*Priority Report*

Cancer Research

# A Novel Algorithm for Simplification of Complex Gene Classifiers in Cancer

Raphael A. Wilson[1], Ling Teng[8], Karen M. Bachmeyer[8], Mei Lin Z. Bissonnette[9], Aliya N. Husain[9], David M. Parham[2], Timothy J. Triche[3], Michele R. Wing[4], Julie M. Gastier-Foster[4], Frederic G. Barr[5], Douglas S. Hawkins[6], James R. Anderson[7], Stephen X. Skapek[1], and Samuel L. Volchenboum[8,10,11]

## Abstract

The clinical application of complex molecular classifiers as diagnostic or prognostic tools has been limited by the time and cost needed to apply them to patients. Using an existing 50-gene expression signature known to separate two molecular subtypes of the pediatric cancer rhabdomyosarcoma, we show that an exhaustive iterative search algorithm can distill this complex classifier down to two or three features with equal discrimination. We validated the two-gene signatures using three separate and distinct datasets, including one that uses degraded RNA extracted from formalin-fixed, paraffin-embedded material. Finally, to show the generalizability of our algorithm, we applied it to a lung cancer dataset to find minimal gene signatures that can distinguish survival. Our approach can easily be generalized and coupled to existing technical platforms to facilitate the discovery of simplified signatures that are ready for routine clinical use. *Cancer Res; 73(18); 5625–32.* ©*2013 AACR.*

## Introduction

High-resolution molecular genetic tools and transcriptome-wide gene expression profiling have revealed molecular subtypes within seemingly uniform cancers (1). Nevertheless, few molecular classifiers have been implemented in the clinic. Supervised analytic methods such as decision trees, support vector machines, and Naïve Bayes classification, all facilitate the generation of multigene classifiers able to separate two or more clinically significant classes. Until now, such algorithms were needed to compensate for the enormous number of feature combinations and the time needed to test all of them. As a result, the literature is rife with complex, multigene signatures of uncertain clinical value. Faster computers and parallelization now make it possible to rapidly test an enormous number of feature combinations. We hypothesized that a brute-force exhaustive search algorithm could be used to supplant a complex gene classifier determined using traditional supervised methods with a much simpler set of features.

We tested this concept on rhabdomyosarcoma, the most common soft tissue sarcoma in children (2, 3). Currently, a compilation of clinical and pathology features separates children with rhabdomyosarcoma into low, intermediate, and high-risk groups with cure rates approximating 90%, 65%, and 25%, respectively (4, 5). This has allowed more "precise," risk-adapted therapy to mollify untoward effects of intensive chemotherapy in the favorable group while intensifying treatment for those less likely to be cured. Most children fall into the intermediate-risk group where outcome prediction is difficult. Retrospective studies (6, 7) suggest the feasibility of a molecular prognostic signature for this group, but the clinical use may be limited by time and cost.

Attempting to simplify molecular classifiers for rhabdomyosarcoma, we leveraged the separation of tumors into two major histologic subtypes with "alveolar" (higher risk) and "embryonal" (lower risk) morphologic features (2). Close to 80% of alveolar tumors contain a chromosomal translocation fusing the 5′ end of the *PAX7* gene (chromosome 1) or *PAX3* gene (chromosome 2) to the 3′ coding sequence of *FOXO1* (chromosome 13; ref. 8; and references therein). As the presence or absence of the *PAX–FOXO1* fusion correlates more closely with gene expression than does the histologic subtype (7, 9), we attempted to define the simplest molecular signature to discriminate *PAX–FOXO1* fusion-positive from fusion-negative

**Authors' Affiliations:** [1]Division of Hematology/Oncology, Department of Pediatrics, University of Texas Southwestern Medical Center and Children's Medical Center, Dallas, Texas; [2]Department of Pathology, University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma; [3]Department of Pathology, University of Southern California and Children's Hospital of Los Angeles, Los Angeles, California; [4]Department of Pathology and Laboratory Medicine, Nationwide Children's Hospital and Department of Pathology and Pediatrics, The Ohio State University, Columbus, Ohio; [5]National Cancer Institute, Bethesda, Maryland; [6]Division of Hematology/Oncology, Department of Pediatrics, University of Washington and Seattle Children's Hospital, Seattle, Washington; [7]Department of Biostatistics, College of Public Health, University of Nebraska Medical Center, Omaha, Nebraska; [8]The Section of Pediatric Hematology/Oncology, Department of Pediatrics, [9]Department of Pathology, [10]Center for Research Informatics, and [11]The Computation Institute, University of Chicago, Chicago, Illinois

