Supplemental Information for:

The Ewing's Sarcoma Oncoprotein EWS/FLI Induces a p53-Dependent Growth Arrest in Primary Human Fibroblasts

Stephen L. Lessnick, Caroline S. Dacwag, and Todd R. Golub

Appendix 1: (below) Detailed description of the methods used in the microarray analyses.

Appendix 2: Excel file containing the complete expression data from the tet-EF cell time course experiment. The data is the output from the Affymetrix GeneChip software, after scaling and thresholding, as described in appendix 1.

Appendix 3: File containing reference pattern for tet-EF cell Knn analysis described in figure 2. File is in .cls format, for use with the GeneCluster program.

Appendix 4a-d: Files containing reference patterns for small-round cell tumor analysis described in Figure 2. Files are in .cls format, for use with the GeneCluster program. Ewing's sarcoma (4a), Burkitt's lymphoma (4b), neuroblastoma (4c), and rhabdomyosarcoma (4d) patterns are included.

Appendix 5: Excel file containing the tet-EF dataset (from appendix 2) limited to those probes that are also represented on the cDNA microarray of Khan et al., 2001

Appendix 6: Excel file containing the small-round cell data (of Khan et al. 2001) limited to those cDNAs that are also represented on the Affymetrix U95Av2 microarray.

Appendix 7a-d: Excel files containing list of small-round cell tumor-specific genes as determined by Knn analysis and permutation testing (at the 1% significance level), as described in appendix 1. The data is the output from the GeneCluster 2.0 program, and as such includes the values from the permuation analysis. Each file also contains the list of the equivalent Affymetrix probe names, and additionally contains the list of overlap genes between each small-round cell tumor and the tet-EF cell data, as described in appendix 1. Ewing's sarcoma (7a), Burkitt's lymphoma (7b), neuroblastoma (7c), rhabdomyosarcoma (7d).

Appendix 8: Excel file containing the list of 117 upregulated genes in tet-EF cells following induction of EWS/FLI expression.

Appendix 9: Excel files of the SOM clusters from tet-EF cells.

Appendix 1: Transcriptional profiling methodologies.

Tet-EF time course sample preparation:
Tet-EF cells were plated at $1x10^6$ cells per 15 cm plate, five plates in total, in tetracycline-containing media. On day zero, one plate of cells was collected, pelleted, and stored frozen in Trizol reagent (Life Technologies). The remaining plates were induced to express EWS/FLI by washing in phosphate-buffered saline and refeeding with tetracycline-free media. On each subsequent day one plate was collected, pelleted, and stored in Trizol reagent (Life Technologies; run 1). The entire induction experiment was repeated on a subsequent week to generate an independent duplicate sample (run 2). Following collection of the final sample, total RNA was isolated using the manufacturer's instructions. RNA quality was assessed by denaturing gel electrophoresis. In all instances, the ribosomal RNA bands were intact.

Microarray hybridization:
The amount of starting total RNA for each reaction was approximately 10 μg. First strand cDNA synthesis was generated using a T7-linked oligo-dT primer, followed by second strand synthesis. An *in vitro* transcription reaction was performed to generate cRNA containing biotinylated UTP and CTP, which was subsequently chemically fragmented at 95°C for 35 minutes. Ten micrograms of the fragmented, biotinylated cRNA was hybridized in MES buffer (2-[N-Morpholino]ethansulfonic acid) containing 0.5 mg/ml acetylated bovine serum albumin (Sigma, St. Louis) to Affymetrix (Santa Clara, CA) U95Av2 arrays at 45°C for 16 hours. U95Av2 arrays contain approximately 12,600 known genes and expressed sequence tags. Arrays were washed and stained with streptavidin-phycoerythrin (SAPE, Molecular Probes). Signal amplification was performed using a biotinylated anti-streptavidin antibody (Vector Laboratories, Burlingame, CA) at 3 μg/ml. This was followed by a second staining with SAPE. Normal goat IgG (2 mg/ml) was used as a blocking agent. Scans were performed on Affymetrix scanners and the expression value for each gene was calculated using Affymetrix GeneChip software. Minor differences in microarray intensity were corrected using a linear scaling method as detailed in the next section.

