SUPPLEMENTAL DATA

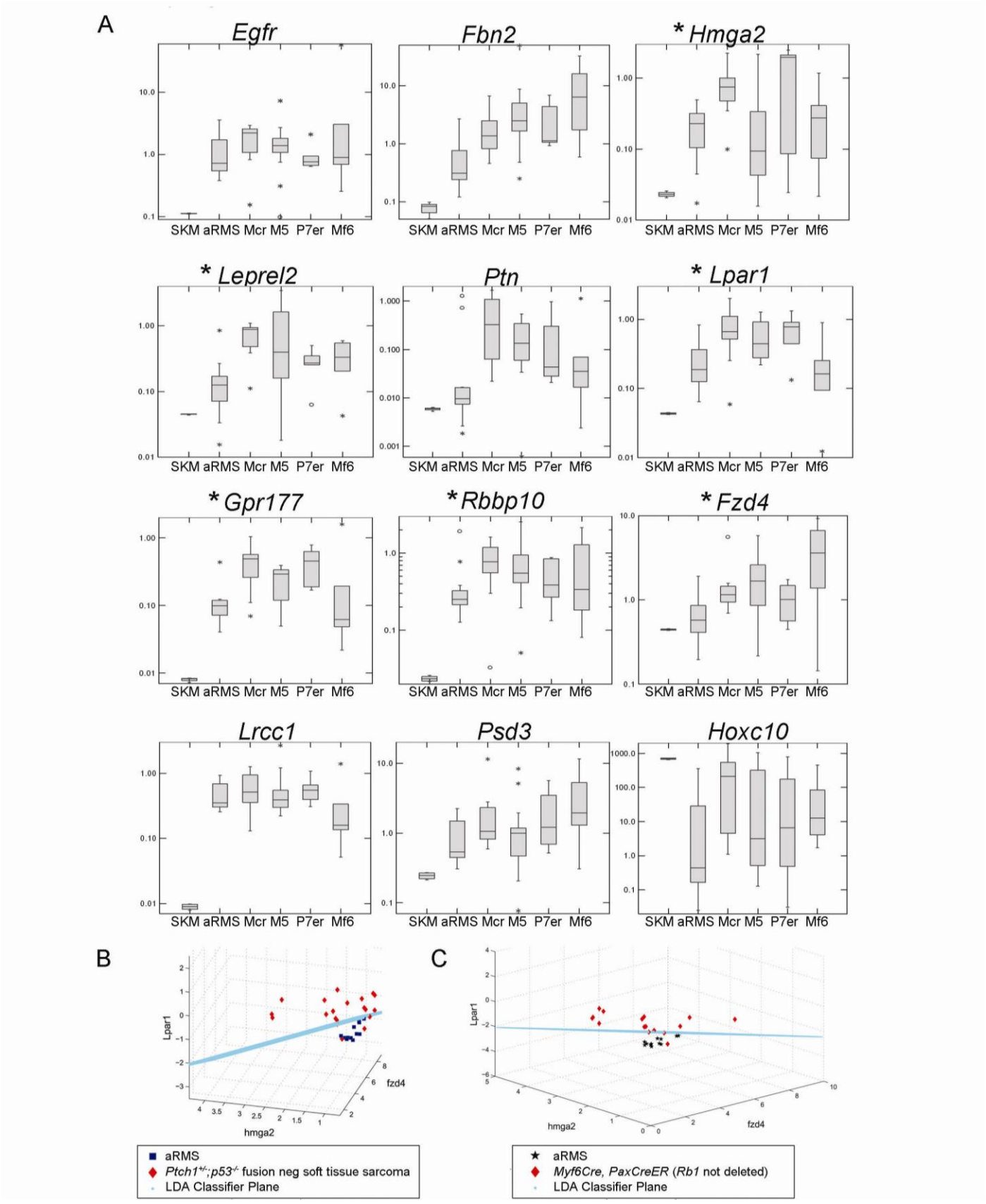


Figure S1 (continued)

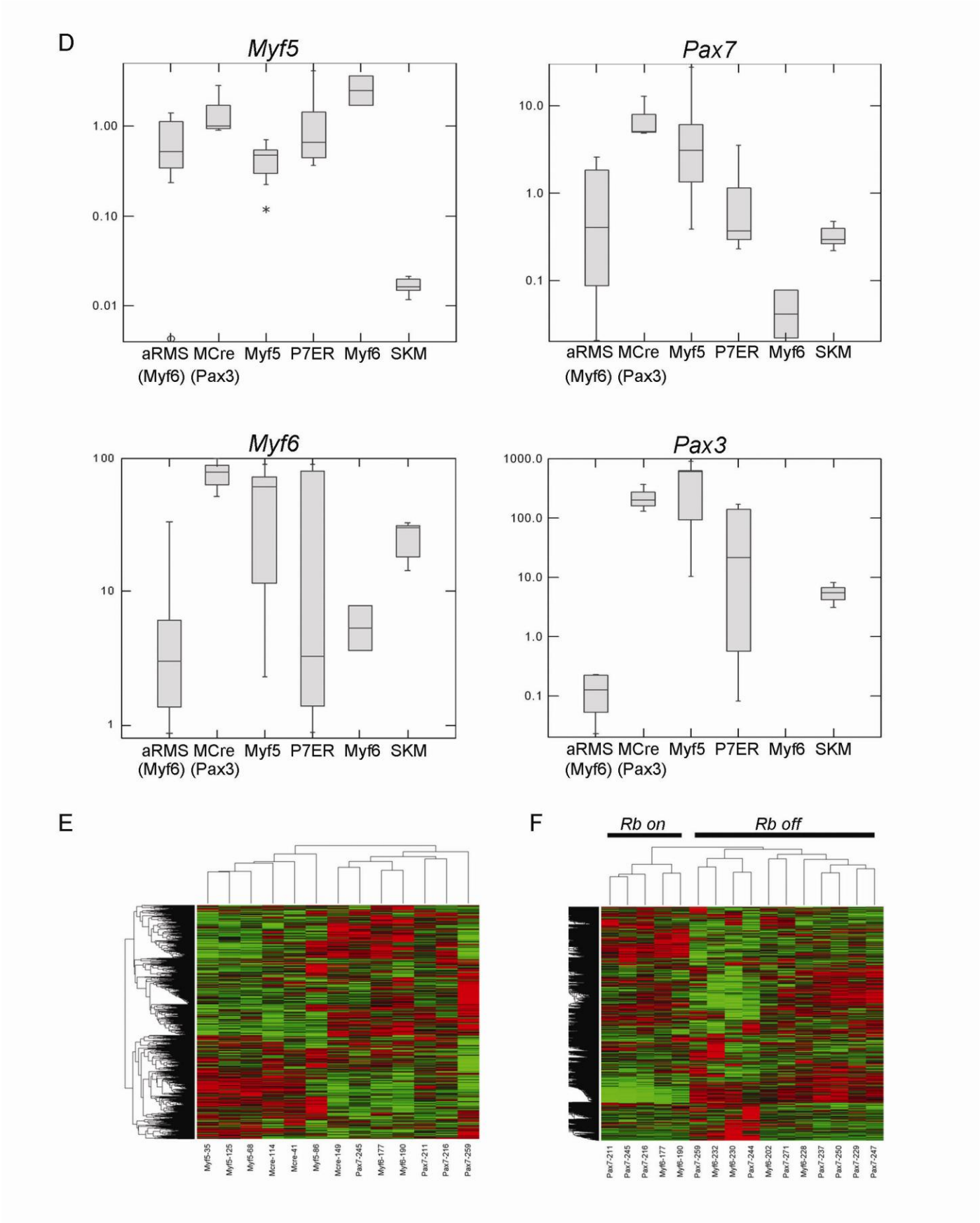


Figure S1, related to Figure 1.

(A) Expression of genes suggested to be specifically over-expressed for human eRMS in murine *Ptch1*$^{+/-}$;*p53*$^{-/-}$ soft tissue tumors versus aRMS or skeletal muscle. SKM, skeletal muscle. aRMS, alveolar rhabdomyosarcoma. Mcr, *MCre*. M5, *Myf5Cre*. P7er, *Pax7CreER*, Mf6, *Myf6Cre*. Statistically-different groups are signified by an asterisk (*).

