UL, Walker RL, Chen Y et al. (2005). Gene expression profiling of human sarcomas: insights into sarcoma biology. *Cancer Res* **65**: 9226–9235.

Beck AH, Espinosa I, Edris B, Li R, Montgomery K, Zhu S et al. (2009a). The macrophage colony-stimulating factor 1 response signature in breast carcinoma. *Clin Cancer Res* **15**: 778–787.

Beck AH, West RB, van de Rijn M. (2009b). Gene expression profiling for the investigation of soft tissue sarcoma pathogenesis and the identification of diagnostic, prognostic, and predictive biomarkers. *Virchows Arch* (e-pub ahead of print 2 May 2009; PMID: 19412622).

Bernards R, Weinberg RA. (2002). A progression puzzle. *Nature* **418**: 823.

Borden EC, Baker LH, Bell RS, Bramwell V, Demetri GD, Eisenberg BL et al. (2003). Soft tissue sarcomas of adults: state of the translational science. *Clin Cancer Res* **9**: 1941–1956.

Bullinger L, Dohner K, Bair E, Frohling S, Schlenk RF, Tibshirani R et al. (2004). Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *N Engl J Med* **350**: 1605–1616.

Edgar R, Domrachev M, Lash AE. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**: 207–210.

Espinosa I, Beck AH, Lee CH, Zhu S, Montgomery KD, Marinelli RJ et al. (2009). Coordinate expression of colony-stimulating factor-1 and colony-stimulating factor-1-related proteins is associated with poor prognosis in gynecological and nongynecological leiomyosarcoma. *Am J Pathol* **174**: 2347–2356.

Fletcher CDM, Unni KK, Mertens F, World Health Organization, International Agency for Research on Cancer. *Pathology and Genetics of Tumours of Soft Tissue and Bone*. IARC Press: Lyon, 427 pp.

Francis P, Namlos HM, Muller C, Eden P, Fernebro J, Berner JM et al. (2007). Diagnostic and prognostic gene expression signatures in 177 soft tissue sarcomas: hypoxia-induced transcription profile signifies metastatic potential. *BMC Genomics* **8**: 73.

Helman LJ, Meltzer P. (2003). Mechanisms of sarcoma development. *Nat Rev Cancer* **3**: 685–694.

Henderson SR, Guiliano D, Presneau N, McLean S, Frow R, Vujovic S et al. (2005). A molecular map of mesenchymal tumors. *Genome Biol* **6**: R76.

Hernando E, Charytonowicz E, Dudas ME, Menendez S, Matushansky I, Mills J et al. (2007). The AKT-mTOR pathway plays a critical role in the development of leiomyosarcomas. *Nat Med* **13**: 748–753.

Jun HJ, Lee J, Lim DH, Park JO, Ahn G, Seo SW et al. (2009). Expression of MET in alveolar soft part sarcoma. *Med Oncol* (e-pub ahead of print 27 May 2009; PMID: 19472090).

Kapp AV, Tibshirani R. (2007). Are clusters found in one dataset present in another dataset? *Biostatistics* **8**: 9–31.

Kattan MW, Leung DH, Brennan MF. (2002). Postoperative nomogram for 12-year sarcoma-specific death. *J Clin Oncol* **20**: 791–796.

Kochetkova M, McKenzie OL, Bais AJ, Martin JM, Secker GA, Seshadri R et al. (2002). CBFA2T3 (MTG16) is a putative breast tumor suppressor gene from the breast cancer loss of heterozygosity region at 16q24.3. *Cancer Res* **62**: 4599–4604.

Lakhani SR, Ashworth A. (2001). Microarray and histopathological analysis of tumours: the future and the past? *Nat Rev Cancer* **1**: 151–157.

Landis-Piwowar KR, Milacic V, Dou QP. (2008). Relationship between the methylation status of dietary flavonoids and their growth-inhibitory and apoptosis-inducing activities in human cancer cells. *J Cell Biochem* **105**: 514–523.

Lapointe J, Li C, Higgins JP, van de Rijn M, Bair E, Montgomery K et al. (2004). Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc Natl Acad Sci USA* **101**: 811–816.

Larramendy ML, Gentile M, Soloneski S, Knuutila S, Bohling T. (2008). Does comparative genomic hybridization reveal distinct differences in DNA copy number sequence patterns between leiomyosarcoma and malignant fibrous histiocytoma? *Cancer Genet Cytogenet* **187**: 1–11.

Lee CH, Espinosa I, Vrijaldenhoven S, Subramanian S, Montgomery KD, Zhu S et al. (2008). Prognostic significance of macrophage infiltration in leiomyosarcomas. *Clin Cancer Res* **14**: 1423–1430.

Lee YF, John M, Falconer A, Edwards S, Clark J, Flohr P et al. (2004). A gene expression signature associated with metastatic outcome in human leiomyosarcomas. *Cancer Res* **64**: 7201–7204.

Lensch MW, Tischkowitz M, Christianson TA, Reifsteck CA, Speckhart SA, Jakobs PM et al. (2003). Acquired FANCA dysfunction and cytogenetic instability in adult acute myelogenous leukemia. *Blood* **102**: 7–16.

Li L, Ren CH, Tahir SA, Ren C, Thompson TC. (2003). Caveolin-1 maintains activated Akt in prostate cancer cells through scaffolding domain binding site interactions with and inhibition of serine/threonine protein phosphatases PP1 and PP2A. *Mol Cell Biol* **23**: 9389–9404.