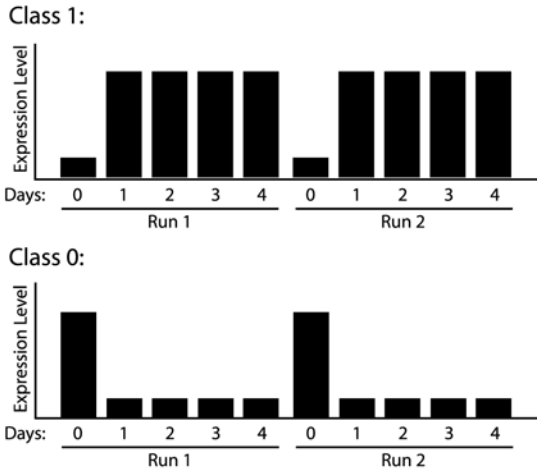
Preprocessing and re-scaling:
The raw expression data as obtained from Affymetrix's GeneChip was re-scaled to account for different chip intensities. Each column (sample) in the dataset was multiplied by *1/slope* of a least squares linear fit of the sample vs. the reference sample. This linear fit is done using only genes that have 'Present' calls in both the sample being re-scaled and the reference. The sample chosen as reference is a typical one (i.e. one with the number of "P" calls closer to the average over all samples in the dataset). Scans were rejected if the scaling factor exceeded a factor of 3.5, fewer than 1000 genes received 'Present' calls, or microarray artifacts were visible.

Following scaling, a ceiling value of 16,000 units was chosen for all experiments because it is at this level that we observe fluorescence saturation of the scanner; values above this cannot be reliably measured. A threshold of 10 units was chosen for all experiments because expression levels below this value demonstrate significant noise and poor reproducibility. Analysis of samples (as described below) was performed on the scaled data to which the ceiling and threshold values had been applied. The complete dataset after scaling and thresholding is included as appendix 2.

Knn analysis of tet-EF cells:

To identify genes that have increased expression following the induction of EWS/FLI in tet-EF cells, we used K-nearest neighbors (Knn) analysis (Golub et al., 1999). Simply stated, Knn analysis determines which genes from a given dataset match an investigator-determined pattern. Each gene in the dataset is compared to the reference pattern and a correlation value is determined mathematically. The genes are then sorted to identify those with the best correlation to the reference pattern.

For the purposes of this analysis, a reference pattern was chosen to identify genes that increase expression following induction of EWS/FLI protein. It is difficult to know, *a priori*, what expression pattern an upregulated gene should have. For example, an upregulated gene could demonstrate near maximal expression immediately following EWS/FLI induction, it could demonstrate a slow, steady increase over the total time of the experiment, or it could demonstrate any number of other patterns. We therefore chose an initial reference pattern in which gene expression rapidly increases following EWS/FLI induction (see figure below). A two class distinction is thus generated, with class 0 consisting of genes whose expression levels correlate with the day 0 timepoint, (i.e., expressed highly at time 0, and decrease their expression level in the subsequent days), and class 1 consisting of genes that correlate with with days 1-4 (i.e., that are expressed at relatively low levels at time 0, and increase their expression level in the subsequent time points; see figure below). The file for the reference pattern is included as appendix 3.



The reference pattern utilizes both experimental runs generated, as shown graphically in the above figure. Thus, the time points are neither blended nor averaged, but rather treated separately.

Genes correlated with the particular class distinctions (e.g. class 0 and class 1) were identified by sorting all of the genes on the array according the signal-to-noise statistic (Golub et al., 1999) $(\mu_{class\ 0} - \mu_{class\ 1})/(\sigma_{class\ 0} + \sigma_{class1})$ where μ and σ represent the mean and standard deviation of expression, respectively, for each class. Permutation of the column (sample) labels was performed to compare these correlations to what would be expected by chance (see the next section).