(B) The human eRMS specific markers *Lpar1*, *Hmga2* and *Fzd4* distinguish murine *Ptch1*$^{+/-}$;*p53*$^{-/-}$, fusion negative soft tissue sarcomas from murine aRMS.

(C) The same eRMS specific markers *Lpar1*, *Hmga2* and *Fzd4* distinguish murine tumors of the *Myf6Cre* or *Pax7CreER* lineages from aRMS when *Rb1* is not deleted; however, this ability to distinguish tumor samples is lost when *Rb1* is homozygously deleted (data not shown in this figure; see Table S2).

(D) Quantitative RT-PCR for *Myf5*, *Pax7*, *Myf6* and *Pax3* genes in *p53*$^{-/-}$ and *Ptch1*$^{+/-}$;*p53*$^{-/-}$ soft tissue tumors (osteosarcomas were excluded). Markers of the origin lineage were not necessarily up-regulated in the tumors that developed in each model, related to Figures 1-3. Expression of each gene was normalized to *Gapdh* expression. The aRMS model shown for comparison is derived from the *Myf6Cre* lineage. P7ER, *Pax7CreER*. Myf5, *Myf5Cre*. Myf6, *Myf6Cre*. SKM, skeletal muscle.

(E) Unsupervised hierarchical clustering of *Ptch1-p53* soft tissue demonstrated no clear relationship between cell of origin and histological subtypes(average intensity>64, 12246 probes).

(F) Unsupervised hierarchical clustering of fusion-negative *Ptch1-p53* or *p53* soft tissue tumors from the *Pax7CreER* or *Myf6Cre* lineages identified a correlation between *Rb1* mutation and global gene expression (average intensity>64, 12309 probes). Taken together, these results suggest that genetic events are more important for the determination of histological subtypes of RMS (phenotype) than the cell of origin.

In panels (A) and (D), error bars represent SEM.

Table **S1**. Classifier Analysis of eRMS vs. non-eRMS. Results for classifiers for tumor phenotypes based on original cell of origin. *This supplemental table relates to Figure 1-3.*

Gene Expression by Lineage	Classifier	resub	loo	cv	boot	.63boot	average
Myf6Cre vs. Myf5Cre, MCre, Pax7CreER							
Myf6	LDA	0.25926	0.25926	0.34524	0.33119	0.36947	0.31288
Myf6	KNN	0.14815	0.2963	0.29286	0.31899	0.25799	0.26286
Myf5Cre vs. Myf6Cre, MCre, Pax7CreER							
Myf5	LDA	0.33333	0	0.4081	0.42411	0.41947	0.317
Myf5	KNN	0.22222	0.2963	0.30952	0.32271	0.27027	0.28421
***Mcre* (*Pax3Cre*) vs. Myf6Cre, Myf5Cre, Pax7CreER**							
Pax3	LDA	0.37037	0.40741	0.40762	0.37634	0.39698	0.39174
Pax3	KNN	0.11111	0.11111	0.11619	0.21192	0.13241	0.13655
Pax7CreER vs. Myf6Cre, Myf5Cre, Mcre							
Pax7	LDA	0.2963	0.2963	0.33333	0.33533	0.29563	0.31138
Pax7	KNN	0.22222	0.48148	0.49238	0.38169	0.37164	0.38988

Table **S2**. Classifier Analysis of eRMS vs. non-eRMS taking into account *Rb1* deletion. Best markers and their errors differentiating various classes. *This supplemental table relates to Figure 1.*

Comparisons	Gene 1	Gene 2	Gene 3	Classifier	Average Error
Deleted *Rb1* vs. *Rb1* in M6, P7ER fnSTS	*Leprel2*	*Fzd4*	*Ptn*	LDA	0.19
Deleted *Rb1* vs. *Rb1* in M6, P7ER fnSTS	*Leprel2*	*Psd3*	*Ptn*	KNN	0.19
aRMS vs. deleted *Rb1* in M6, P7ER fnSTS	*Gpr177*	*Hmga2*	*Egfr*	LDA	0.21
aRMS vs. deleted *Rb1* in M6, P7ER fnSTS	*Gpr177*	*Egfr*	*Hmga2*	KNN	0.195
aRMS vs. non-deleted *Rb1* in M6, P7ER fnSTS	*Hmga2*	*Lpar1*	*Fzd4*	LDA	**0.08**
aRMS vs. non-deleted *Rb1* in M6, P7ER fnSTS	*Gpr177*	*Hmga2*	*Lrcc1*	KNN	0.13

M6, Myf6Cre. P7ER, Pax7CreER. fnSTS, fusion negative soft tissue sarcoma (*e.g.*, eRMS).

Table **S3**. Supervised Clustering of Microarray Results. *This supplemental table relates to Figure 1.* Please find this table in the corresponding, accompanying Excel file.

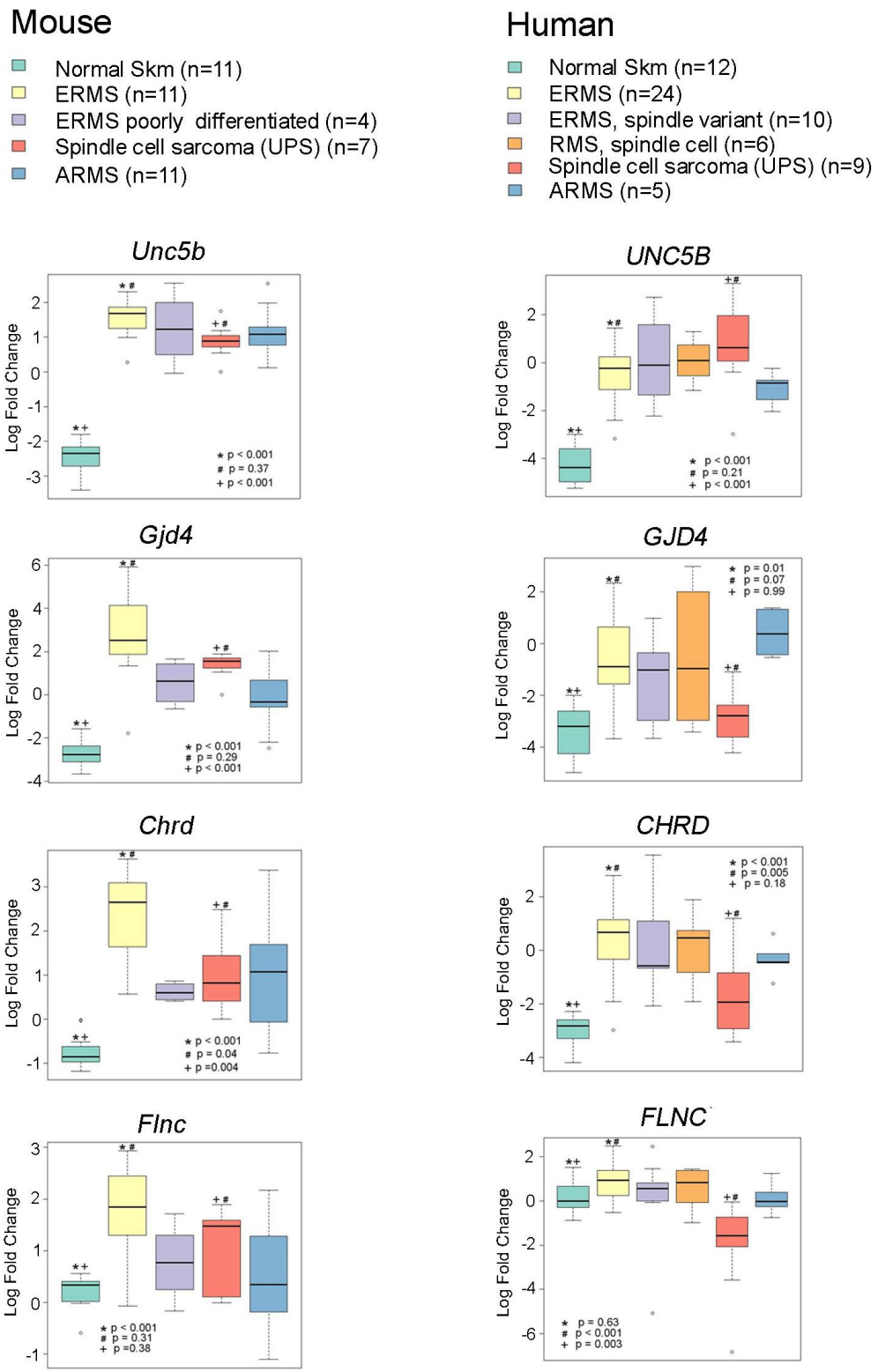


Figure S2 Global gene expression relationships based on biological features, related to Figure 4. Markers found to be differentially expressed by eRMS or undifferentiated spindle cell sarcomas (USCS/UPS) for mouse and human. Error bars represent SEM.