**Note:** Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

**Corresponding Authors:** Samuel L. Volchenboum, University of Chicago, 900 E. 57th Street, KCBD 5130, Chicago, IL 60637. Phone: 773-702-4303; Fax: 773-834-1329; E-mail: slv@uchicago.edu; and Stephen X. Skapek, Division of Hematology/Oncology, Department of Pediatrics, University of Texas Southwestern Medical Center and Children's Medical Center, Dallas, TX 75390. E-mail: stephen.skapek@utsouthwestern.edu

**doi:** 10.1158/0008-5472.CAN-13-0324

disease as a first step toward developing a useful molecular tool to discern prognosis.

## Materials and Methods

### Manipulation and analysis of published gene expression signatures

All gene expression analyses were conducted in R/BioConductor using the "affy" and "affyPLM" packages and the Robust Multiarray Average (RMA) normalization algorithm. Principal component analysis (PCA) was conducted using the prcomp function in R (http://bit.ly/O10aUw). Subsequent analyses were conducted using scripts written in Python 2.7. An exhaustive search was conducted by iterating over all possible combinations of 1, 2, and 3 genes from a given subset. Supervised analysis was conducted using linear support vector classification (SVC) using the scikit-learn module for Python (http://bit.ly/rEVRFS; ref. 10) to construct a hyperplane best separating the fusion-negative and fusion-positive genes and to test the resulting classifier model. To determine the most differentially expressed genes, a *t* test was conducted, and the genes were sorted according to the mean of the distribution. The most positive *t* test values were associated with the fusion-negative samples. All possible pairs from the 1,000 probes at the top (most fusion negative) and bottom (most fusion positive) of the list were used to make pairs for testing (i.e., 1,000 top vs. 1,000 bottom = 1,000,000 combinations). All software was run on standard personal computer configurations. Performance on standard desktop and laptop hardware was adequate for analysis. The iterative search of 1 million gene dyads from the 2,000 most differentially expressed genes ran in about 6 hours.

Two publicly available rhabdomyosarcoma gene expression datasets were obtained. The first consisted of gene expression data from 101 patients with rhabdomyosarcoma (ref. 9; available for download here: http://bit.ly/ISG5uF). The metagenes corresponding to fusion-positive (F1) and fusion-negative disease (F2) from Williamson and colleagues (9), each with 25 genes, were chosen for further study, as detailed above. A second set of publicly available data consisted of 139 published alveolar and embryonal rhabdomyosarcoma expression profiles (7; ref. available for download here: http://stat.ethz.ch/R-manual/R-patched/library/stats/html/prcomp.html). Gene expression data in these cases were normalized, and PCA was conducted using 50 genes in the Williamson F1/F2 metagenes. In addition, the most discriminatory combinations, as determined for the Williamson data through exhaustive search, were tested on the Davicioni dataset using the linear SVC methods defined above. Finally, we obtained a set of exon array data on 49 rhabdomyosarcoma tumor specimens for which fusion status was known (J.R. Anderson and T.J. Triche, unpublished data). The data were normalized using Affymetrix Power Tools, unsupervised and supervised PCA was conducted, and our classifiers were tested on these data.

To conduct an exhaustive search over all possible pairs from the 54,613 genes in the microarray, we ran our iterative search algorithm in parallel on a high-performance computing cluster consisting of 240 compute cores (twenty 12-core Intel Westmere nodes with 24 Gb RAM). The Swift parallel scripting language (11) was used to distribute the jobs over the compute nodes. Evaluating all 1.5 billion gene pairs took less than 6 hours. To find the most interesting pairs, the resulting list of gene dyads with classification efficiency more than 95% was parsed to remove the pairs containing the most highly differentially expressed genes. The remaining pairs were tested on the Davicioni and exon array datasets (Supplementary Table S4 and Supplementary Fig. S3). When testing gene pairs, the same probe ID was used where possible. If not available, the same gene symbol was used, and the best classification efficiency was used when there were multiple probes. For the exon array data, the gene symbol was used to query the expression levels. Not all genes were available across all samples.

Computational analyses for the nCounter data were carried out using normalized gene expression with the six housekeeping genes removed. The samples, each with 46 features, were subjected to unsupervised analysis with PCA. Using linear SVC, the gene combinations found above for the Williamson data were also tested on the nCounter data. Finally, all possible combinations of 1, 2, and 3 of the 46 genes were tested using linear SVC to search for the most robust classifier. For design and testing of the nCounter CodeSet and acquisition of the nCounter data, see Supplementary Materials and Methods.