Permutation Test and Neighborhood Analysis for Marker Genes:

A permutation test was used to calculate whether the top marker genes with respect to a biologically meaningful phenotype (e.g. following induction of EWS/FLI) were statistically significant (Golub et al., 1999). To do this we compared the top signal-to-noise scores for top marker genes and compared them with the corresponding ones for random permutation versions of the class labels (phenotype). Typically 1000 random permutations were used to build histograms for the top marker, the second best, etc. Based on this histogram we determined the 5% and 1% significance levels and compared them with the values obtained for the real data set.

This procedure is motivated by considering the following question: what is the likelihood that the set of markers genes, for example selected by signal-to-noise or any other distance or correlation measure, of a phenotype of interest represent chance correlations and not any biological significant match? If one moves down the list of markers, how many could one consider as being significantly correlated and not the results of chance correlations?

In detail the permutation test procedure is as follows:

- Generate signal-to-noise $(\mu_{class\ 0} - \mu_{class\ 1})/(\sigma_{class\ 0} + \sigma_{class1})$ scores for all genes using the actual class labels (phenotype) and sort them accordingly. The best match (k=1) is the gene “closer” or more correlated to the phenotype using the signal-to-noise as a distance function. In fact one can imagine the reciprocal of the signal-to-noise as a “distance” between the phenotype and each gene as shown in the figure below.
- Generate 1000 random permutations of the class labels (phenotype). For each case of randomized class labels generate signal-to-noise scores and sort genes accordingly.
- Build a histogram of signal-to-noise scores for each value of k. For example, one for all the 1000 top markers (k=1), another one for the 1000 second best (k=2) etc. These histograms represent a reference statistic for the best match, second etc. and for a given value of k different genes

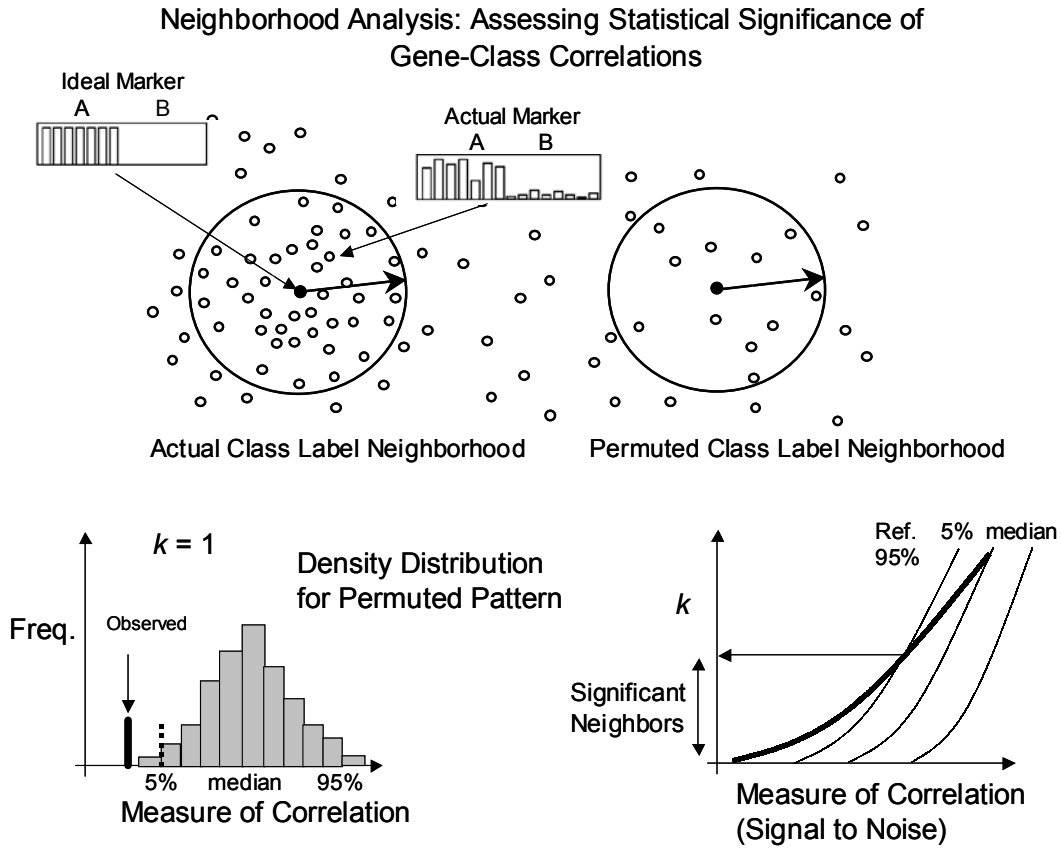
contribute to it. Notice that the correlation structure of the data is preserved by this procedure. Then for each value of k one determines the 5% and 1% significance levels. See the bottom diagrams in the figure.