Table **S4**. SVM: Table of Classification Results. *This supplemental table relates to Figure 7 to validate five genes that as a profile differentiate SKM, eRMS, USCS/UPS and aRMS in humans.*

	Observed by Pathologist							
Predicted by Model	Normal SkM	ERMS	ERMS,spindle cell variant	RMS,spindle cell	Spindle cell sarcoma	ARMS	Total	Accuracy
Normal SKM	12	0	1	1	0	0	14	0.71 (0.056)
ERMS	0	23	7	4	1	0	35	
ERMS,spindle cell variant	0	1	0	0	1	0	2	
RMS,spindle cell	0	0	0	0	0	0	0	
Spindle cell sarcoma	0	0	1	1	7	0	9	
ARMS	0	0	1	0	0	5	6	
Total	12	24	10	6	9	5	66	

Table **S5**. SVM: Table of Classification Statistics. *This supplemental table relates to Figure 7 to validate five genes that as a profile differentiate SKM, eRMS, USCS/UPS and aRMS in humans.*

Contrast	Statistic	Estimate	95% CI	Joint CI
Normal SKM *vs.* others	True Positive Fraction (TPF or Sensitivity)	1 (0)	(0.81, 1)	(0.78, 1) x (0.01, 0.12)[3]
	False Positive Fraction (FPF)	0.04 (0.023)	(0.01, 0.11)	
	Specificity (1 - FPF)	0.96 (0.023)	(0.89, 0.99)	(0.78, 1) x (0.89, 0.99)[4]
	Positive Predictive Value (PPV)	0.86 (0.043)	(0.62, 0.97)	(0.62, 0.97) x (0.96, 1)[5]
	Negative Predictive Value (NPV)	1 (0)	(0.95, 1)	
ERMS *vs.* others	True Positive Fraction (TPF or Sensitivity)	0.96 (0.025)	(0.82, 1)	(0.8, 1) x (0.09, 0.3)[3]
	False Positive Fraction (FPF)	0.29 (0.056)	(0.17, 0.43)	
	Specificity (1 - FPF)	0.71 (0.056)	(0.57, 0.83)	(0.8, 1) x (0.57, 0.83)[4]
	Positive Predictive Value (PPV)	0.66 (0.058)	(0.49, 0.8)	(0.49, 0.8) x (0.92, 1)[5]
	Negative Predictive Value (NPV)	0.97 (0.022)	(0.86, 1)	
ERMS,spindle cell variant *vs.* others	True Positive Fraction (TPF or Sensitivity)	0 (0)	(0, 0.22)	(0, 0.26) x (0.01, 0.11)[3]
	False Positive Fraction (FPF)	0.04 (0.023)	(0.01, 0.11)	
	Specificity (1 - FPF)	0.96 (0.023)	(0.89, 0.99)	(0, 0.26) x (0.89, 0.99)[4]
	Positive Predictive Value (PPV)	0 (0)	(0, 0.67)	(0, 0.67) x (0.75, 0.92)[5]
	Negative Predictive Value (NPV)	0.84 (0.045)	(0.74, 0.92)	
RMS,spindle cell *vs.* others	True Positive Fraction (TPF or Sensitivity)	0 (0)	(0, 0.33)	(0, 0.39) x (0, 0.05)[3]
	False Positive Fraction (FPF)	0 (0)	(0, 0.04)	
	Specificity (1 - FPF)	1 (0)	(0.96, 1)	(0, 0.39) x (0.96, 1)[4]
	Positive Predictive Value (PPV)		(0, 1)	(0, 1) x (0.82, 0.96)[5]
	Negative Predictive Value (NPV)	0.91 (0.035)	(0.82, 0.96)	
Spindle cell sarcoma *vs.* others	True Positive Fraction (TPF or Sensitivity)	0.78 (0.051)	(0.46, 0.95)	(0.41, 0.96) x (0.01, 0.11)[3]
	False Positive Fraction (FPF)	0.04 (0.023)	(0.01, 0.11)	
	Specificity (1 - FPF)	0.96 (0.023)	(0.89, 0.99)	(0.41, 0.96) x (0.89, 0.99)[4]
	Positive Predictive Value (PPV)	0.78 (0.051)	(0.46, 0.95)	(0.46, 0.95) x (0.9, 0.99)[5]
	Negative Predictive Value (NPV)	0.96 (0.023)	(0.89, 0.99)	
ARMS vs. others	True Positive Fraction (TPF or Sensitivity)	1 (0)	(0.62, 1)	(0.55, 1) x (0, 0.08)[3]
	False Positive Fraction (FPF)	0.02 (0.016)	(0, 0.07)	
	Specificity (1 - FPF)	0.98 (0.016)	(0.93, 1)	(0.55, 1) x (0.93, 1)[4]
	Positive Predictive Value (PPV)	0.83 (0.046)	(0.44, 0.98)	(0.44, 0.98) x (0.96, 1)[5]
	Negative Predictive Value (NPV)	1 (0)	(0.96, 1)	

[1] Support Vector Machine (SVM) using a Gaussian Radial Basis Function (RBF) kernel function (Scholkopf et al., 1997).

[2] F-score and Supported Sequential Forward Search method (F_SSFS) for gene variable reduction (Lee, 2009).

[3] For (TPF, FPF); [4] For (Sensitivity, Specificity); [5] For (PPV, NPV).