Normalized expression data for 442 patients with non–small cell lung cancer (12) were provided by Yang Xie (University of Texas Southwestern Medical Center, Dallas, TX). Our exhaustive search algorithm was similarly applied to this dataset. First, the samples were filtered to include only those patients who had received chemotherapy. The samples were classified as chemo unresponsive (survival < 24 months, 27 samples) and chemoresponsive (survival > 48 months, 41 samples or survival > 60 months, 32 samples). The algorithm was used to test all the single (44,000) and the top 1,000 two-gene combinations (1,999,000) using SVC as in our original rhabdomyosarcoma studies. The top 100 three-gene combinations (1,313,400) and top 40 four-gene combinations (658,008) were also tested for classification efficiency.

## Results

Two separate teams previously used transcriptome data from archived, freshly frozen specimens to develop complex gene classifiers that can discriminate fusion-positive from fusion-negative rhabdomyosarcoma (7, 9). After recapitulating the PCA using the 50-gene Williamson F1/F2 classifier (Fig. 1A), an exhaustive search algorithm using linear SVC was used to iterate among all the possible two-, three-, and four-gene combinations of the 50 genes to define a simpler classifier. Twenty-eight two- and three-gene combinations correctly classified at least 98% of the samples (Table 1; Fig. 2A and Supplementary Fig. S1). One persistent outlier (Fig. 2A and Supplementary Figs. S1–S3A, green dot) harbored a recently described aberrant fusion involving the *NCOA1* gene (13, 14), and it consistently classified with the fusion-negative samples, as previously described (9). The 630 four-gene combinations with classification efficiency of more than 99% (Supplementary Table S1) contained one of the three-gene combinations (Table 1) 61% of the time.

We then expanded our search for two-gene classifiers to all 54,613 genes in the microarray. We first limited our search to