- Compare the actual signal-to-noise scores with the different significance levels obtained for the histograms of permuted class labels for each value of k. This test helps to assess the statistical significance of gene markers in terms of target class-correlations.

In the results section the values for permutation tests of marker genes are reported in tables with this format:

Distinction	Distance	Perm 1%	Perm 5%	Feature	Desc
class 0	0.96694607	1.0144908	0.8333578	M93119_at	INSM1 Insulinoma-associated 1
class 0	0.9096911	0.8600172	0.7669801	M30448_s_at	Casein kinase II beta subunit
class 0	0.90010124	0.85051423	0.7251496	S82240_at	RhoE
class 0	0.832689	0.84354156	0.7071885	U44060_at	Homeodomain protein (Prox 1)
class 0	0.83225346	0.8009565	0.68034023	D80004_at	KIAA0182 gene
.....	.....	.....	.....	.....	.....
class 1	1.6520017	0.9831643	0.84544426	X86693_at	High endothelial venule
class 1	1.2436218	0.88150144	0.7559189	M93426_at	PTPRZ Protein tyrosine phosphatase, receptor-type, zeta polypeptide
class 1	1.2317128	0.86047184	0.70928395	U48705_rna1_s_at	Receptor tyrosine kinase DDR gene
class 1	1.2259983	0.8433512	0.68909335	X86809_at	Major astrocytic phosphoprotein PEA-15
class 1	1.214929	0.8281318	0.6849929	U45955_at	Neuronal membrane glycoprotein M6b mRNA, partial cds
class 1	1.2095517	0.79365546	0.6711517	U53204_at	Plectin (PLEC1) mRNA
.....	.....	.....	.....	.....	.....

The distinction represents the class for which the markers are high (and low in the other classes). Distance is the signal to noise to the actual phenotype. Perm. 1% and 5% and the corresponding percentiles (significance levels) in the histograms of random permutation signal to noise scores for a given value of k. Feature is the gene accession number and Description the gene name and annotation.



Analysis of SRCT cDNA array data: Khan et al., generated cDNA microarray data for 63 small round cell tumors (SRCT) of childhood, including 23 Ewing's sarcomas, 8 Burkitt's lymphomas, 12 neuroblastomas, and 20 rhabdomyosarcomas (Khan et al., 2001). The list of 2308 genes included in the data analysis was downloaded from http://www.nhgri.nih.gov/DIR/Microarray/Supplement and converted into a .res file for use in the GeneCluster 2.0 program (Tamayo et al., 1999). These data were pre-filtered (by Khan et al.) to remove uninformative features.

The cDNA microarray dataset of SRCT samples and the Affymetrix microarray dataset of tet-EF expression are in different formats. The cDNA microarray dataset contains IMAGE clone numbers to identify genes, while the Affymetrix microarray dataset uses Genbank accession numbers for gene identification. To directly compare the two datasets, each probe was first converted to a Genbank entry, and was subsequently mapped to a Unigene cluster. Unigene clusters which are present on both microarray platforms were identified. Each cluster was then "back-mapped" onto their original dataset. In this way were were able to limit the datasets to only those genes that were interrogated by both platforms. We use the term "probes" to indicate Affymetrix oligonulceotide probesets corresponding to an individual gene. We use the term "cDNA" to indicate the cDNA probe from the SRCT samples corresponding to an individual gene. The limited datasets are included as appendices 5 and 6.