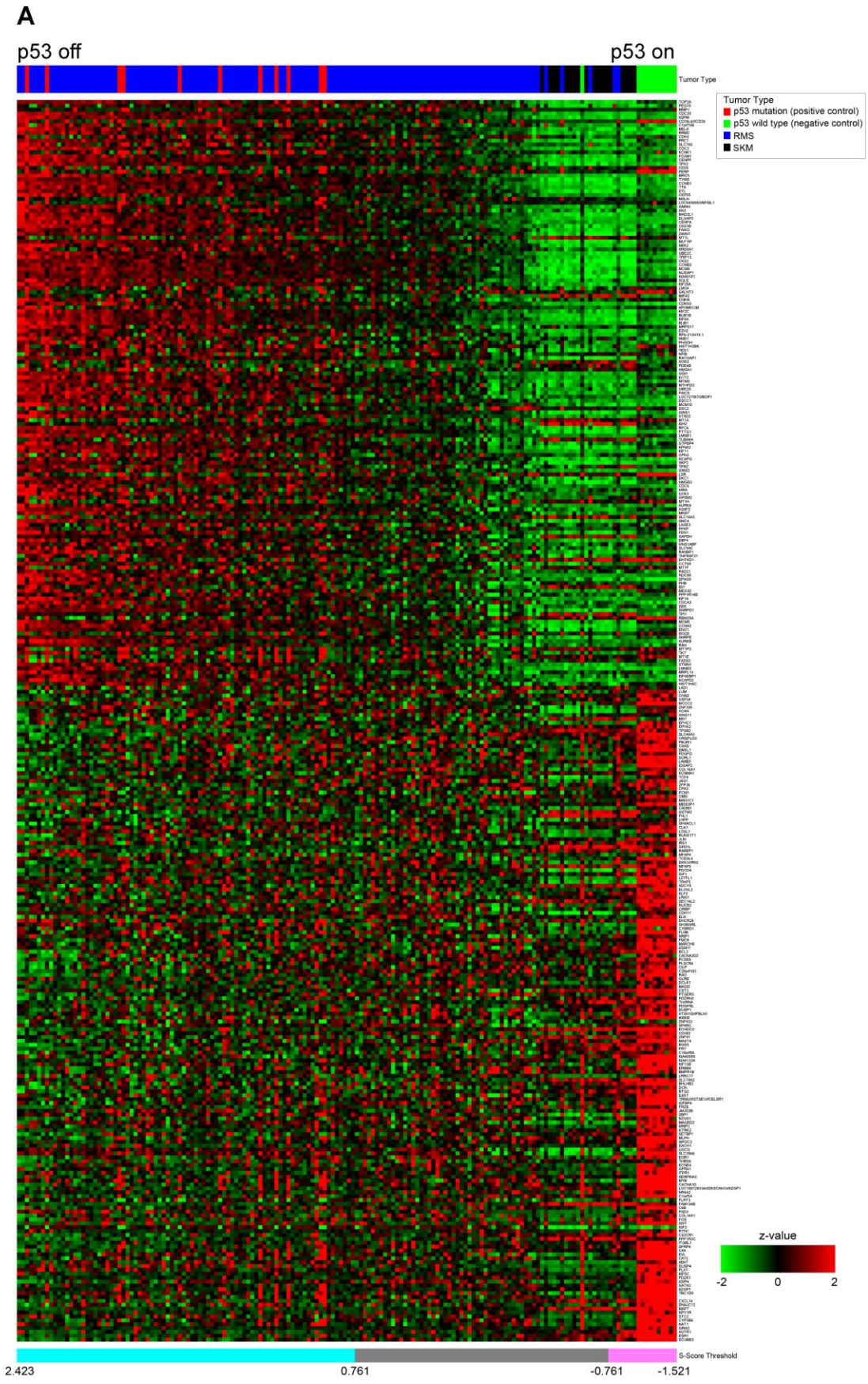

A
p53 off
p53 on
Tumor Type
Tumor Type
p53 mutation (positive control)
p53 wild type (negative control)
RMS
SKM
z-value
-2
0
2
S-Score Threshold
2.423
0.761
-0.761
-1.521

Figure S3 (continued)

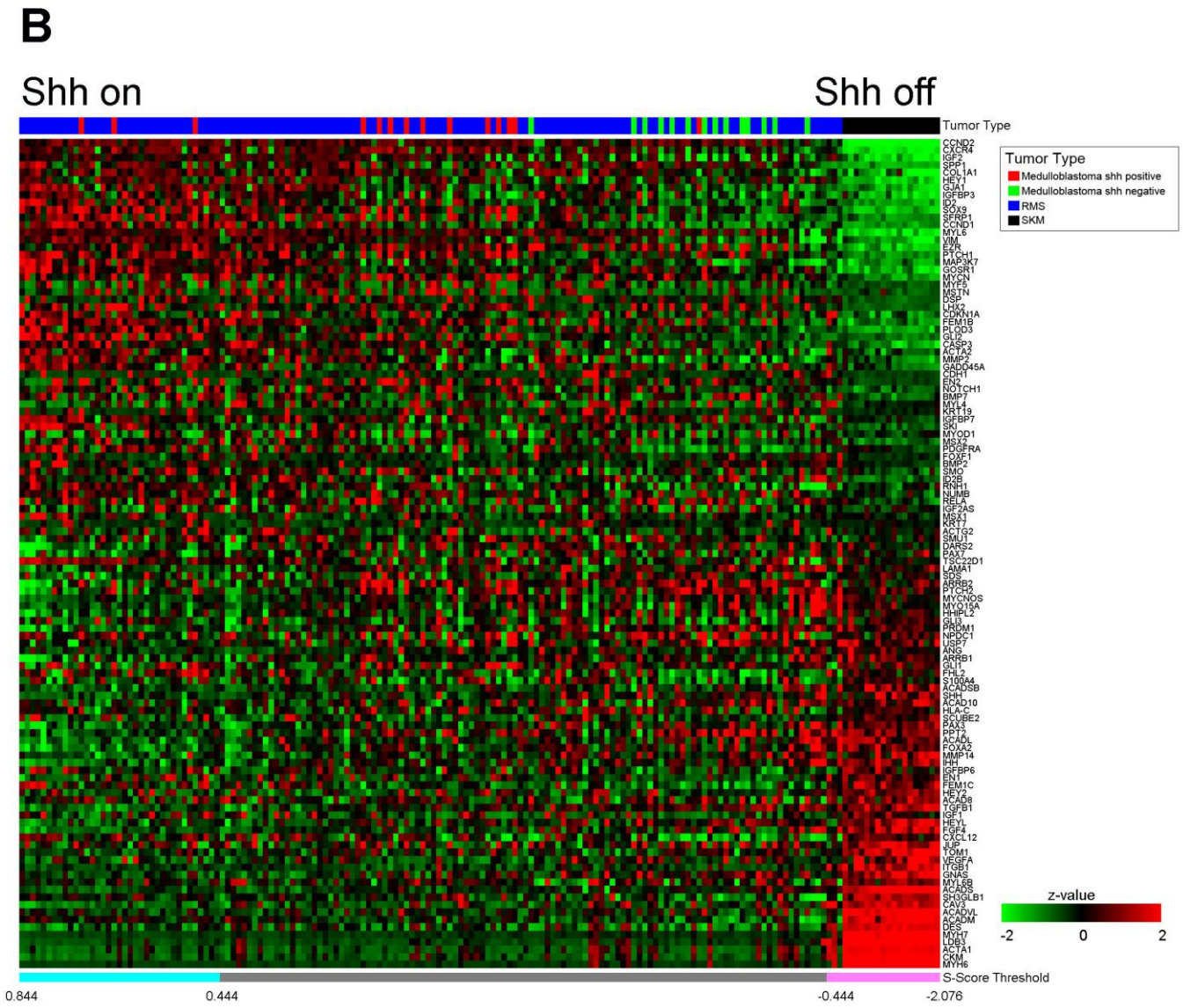


Figure S3 (continued)