Meza-Zepeda LA, Kresse SH, Barragan-Polania AH, Bjerkehagen B, Ohnstad HO, Namlos HM et al. (2006). Array comparative genomic hybridization reveals distinct DNA copy number differences between gastrointestinal stromal tumors and leiomyosarcomas. *Cancer Res* **66**: 8984–8993.

Nakayama R, Nemoto T, Takahashi H, Ohta T, Kawai A, Seki K et al. (2007). Gene expression analysis of soft tissue sarcomas: characterization and reclassification of malignant fibrous histiocytoma. *Mod Pathol* **20**: 749–759.

Nielsen TO, West RB, Linn SC, Alter O, Knowling MA, O'Connell JX et al. (2002). Molecular characterisation of soft tissue tumours: a gene expression study. *Lancet* **359**: 1301–1307.

Ottone T, Hasan SK, Montefusco E, Curzi P, Mays AN, Chessa L et al. (2009). Identification of a potential 'hotspot' DNA region in the RUNX1 gene targeted by mitoxantrone in therapy-related acute myeloid leukemia with t(16;21) translocation. *Genes Chromosomes Cancer* **48**: 213–221.

Perot G, Derre J, Coindre JM, Tirode F, Lucchesi C, Mariani O et al. (2009). Strong smooth muscle differentiation is dependent on myocardin gene amplification in most human retroperitoneal leiomyosarcomas. *Cancer Res* **69**: 2269–2278.

Pervaiz N, Colterjohn N, Farrokhyar F, Tozer R, Figueredo A, Ghert M. (2008). A systematic meta-analysis of randomized controlled trials of adjuvant chemotherapy for localized resectable soft-tissue sarcoma. *Cancer* **113**: 573–581.

Potti A, Dressman HK, Bild A, Riedel RF, Chan G, Sayer R *et al*. (2006). Genomic signatures to guide the use of chemotherapeutics. *Nat Med* **12**: 1294–1300.

Quade BJ, Wang TY, Sornberger K, Dal Cin P, Mutter GL, Morton CC. (2004). Molecular pathogenesis of uterine smooth muscle tumors from transcriptional profiling. *Genes Chromosomes Cancer* **40**: 97–108.

Ramaswamy S, Ross KN, Lander ES, Golub TR. (2003). A molecular signature of metastasis in primary solid tumors. *Nat Genet* **33**: 49–54.

Ren B, Yu YP, Jing L, Liu L, Michalopoulos GK, Luo JH *et al*. (2003). Gene expression analysis of human soft tissue leiomyosarcomas. *Hum Pathol* **34**: 549–558.

Rong S, Jeffers M, Resau JH, Tsarfaty I, Oskarsson M, Vande Woude GF. (1993). Met expression and sarcoma tumorigenicity. *Cancer Res* **53**: 5355–5360.

Sarcoma Meta-analysis Collaboration (2000). Adjuvant chemotherapy for localised resectable soft tissue sarcoma in adults. *Cochrane Database Syst Rev*. Art. No. CD001419.

Seale P, Bjork B, Yang W, Kajimura S, Chin S, Kuang S *et al*. (2008). PRDM16 controls a brown fat/skeletal muscle switch. *Nature* **454**: 961–967.

Segal NH, Pavlidis P, Antonescu CR, Maki RG, Noble WS, DeSantis D *et al*. (2003). Classification and subtype prediction of adult soft tissue sarcoma by functional genomics. *Am J Pathol* **163**: 691–700.

Shmulevich I, Hunt K, El-Naggar A, Taylor E, Ramdas L, Laborde P *et al*. (2002). Tumor specific gene expression profiles in human leiomyosarcoma: an evaluation of intratumor heterogeneity. *Cancer* **94**: 2069–2075.

Sidransky D. (2002). Emerging molecular markers of cancer. *Nat Rev Cancer* **2**: 210–219.

Skubitz KM, Skubitz AP. (2003). Differential gene expression in leiomyosarcoma. *Cancer* **98**: 1029–1038.

Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H *et al*. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA* **98**: 10869–10874.

Sotiriou C, Piccart MJ. (2007). Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? *Nat Rev Cancer* **7**: 545–553.

Tibshirani R, Hastie T, Narasimhan B, Chu G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci USA* **99**: 6567–6572.

Tibshirani R, Wang P. (2008). Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics* **9**: 18–29.

Toro JR, Travis LB, Wu HJ, Zhu K, Fletcher CD, Devesa SS. (2006). Incidence patterns of soft tissue sarcomas, regardless of primary site, in the surveillance, epidemiology and end results program, 1978–2001: An analysis of 26 758 cases. *Int J Cancer* **119**: 2922–2930.

Tusher VG, Tibshirani R, Chu G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* **98**: 5116–5121.

Wang W. (2007). Emergence of a DNA-damage response network consisting of Fanconi anaemia and BRCA proteins. *Nat Rev Genet* **8**: 735–748.

Weigelt B, Hu Z, He X, Livasy C, Carey LA, Ewend MG *et al*. (2005). Molecular portraits and 70-gene prognosis signature are preserved throughout the metastatic process of breast cancer. *Cancer Res* **65**: 9155–9158.

Weigelt B, van't Veer LJ. (2004). Hard-wired genotype in metastatic breast cancer. *Cell Cycle* **3**: 756–757.

Weiss SW, Goldblum JR. (2008). *Enzinger and Weiss's Soft Tissue Tumors*, 5th edn. Mosby/Elsevier: Philadelphia.

Zou H, Hastie T, Tibshirani R. (2006). Sparse principal component analysis. *J Comput Graphical Stat* **15**: 265–286.

Supplementary Information accompanies the paper on the Oncogene website (http://www.nature.com/onc)