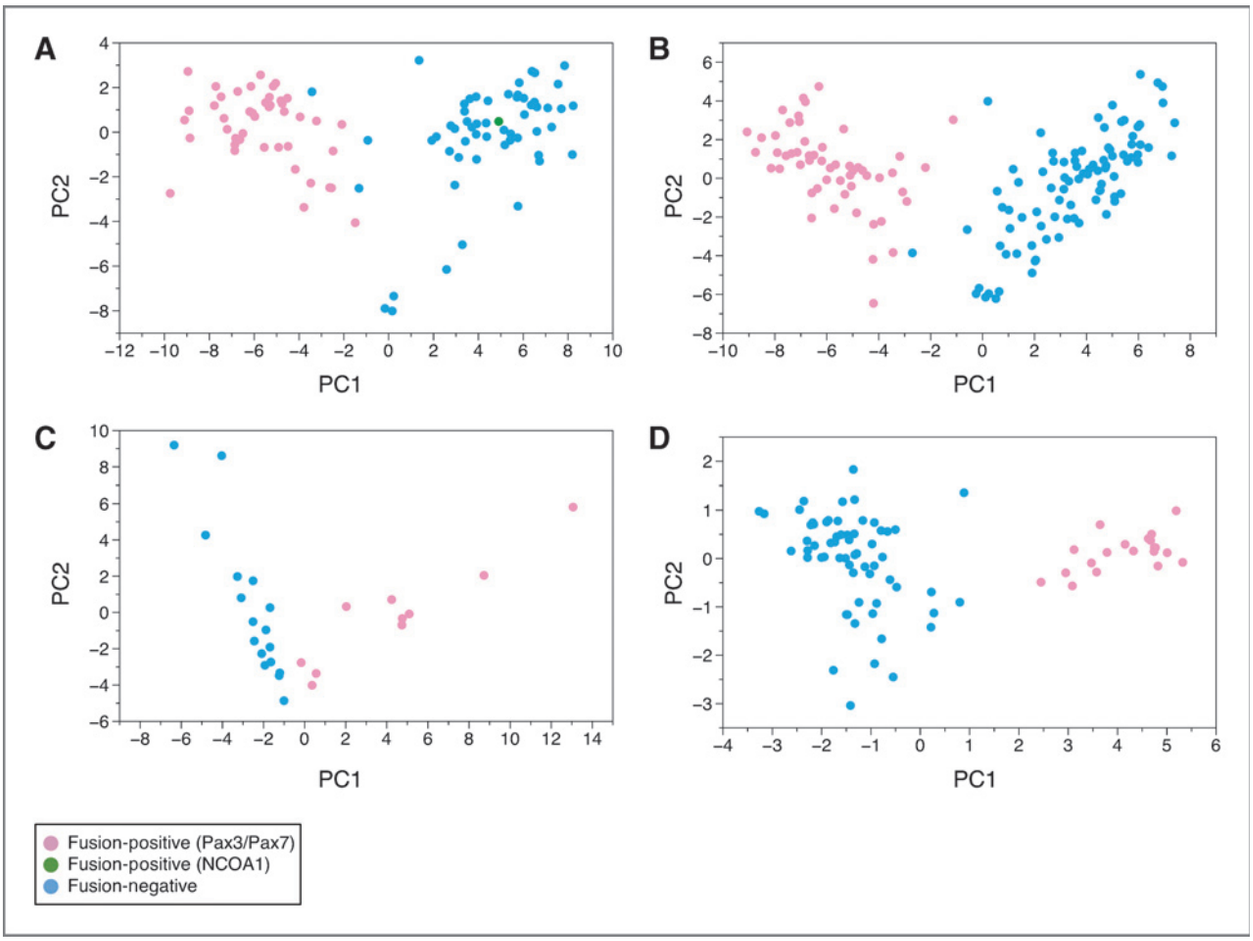


Figure 1. Principal component analyses of the Williamson (A), Davicioni (B), nCounter (C), and exon array (D) datasets using the 50-gene Williamson F1/F2 classifier. Graphs of the first 2 principal components show excellent separation between fusion-positive (pink) and fusion-negative (blue) samples in all 4 datasets. Note that the sample with PAX3–NCOA1 fusion gene in the Williamson dataset (A, green dot) clusters with the fusion-negative samples.

the 1,000,000 dyads formed by the 1,000 most differentially expressed genes from fusion-positive and fusion-negative samples, as determined by *t* test. Thirteen two-probe/gene combinations resulted in classification of more than 99% (Table 1; Supplementary Table S2 and Supplementary Fig. S2), whereas 3,326 two-probe/gene combinations differentiated the fusion-positive from fusion-negative samples with >98% accuracy (Supplementary Table S2). Only 29 of the 704 genes found through the exhaustive search were present in the 50-gene Williamson metagene (Supplementary Table S3). We then leveraged a high-performance computing cluster to expand our search to all 1.5 billion possible gene dyads, identifying an additional set of discriminating gene pairs (Table 1; Supplementary Table S4 and Supplementary Fig. S3).

We validated these combinations using another publicly available rhabdomyosarcoma dataset (7). PCA of this second set using the 50-gene Williamson classifier revealed good separation of the samples according to fusion status (Fig. 1B). The top-scoring two- and three-gene combinations identified from the analysis of the Williamson data were tested (Fig. 2B; Table 1 and Supplementary Table S2), and the classification scores ranged from 60.1% to 100%. Of these, 60 pairs yielded a classification of more than 98% in both datasets.

Because we substantially limited the number of class-defining analytes to just two transcripts, we considered whether certain pairs would discriminate fusion status using a lower throughput assay applied to formalin-fixed, paraffin-embedded (FFPE) material. We chose the nCounter assay (NanoString) based on its capacity to accurately quantify rhabdomyosarcoma gene expression in FFPE material in which RNA integrity is substantially compromised (M. Wing and J. Gastier-Foster, unpublished data; Supplementary Fig. S4). We evaluated 16 embryonal and 11 alveolar rhabdomyosarcoma cases (Supplementary Table S5) using 46 genes from the Williamson F1/F2 metagenes (Supplementary Table S6). Of particular note, the histologic diagnoses in these cases were inaccurate as two "embryonal" tumors were fusion positive and three "alveolar" specimens lacked a detectable fusion. Despite low RNA concentration (~30–500 ng/μL) and integrity ($RIN_{avg}$ ~2.3, $r = 0.12$), expression of the 46 signature genes was detected in each case (Supplementary Fig. S5B). PCA of the nCounter data showed excellent separation of the fusion-positive and