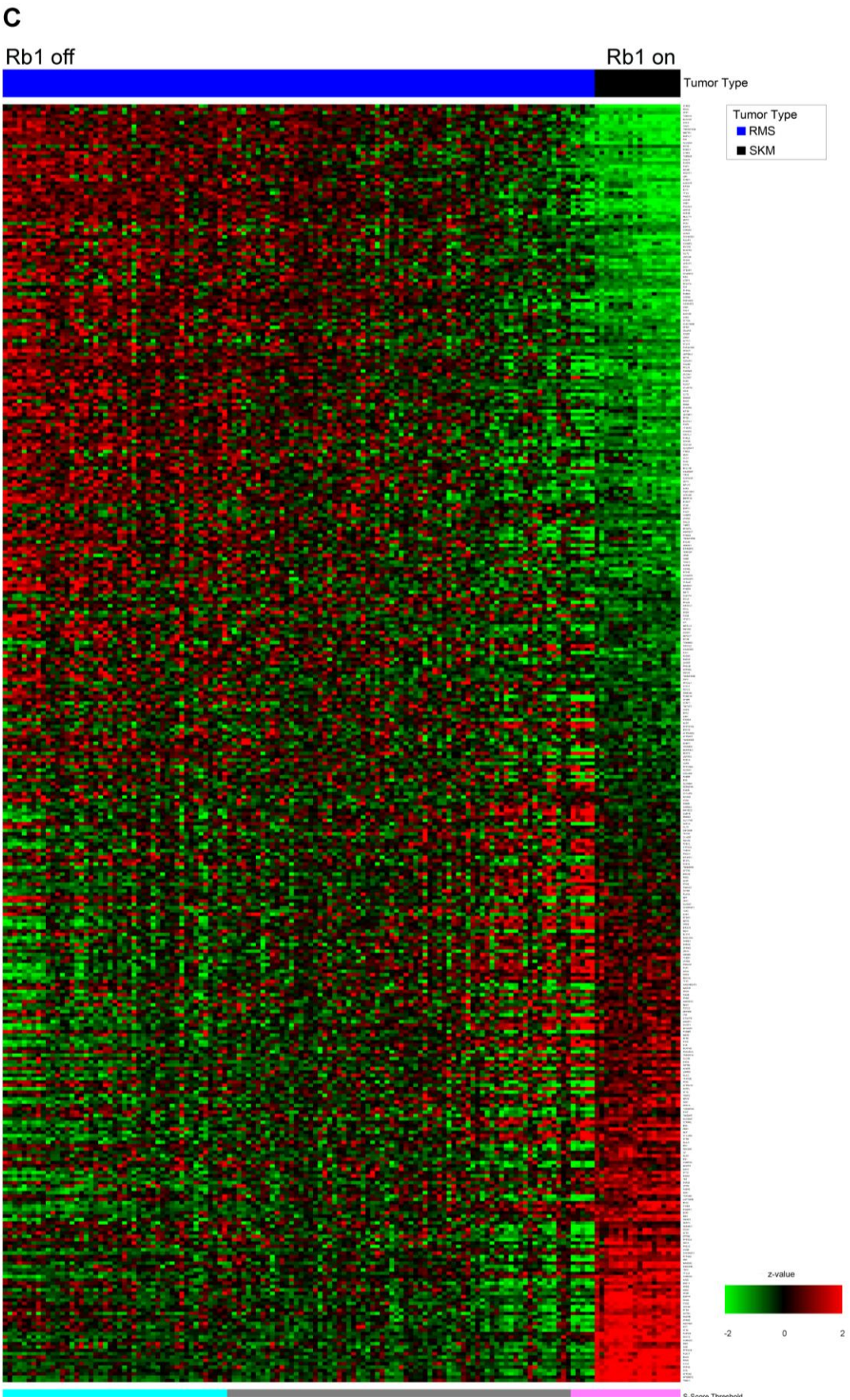


Figure S3 (continued)

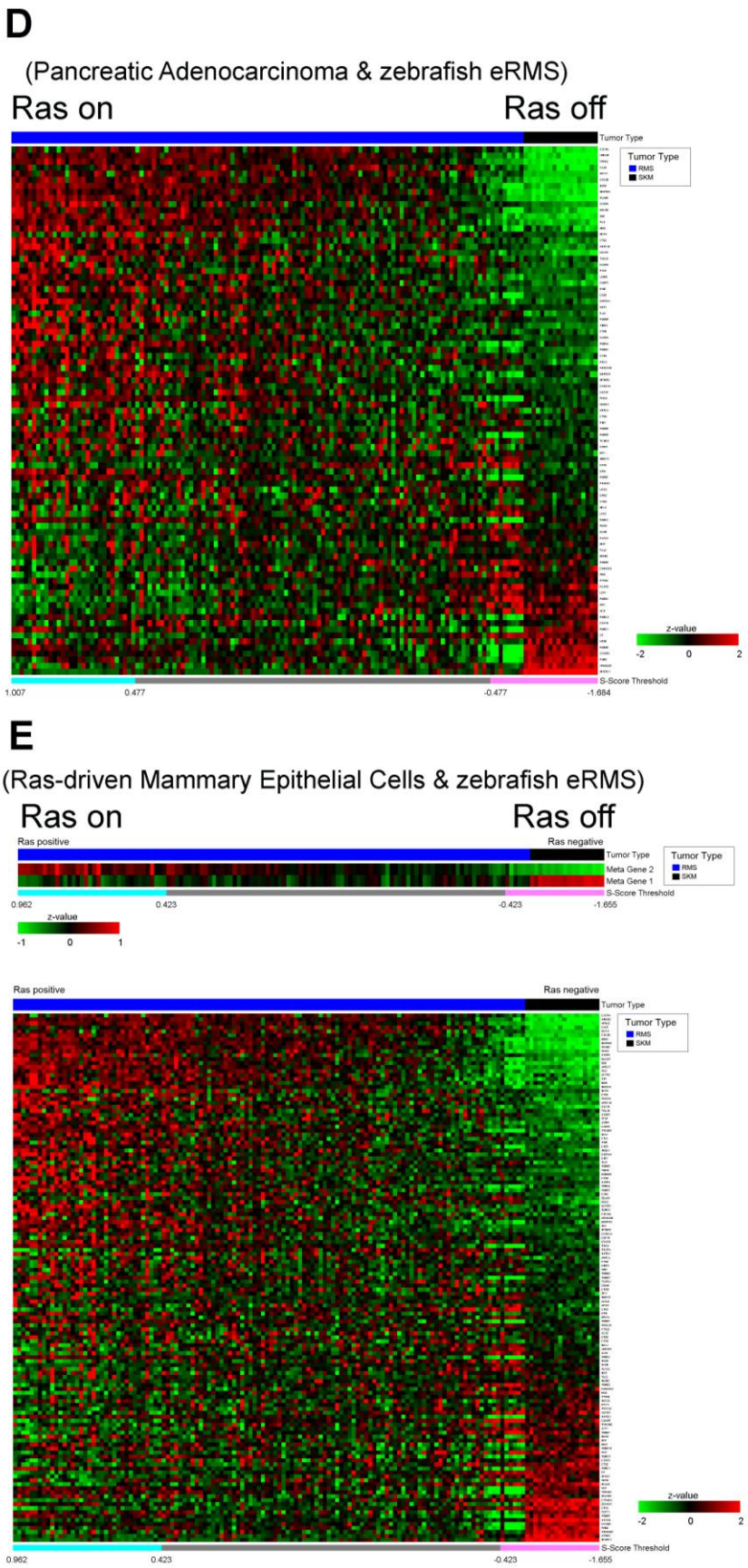


Figure S3 Heatmaps of human fusion-negative rhabdomyosarcomas for gene signatures representing the p53, Shh, Rb1 or Ras pathways, related to Figure 8.

(A) The p53 pathway heatmap was generated based on 320 significant genes (Benjamini and Hochberg adjusted p-value <0.05 & fold change > 2) with 183 genes were in metagene 1, whereas the other 136 genes were in metagene 2. In total, 65 of 111 fusion-negative RMS primary tumor samples (59%) exhibited a gene expression signature consistent with the 'p53 off' state for which the S-score was greater than 0.761 (see Supplemental Experimental Procedures). All human breast cancer control samples with known p53 mutations also exhibited S-scores greater than 0.761.

(B) For the Shh signaling pathway heatmap, 111 genes were employed, 56 of which were in metagene 1, the other 55 genes were in metagene 2. Overall, 32 of 111 (29%) of tumors exhibited a gene expression signature consistent with a 'Shh on' overdrive state for which S-score was greater than 0.444. The p-value for comparison of the S-score for Shh+ medulloblastoma controls and for Shh- medulloblastoma controls was 2.62×10^{-5} as calculated by the Wilcoxon rank sum test.

(C) For the Rb1 pathway, 381 genes were used to construct the heatmap, with 157 in metagene 1, and 224 genes in meta gene 2. For the Rb1 pathway, 42 of 111 (38%) of samples demonstrated an 'Rb1 state' with an S-score greater than 0.345. All heatmaps for every single gene in all three pathways are provided in Supplemental Figure S5.

(D) To evaluate the Ras pathway, 87 genes common to zebrafish eRMS and human Ras-driven pancreatic cancer were used to construct the heatmap. A corresponding metagene analysis is given in Figure 8D.

(E) As a secondary way to evaluate the Ras pathway, 112 genes common to zebrafish eRMS and human Ras-driven mammary epithelial cells were used to construct the heatmap, with 45 genes in metagene 1, and 88 genes in metagene 2. (see also Figure 8 legend and Results).

Table **S6**. p53 Signature Genes and Results. *This supplemental table gives the detailed descriptions of gene lists and samples used to generate the metagene profiles in Figure 8A.* Please find this table in the corresponding, accompanying Excel file.

Table **S7**. SHH Signature Genes and Results. *This supplemental table gives the detailed descriptions of gene lists and samples used to generate the metagene profiles in Figure 8B.* Please find this table in the corresponding, accompanying Excel file.

Table **S8**. Rb1 Signature Genes and Results. *This supplemental table gives the detailed descriptions of gene lists and samples used to generate the metagene profiles in Figure 8C.* Please find this table in the corresponding, accompanying Excel file.

Table **S9**. Ras Signature Genes and Results. *This supplemental table gives the detailed descriptions of gene lists and samples used to generate the metagene profiles in Figure 8D.* Please find this table in the corresponding, accompanying Excel file.

Table **S10**. Human Sample Subgroups for p53, SHH, Rb1 and Ras Signatures (99% CI). This supplemental table gives the detailed descriptions of human tumor samples used to generate the Venn diagram in Figure 8E. Please find this table in the corresponding, accompanying Excel file.