We used Knn analysis (as described above) to identify genes whose expression is higher in each of the tumor types than in the other tumor types represented in the dataset. For example, we used Knn analysis to identify genes whose expression is "high in Ewing's sarcoma, low in other tumor types." Permutation testing (as described above) was used to define the list of specific genes that correlate with the tumor type better than is expected by chance alone. We repeated this process for each tumor type, in each instance comparing those samples to the remaining samples. The files for the reference patterns are included as appendices 4a-d.

To allow for comparisons of the resultant gene lists between the SRCT cDNA microarray and the tet-EF cell microarray data, we identified the Affymetrix probes that corresponded to each cDNA present on the SRCT microarray. In many instances the Affymetrix array uses more than one probeset for each gene, and so the number of probes identified exceeds the number of cDNAs present on the SRCT array.

We identified 686 probes whose expression correlated with Ewing's sarcoma at the 1% significance level, or better. Similarly, there were 744 genes for Burkitt's lymphoma, 772 genes for neuroblastoma, and 931 genes for rhabdomyosarcoma. The output from the GeneCluster program was converted to a Microsoft Excel spreadsheet and is included as appendices 7a-d of the supplemental information.

We compared the lists of probes generated by Knn analysis and permutation testing for each of the SRCT types to the data obtained for the tet-EF cells (as described above). There were 117 probes that were upregulated in the tet-EF cells (at the 1% significance level), taken from the total of 2529 available probes to query. Thus, 117 of 2529 probes, or 4.6% of the probes were upregulated following EWS/FLI expression. The list of these genes is included as appendix 8.

We reasoned that if there were no similarity between the tumor dataset and the tet-EF cell dataset (the null hypothesis), then one would expect to identify 4.6% overlap between the tumor-specific data and the tet-EF data by chance alone. Thus, in the case of Ewing's sarcoma, for example, there were 686 probes that were markers of Ewing's sarcoma at the 1% significance level. The expected overlap between Ewing's-specific genes and tet-EF upregulated genes would be 4.6% of 686 genes, or 32 genes. We performed Chi-square analysis to determine a p value for the observed overlap as compared to the computationally determined expected overlap for each tumor type.

<u>Cluster analysis using self-organizing maps (SOMs):</u>
Knn analysis and permutation testing is a useful tool to identify genes that behave in a predetermined manner (e.g., that increase in response to EWS/FLI expression in a particular pattern). Cluster analysis, on the other hand, allows

one to identify the predominant patterns of gene expression without predetermination. Self-organizing maps (SOMs) were generated using our Genecluster software (Tamayo et al., 1999).

The data set for tet-EF cells induced to express EWS/FLI were scaled, and had thresholds and ceilings applied as described above. Because we were interested in identifying genes that change during the course of the experiment, we filtered the list to include only those genes that had at least a 2.5 fold increase or decrease in at least two time points as compared to the day 0 sample of that particular experimental run.

This filtering resulted in 695 genes that demonstrated significant variation. We used the GeneCluster 2.0 program to generate three clusters (a 1x3 SOM). Simply stated, with the SOM, one randomly chooses the geometry of the grid (e.g., a 1 x 3 grid) and maps it into the k-dimensional feature space. Initially the features are randomly mapped to the grid but during training the mapping is iteratively adjusted to reflect the data structure. Thus, the predominant patterns of gene expression are identified in an unbiased fashion.

The 1x3 SOM generated three clusters that neatly subdivided the data into genes that were upregulated, genes that were downregulated, and genes that had a variable expression pattern. The genes present in each cluster are included in appendix 9 of the supplemental information.

References:

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A.*, et al.* (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science *286*, 531-537.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., and Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med *7*, 673-679.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci U S A *96*, 2907-2912.