**Table 1.** Top-scoring classifiers from Williamson dataset

| Symbol | Symbol | Symbol | Classification Williamson data (9) | Classification Davicioni data (7) | Classification nCounter data | Classification Exon array data |
|---|---|---|---|---|---|---|
| **Exhaustive search of F1/F2 metagene dyads and triads** | | | | | | |
| PGRMC1 | CELA2A/2B | SPRED2 | 99.0 | 95.8 | 92.6 | 100 |
| DAPK1 | CELA2A/2B | SPRED2 | 99.0 | 98.6 | 96.3 | 100 |
| FBN2 | CELA2A/2B | SPRED2 | 99.0 | 97.2 | 92.6 | 100 |
| MFAP2 | CELA2A/2B | SPRED2 | 99.0 | 95.1 | 92.6 | 100 |
| ZIC1 | CELA2A/2B | SPRED2 | 99.0 | 97.2 | 92.6 | 100 |
| CELA2A/2B | ASS1 | RBM13/MAK16 | 99.0 | 97.9 | 88.9 | — |
| CELA2A/2B | ALK | SPRED2 | 99.0 | 97.9 | 88.9 | 100 |
| CELA2A/2B | RBM13/MAK16 | SPRED2 | 99.0 | 95.8 | 92.6 | 98.8 |
| CELA2A/2B | RBM13/MAK16 | RAP1GAP2 | 99.0 | 97.2 | 92.6 | 98.8 |
| CELA2A/2B | SPRED2 | SAE1 | 99.0 | 95.1 | 92.6 | 98.8 |
| CELA2A/2B | SPRED2 | TRPS1 | 99.0 | 95.8 | 92.6 | 98.8 |
| CELA2A/2B | SPRED2 | PTBP2 | 99.0 | 94.4 | 92.6 | 100 |
| CELA2A/2B | SPRED2 | SOBP | 99.0 | 95.8 | 92.6 | 100 |
| CELA2A/2B | SPRED2 | WDYHV1 | 99.0 | 95.1 | — | 98.8 |
| CELA2A/2B | SPRED2 | PGBD5 | 99.0 | 95.8 | 92.6 | 98.8 |
| CELA2A/2B | SPRED2 | ASAP1 | 99.0 | 95.8 | 92.6 | 100 |
| CELA2A/2B | ASAP1 | EIF4EBP1 | 99.0 | 95.8 | 92.6 | 98.8 |
| FBN2 | TFAP2B | — | 98.0 | 97.9 | 88.9 | 97.5 |
| FZD7 | PIPOX | — | 98.0 | 98.6 | 96.3 | 97.5 |
| NELL1 | TFAP2B | — | 98.0 | 97.9 | 88.9 | 98.8 |
| CELA2A/2B | SPRED2 | — | 98.0 | 94.4 | 92.6 | 98.8 |
| CELA2A/2B | SAE1 | — | 98.0 | 94.4 | 92.6 | 98.8 |
| HOXC6 | PIPOX | — | 98.0 | 99.3 | 100 | 100 |
| RBM13/MAK16 | PIPOX | — | 98.0 | 97.9 | 100 | 95.1 |
| SPRED2 | PIPOX | — | 98.0 | 97.9 | 96.3 | 97.5 |
| TFAP2B | TRPS1 | — | 98.0 | 99.3 | 88.9 | 97.5 |
| TFAP2B | PGBD5 | — | 98.0 | 97.9 | 85.2 | 98.8 |
| TRPS1 | PIPOX | — | 98.0 | 98.6 | 96.3 | 100 |
| **Exhaustive search of 1M dyads from top 1,000 genes - selected additional pairs found** | | | | | | |
| TRIL | PIPOX | — | 99.0 | 100 | — | 96.3 |
| NCOA2 | PIPOX | — | 99.0 | 98.6 | — | — |
| HOXB3 | PIPOX | — | 99.0 | 97.9 | — | — |
| NSMAF | ARIIGAP26 | — | 99.0 | 86.7 | — | — |
| RGS7 | NOS1 | — | 99.0 | 77.6 | — | — |
| RGS7 | NOS1 | — | 99.0 | 77.6 | — | — |
| HS6ST1 | NOS1 | — | 99.0 | 72.7 | — | — |
| AMMECR1 | PITPNM3 | — | 99.0 | 62.2 | — | — |
| COL27A1 | PIPOX | — | 99.0 | — | — | — |
| LOC344595 | SYN2 | — | 99.0 | — | — | — |
| B3GALTL | NOS1 | — | 99.0 | — | — | 100 |
| **Exhaustive search of all 1.5B gene combinations - selected additional pairs found** | | | | | | |
| KCND3 | THSD4 | — | 99.0 | 93.7 | — | 98.7 |
| TULP4 | GBP2 | — | 99.0 | 93.0 | — | — |
| SLC25A29 | SLC46A3 | — | 99.0 | — | — | 96.3 |
| PDE4DIP | PEG3 | — | 96.0 | 87.4 | — | 95.1 |
| FAM110B | PEG3 | — | 95.0 | 94.4 | — | 93.8 |
| MYOZ2 | MYLPF | — | 95.0 | 88.8 | — | 90.1 |
| TTN | MYLPF | — | 95.0 | 96.5 | — | 90.1 |
| GPX7 | PEG3 | — | 95.0 | 88.8 | — | 93.8 |
| MYCN | EPAS1 | — | 95.0 | 89.5 | — | 93.8 |

NOTE: Classification, percentage of samples correctly classified. If no classification listed, probe/gene values were not available.