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Mice

The conditional alleles for *Pax3:Fkhr* knock-in(MMHCC Strain Code 01XBM B6), *Patched1* deletion(MMHCC Strain Submission ID #232), *p53* deletion, and *Rb1* deletion have been previously described(Keller et al., 2004a; Keller and Capecchi, 2005; Marino et al., 2003; Marino et al., 2000; Nishijo et al., 2009; Taniguchi et al., 2009). Myogenic Cre lines, including *Myf6Cre*(MMHCC Strain Code 01XBL B6), *Myf5Cre*, *Pax7CreER*(MMHCC Strain Submission ID #231), and *MCre*, were also described previously(Brown et al., 2005; Keller et al., 2004a; Keller and Capecchi, 2005; Keller et al., 2004b). For the *Pax7CreER* line, tamoxifen was administered intraperitoneally for 5 consecutive days at 4 weeks of age. Kaplan-Meier survival analysis of the mice was performed with the end-point being the development of rhabdomyosarcoma (first visible sign of a tumor). The log-rank test was utilized to determine the statistical significance ($p<0.05$). Both analyses were performed with Systat12 software (Systat, Chicago, IL).

Histology and immunohistochemical staining

Fixed tissues were paraffin-embedded and sectioned at 3.5µm thickness. Paraffin sections were stained with hematoxylin and eosin (H&E) as previously described, or by Gomori Trichrome. For MyoD and Myogenin immunohistochemistry, staining was performed using the M.O.M. Immunodetection Kit Staining Procedure (Vector Laboratories, Burlingame, CA) following the manufacturer's instructions using antigen unmasking. The Myogenin monoclonal primary antibody (F5D supernatants; Developmental Hybridoma Studies Bank, Iowa City, IA) was used at a concentration of 1:50. The Desmin monoclonal primary antibody (Sigma Aldrich, St. Louis, MO, USA) was used at a concentration of 1:200.

Diagnostic criteria of eRMS ranged from morphological features of spindle cells without obvious rhabdomyoblastic differentiation to neoplasms composed of an admixture of primitive-appearing or spindle-shaped cells with variable numbers of epithelioid or spindle cell shaped rhabdomyoblasts. Cross-striations were another supportive diagnostic feature. Expression of Desmin and Myogenin or MyoD was an additional criterion. UPS morphological criteria allowed a wider range of spindle cell morphologies, yet lacking rhabdomyoblasts or the immunohistochemical expression of Desmin, myogenin and MyoD.

Cell culture and immunocytochemical staining

For primary culture of mouse tumors, tumor tissues were digested with 1% collagenase IV (Sigma Aldrich) overnight, rinsed with PBS, and then plated on 10cm dishes. Cells were cultured in Dulbecco's modified eagle media (DMEM, Sigma Aldrich) supplemented with 10% FBS. For induction of myogenic differentiation, cells were cultured under 2% horse serum for 5 days. For immunocytochemical staining, antibodies against Ki67 and MHC were from Santa Cruz Biotechnology (Santa Cruz, CA) and DSHB, respectively. Nuclei were counter-stained with DAPI and observed under a Leica TCS-LSI confocal microscope (Leica, Bannockburn, IL).

In vitro growth assays

The CellTiter-Glo Luminescent Cell Viability Assay (Promega, Fitchburg, WI) was utilized according to the manufacturer's specifications. Mouse rhabdomyosarcoma primary cell cultures were plated at 5×10^3 cells per well in 96-well plates. After 24 hours, the cells were washed thrice then incubated with DMEM and 1% or 10% FBS for 1 – 5 days. The effects on cell viability were assessed using the CellTiter-Glo Luminescent Cell Viability Assay and the SpectraMax M5 luminometer machine (Molecular Devices, Sunnyvale, CA). Three replicates were performed for each data point.

Real-Time RT-PCR

Quantitative reverse transcription-PCR (qRT-PCR) analyses for Figure S1 were performed by SYBR Green assay (PE Applied Biosystems, Foster City, CA). Primers for these mouse genes (in 5' to 3' orientation) were: *Egfr* (cagatggatgtcaaccctgaag and tggagagtgtgtctttaaattcacc); *Fbn2* (tcaattcagcagtgtagcgt and caagcacagcggttaggg); *Fzd4* (gcagttcttcctttgttcggt and ccaaattctctcaggactggtt); *Gpr177* (gcatcttcatcattatggtgtggt and catggaaatcccaagggcaaa); *Hmga2* (aaatggccacaacaagtcgt and tctcccttcaaaagatccaactg); *Hoxc10* (cggataacgaagctaaagagga and gcgtctggtgtttagtataggg); *Leprel2* (gaccacgagaggacatcca and ccgggtccttgaagctagt); *Lpar1* (tcatggtggttctctacgct and agcagacaataaaggcaccaa); *Lrrc1* (gaaatcagctgtctgaattacctc and ccctcaggaattgtttctagca); *Psd3* (ttcaagagatcggacgtt and acctgagagactgatcca);

Ptn (tggagctgagtgcaagtacc and ggcgtcttttaatccagcatct); *Rrbp1* (agcaagtgtgaagagctgagtag and actccagctccttgagacga); *Myf5* (ccttgctcagctccctcaa and gccatccgctacattgagag); *Myf6* (gcctcgtgataactgctaagg and gttccaaatgctggctgagt); *Pax3* (gcactattccttcgaacgca and ggttggtcagaagtcccatt); *Pax7* (cctcagtgagttcgattagcc and ggtagtgggtcctctcgaag). For Figure 6, qRT-PCR was performed using custom Format-24 Taqman arrays (ABI and Assuragen, Austin, TX) using mouse or human *GAPDH* as a control for relative gene expression, and 18S RNA as a quality control. Statistical considerations are given in the Supplemental Methods. Probesets for mouse samples were 18S-Hs99999901_s1, Adssl1-Mm00475814_m1, Bmp4-Mm00432087_m1, Cav1-Mm00483057_m1, Chrd-Mm00438203_m1, Chrng-Mm00437419_m1, Ckm-Mm00432556_m1, Dlk1-Mm00494477_m1, Flnc-Mm00471824_m1, Gapdh-Mm99999915_g1, Gjd4-Mm00462088_m1, Hes6-Mm00517097_g1, Jag1-Mm00496902_m1, Myf5-Mm00435125_m1, Myf6-Mm00435126_m1, Myh3-Mm01332463_m1, Myl4-Mm00440378_m1, Notch3-Mm00435270_m1, Pax3-Mm00435493_m1, Pax7-Mm00834082_m1, Sct-Mm00441235_g1, Sema3f-Mm00441325_m1, Tbx2-Mm00436915_m1 and Unc5b-Mm00504054_m1. Probesets for mouse samples were 18S-Hs99999901_s1, ADSSL1-Hs00411846_m1, BMP4-Hs00370078_m1, CAV1-Hs00971716_m1, CHRD-Hs00415315_m1, CHRNG-Hs00183228_m1, CKM-Hs00176490_m1, DLK1-Hs00171584_m1, FLNC-Hs00155124_m1, GAPDH-Hs99999905_m1, GJD4-Hs00542133_m1, HES6-Hs00936587_g1, JAG1-Hs01070036_m1, MYF5-Hs00271574_m1, MYF6-Hs00231165_m1, MYH3-Hs01074230_m1, MYL4-Hs00267321_m1, NOTCH3-Hs01128541_m1, PAX3-Hs00240950_m1, PAX7-Hs00242962_m1, SCT-Hs00360814_g1, SEMA3F-Hs00188273_m1, TBX2-Hs00172983_m1 and UNC5B-Hs00900710_m1.

Pattern Recognition

The classifiers that have been used are LDA (Linear Discriminant Analysis) and KNN (K nearest neighbors) classifier. LDA tries to find the linear combination of the features that best differentiates the classes(Duda and Hart, 2001). KNN rule classifies a sample by assigning it the label most frequently represented among the *k* nearest samples; in other words, a decision is made by examining the labels on the *k* nearest neighbors and taking a vote(Duda and Hart, 2001). In our analysis, the neighbors have been decided based on the smallest Euclidean distances and k has been taken to be 3. KNN is a non-parametric classification technique and is considered to be simple and robust. The error estimation methods used for measuring the accuracy of the classifiers are (1) Re-substitution (resub): The same samples are used for training and testing. (2) Leave one out (loo): One sample at a time is left out for testing and rest are used for training the classifier. (3) 5 fold cross validation (cv): The data is randomly divided into 5 folds and 4 folds are used for training and the 5th fold for testing. This is repeated for x times. In our case, x was 10. (4) Bootstrap (boots): N samples are selected from the data with replacement. The samples not selected are used for testing. This is repeated for y times. In our case, y was 10. (5) .632 bootstrap (.632boots): the error is computed as .632*bootstrap error + .368 * re-substitution error(Kohavi, 1995).

Gene Expression Microarray Analysis

Mouse gene expression data were generated using Illumina Mouse Ref-8 BeadChip v1.1 (Illumina, San Diego, CA). Datasets were deposited in the GEO database (GSE22520). Rank invariant set normalization was performed on the log2-transformed expression value. To determine differential expression between phenotypes, we applied a *t*-test to the normalized expression data. Signature gene sets for a given comparison were selected by adjusted p-value <0.05 (Benjamini-Hochberg correction for multiple tests) and fold change >2. For functional annotation and enrichment, the significant genes of each comparison were uploaded to the **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery (DAVID, http://david.abcc.ncifcrf.gov) to identify enriched biological themes such as KEGG, BioCarta pathways, and Gene Ontology (GO) terms. Hierarchical clustering was performed to generate "heatmaps" by using Pearson correlation coefficient and average linkage for both gene and sample clustering. All bioinformatics tasks were performed with MATLAB/Bioinformatics Toolbox (Mathworks Inc, Natick, MA), unless otherwise noted. We also performed Principal Component Analysis (PCA) on all mouse tissue samples, with the 345 signature differential expressed genes between mouse eRMS and mouse UPS (selected with the criteria of raw p-value <0.05 & fold change > 2 (see Results)), samples expression levels were projected to change greater than the first 3 principal components and then plotted in 3D space for visualization. A similar approach to the human fusion negative soft tissue sarcomas was taken using the criteria of p-value < 0.05 & fold change > 1.5.

All microarray gene expression data are deposited in the NCBI Gene Expression Omnibus database.

Subtype Score (S-score)

In order to confirm the tumor subtype identified by pathologist with the gene expression profiling result, a subtype scoring method was developed to quantify each sample's consistency according to the training result. We briefly describe the method as follows.

Assuming there are *n* genes in the signature gene set as described before, and let x = {$r_{i,j}$, c_i}p, where $r_{i,j}$ is the $\log_2$-transformed expression level of gene *j* at sample *i*, and c_i is the sample class label where $c_i \in$ {-1, 1}, or c_i = 1 if sample *i* belongs to subtype A; otherwise -1. Let, $\rho_{i,j}$ to be the Pearson correlation coefficient between gene *j* and c_i for each gene, or $\rho_{i,j} = \mathrm{cov}(r_{i,j}, c_i)/(\sigma_i \sigma_c)$. p_i is the *t*-test probability, *p*-value, of gene *j* between two subtypes. The Subtype Score (S-score) is defined as,

$$s_i = \frac{1}{K}\sum_{j=1}^{n} sign(\rho_j) p_j^* z_{i,j},$$

where

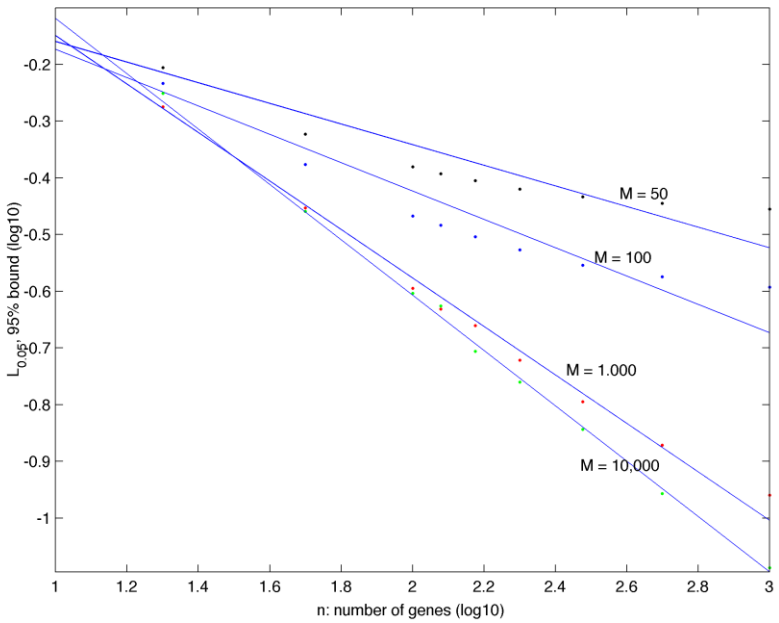
$$p_j^* = \begin{cases} 3, \text{ if } -\log_{10}(p_j) > 4 \\ 0, \text{ if } -\log_{10}(p_j) > 1 \end{cases}$$

$$z_{i,j} = \left(r_{i,j} - \mu_j\right)/\sigma_j^*,$$

$$K = \sum_{j=1}^{n} p_j^* \text{ and } \sigma_j^* = \sqrt{\frac{(n_1 - 1)\sigma_1 + (n_2 - 1)\sigma_2}{n_1 + n_2 - 2}}$$

μ_j and σ_j^* are mean and pooled sample standard deviation of gene *j* across all samples, respectively. Notice that S-score s_i is positive when its standardized expression value *z* and correlation have the same directional sign (positive or negative). The S-score simply average genes with significant test statistics, considering their directional effect across two subtypes. The calculation of S-score is implemented in MATLAB.

No-Call Region of S-score: while the upper bound of the S-score, s_i, depends on the differential gene expression level, the lower bond can be determined by assuming all selected genes' expression levels are randomly distributed (no discernable differential expression). Let $z_{i,j}$ to be normally distributed with *N*(0, 1), and suppose there are *n* genes and total of *M* samples balanced in groups 1 and 2. The simulation results are shown in the graph following this paragraph for *n* = 50 to 1000, and *M* = 10,000, 1,000, 100 and 50. In addition, by assuming s_i^* to be approximately normally distributed and according to the large number theorem, we have the critical value, $L_{0.05} = \pm 1.97\sigma_s$, where σ_s is the standard deviation of s^* obtained from random gene expression value. Samples with S-score less than $L_{0.05}$ will not be classified (no-call group) to control the call-error to be less than 5% due to the random events. As expected, the bounds will be smaller with larger number of genes (chances of making "no call" shall be smaller with more genes in the gene set), or more samples. For example, for *n* = 100 and *M* = 100, we will not make a call for a given sample to be in either group 1 or 2, if s_i is outside (-0.34, 0.34), to guarantee no more than 5% of classification error. For the case of 13 ERMS samples vs. other non-ERMS samples (10), for a total of 345 signature genes, we have the no-call regions of (-0.51, 0.51).



Relationship between number of genes, number of samples, and the no-call boundary $L_{0.05}$.

p53, Shh, Rb1 and Ras signatures in human fusion negative rhabdomyosarcomas We downloaded the public domain datasets for fusion negative rhabdomyosarcoma reported by Davicioni *et al*(Davicioni et al., 2006) from (https://array.nci.nih.gov/caarray/project/details.action?project.experiment.publicIdentifier=trich-00099) as well as eRMS rhabdomyosarcoma datasets from Lae *et al*(Lae et al., 2007) and Wachtel *et al*(Wachtel et al., 2004). These fusion negative RMS and eRMS datasets were designated as the test samples, whereas normal skeletal muscles (SKM) samples reported by Bakay *et al*(Bakay et al., 2002) were used as the control group. We also downloaded signature specific datasets as described in the paragraphs below. All the data had been performed on Affymetrix U133A array platform (Affymetrix, Santa Clara, CA). Sample IDs used are given in Table S6.

To examine whether human fusion negative RMS and eRMS tumors had evidence of **p53 loss of function**, we downloaded positive control datasets for the p53 loss of function gene signature in breast cancer(Miller et al., 2005), available in the GEO database. The breast cancer samples for p53 loss of function gene signature were treated as the positive controls, and the breast cancer samples without evidence of p53 loss of function were treated as the negative controls. Normal SKM was also treated as a control group, whereas gene-wise *t*-test were performed between the two groups to derive the p53 loss of function signature genes specific to muscle (**Benjamini and Hochberg adjusted p-value<0.05 & fold change >2**). The two groups were also the training subtypes of S-score and all the samples were sorted based on their S-score. For metagene representations, the average z-values of SKM samples that are greater than 0 were represented as metagene 1, while those less than 0 were represented as metagene 2. The same methodology was used for representing metagenes for Shh and Rb1 signatures.

To examine whether human fusion negative RMS and eRMS tumors had evidence of **Shh gain of function**, we downloaded gene expression datasets for medulloblastoma samples known to exhibit a Shh gain of function signature(Thompson et al., 2006), available at http://www.stjuderesearch.org/data/medulloblastoma/. Sample IDs used are given in Supplemental Table S5B2. All the data had been performed on Affymetrix U133A array platform (Affymetrix). Because of concerns that brain and muscle specific Shh gain of function signatures may have marked differences, a list of genes activated and suppressed by Shh in muscle and similar cell types or tissues was created from a literature search, with hand-curated cross-correlation between probesets for mouse and human gene expression

analysis platforms (Supplemental Table S7). An S-score was constructed based on 124 RMS samples and 18 skeletal muscle samples (as described above) with the manually-curated Shh signature set for muscle. We then calculated the combined samples (total of 124+18+14+14 = 170 samples), and ordered them according to the S-score. Samples with S-score within 0.44 to -0.44 were assigned to be in "no-call group". To produce a heatmap, we also order the Shh signature genes according to their relative expression level (log-transformed ratio between RMS and SKM).

To examine whether human fusion negative RMS and eRMS tumors had evidence of **Rb1 loss of function**, we took genes from the supervised hierarchial clustering of *Rb1* wildtype and homozygous *Rb1* deleted fusion negative mouse sarcomas presented in Supplemental Figure 3B and matched these genes to probesets for human gene expression microarrays using hand-curation (Supplemental Table S8). S-scores were constructed and then evaluated for all RMS and SKM samples. All data were derived from Affymetrix U133A GeneChip. For heatmap display, samples were ordered based on their S-score, and genes were ordered based according to their relative expression level (log-transformed ratio between SKM and RMS).

To examine whether human fusion negative RMS and eRMS tumors had evidence of **Ras activation**, we used previously established gene lists for the activated Ras signature of zebrafish eRMS (Langenau et al., 2007):

1. Ras signature in fish eRMS common to Ras driven pancreatic cancer (87 unique genes), or

2. Ras signature in fish eRMS common to Ras activated mammary epithelial cells (112 unique genes)

Probes were matched manually between zebrafish and human platforms. For the first gene list (pancreatic), we used an S-score threshold of +/- 0.477 to define a 99% confidence interval that discerns samples with a Ras signature from samples without such a signature. For the second gene list (mammary), we used an S-score threshold of +/- 0.435 to define a 99% confidence interval that discerns samples with a Ras signature from samples without such a signature.

Quantitative RT-PCR Expression Analysis

Average gene expressions were collected on 22 genes for 66 patients with various tumor types. The 6 tumor classifications were Normal Skm (n=12), ERMS (n=24), ERMS spindle cell variant (n=10), RMS spindle cell (n=6), Spindle cell sarcoma (n=9), and ARMS (n=5). Across all patients for the 22 genes (n=1452), there was 4.5% were randomly missing a gene expression. K-nearest neighbor method was used to impute the 66 missing gene expressions. We reduced the number of genes used by the SVM to classify subjects using the F_SSFS method (Lee, 2009). The F_SSFS method selects genes on the basis of an F-score to quantify the separation between the values of the gene and each of the tumor categories. Only those genes with an F-score greater than a certain predetermined threshold, which is based on the dimension of the data, are considered. The F_SSFS method then uses forward selection with genes whose F-score exceeds the threshold added to the selected subset that will be used in the final reduced classifier on the basis of the improvement in the accuracy of the SVM. To validate the final reduced classifier, we computed sensitivity, specificity, positive predictive value, negative predicative value and accuracy. After identifying the optimal subset of genes, a Leave-One-Out cross validation method was used to test the validity and accuracy of the SVM. These same procedures were then applied to a subset of the data with the intermediate diagnosis groups, ERMS spindle cell variant and RMS spindle cell removed.

SUPPLEMENTAL REFERENCES

Bakay, M., Zhao, P., Chen, J., and Hoffman, E. P. (2002). A web-accessible complete transcriptome of normal human and DMD muscle. Neuromuscul Disord 12 Suppl 1, S125-141.

Duda, R. O., and Hart, P. (2001). Pattern Classification., 2nd edn: New York: John Wiley & Sons).

Keller, C., and Capecchi, M. R. (2005). New genetic tactics to model alveolar rhabdomyosarcoma in the mouse. Cancer Res 65, 7530-7532.

Keller, C., Hansen, M. S., Coffin, C. M., and Capecchi, M. R. (2004b). Pax3:Fkhr interferes with embryonic Pax3 and Pax7 function: implications for alveolar rhabdomyosarcoma cell of origin. Genes Dev 18, 2608-2613.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. International Joint Conference on Artificial Intelligence (IJCAI), 1137-1145.

Lee, M. (2009). Using support vector machine with a hybrid feature selection method to the stock trend prediction. Expert Systems with Applications, 10896-10904.

Marino, S., Hoogervoorst, D., Brandner, S., and Berns, A. (2003). Rb and p107 are required for normal cerebellar development and granule cell survival but not for Purkinje cell persistence. Development 130, 3359-3368.

Marino, S., Vooijs, M., van Der Gulden, H., Jonkers, J., and Berns, A. (2000). Induction of medulloblastomas in p53-null mutant mice by somatic inactivation of Rb in the external granular layer cells of the cerebellum. Genes Dev 14, 994-1004.

Miller, L. D., Smeds, J., George, J., Vega, V. B., Vergara, L., Ploner, A., Pawitan, Y., Hall, P., Klaar, S., Liu, E. T., and Bergh, J. (2005). An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proc Natl Acad Sci U S A 102, 13550-13555.

Nishijo, K., Chen, Q. R., Zhang, L., McCleish, A. T., Rodriguez, A., Cho, M. J., Prajapati, S. I., Gelfond, J. A., Chisholm, G. B., Michalek, J. E., et al. (2009). Credentialing a preclinical mouse model of alveolar rhabdomyosarcoma. Cancer Res 69, 2902-2911.

Scholkopf, B., Kah-Kay, S., Burges, C. J. C., Girosi, F., Niyogi, P., Poggio, T., and Vapnik, V. (1997). Comparing support vector machines with Gaussian kernels to radial basis function classifiers. Signal Processing, IEEE Transactions on 45, 2758-2765.

Taniguchi, E., Cho, M. J., Arenkiel, B. R., Hansen, M. S., Rivera, O. J., McCleish, A. T., Qualman, S. J., Guttridge, D. C., Scott, M. P., Capecchi, M. R., and Keller, C. (2009). Bortezomib reverses a post-translational mechanism of tumorigenesis for patched1 haploinsufficiency in medulloblastoma. Pediatr Blood Cancer 53, 136-144.

Thompson, M. C., Fuller, C., Hogg, T. L., Dalton, J., Finkelstein, D., Lau, C. C., Chintagumpala, M., Adesina, A., Ashley, D. M., Kellie, S. J., et al. (2006). Genomics identifies medulloblastoma subgroups that are enriched for specific genetic alterations. J Clin Oncol 24, 1924-1931.

Wachtel, M., Dettling, M., Koscielniak, E., Stegmaier, S., Treuner, J., Simon-Klingenstein, K., Buhlmann, P., Niggli, F. K., and Schafer, B. W. (2004). Gene expression signatures identify rhabdomyosarcoma subtypes and detect a novel t(2;2)(q35;p23) translocation fusing PAX3 to NCOA1. Cancer Res 64, 5539-5545.