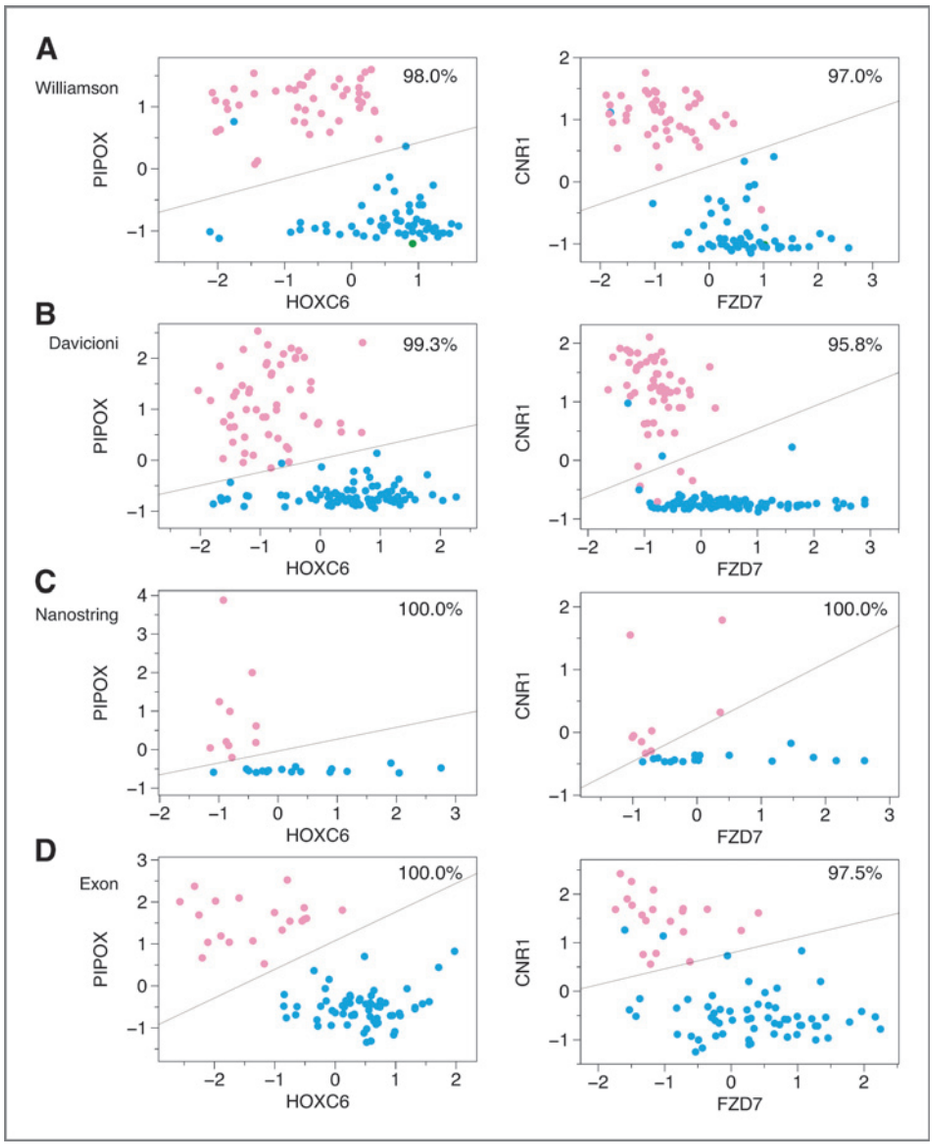


Figure 2. Gene dyads separate fusion-positive from fusion-negative rhabdomyosarcoma. Representative graphs showing normalized expression of *PIPOX* and *HOXC6* (left) and *CNR1* and *FZD7* (right). Individual specimens are known to be *PAX–FOXO1* fusion positive (pink) or negative (blue); one fusion-positive specimen expresses an alternative *PAX3–NCOA1* fusion (green). Clustering is based on microarray expression data from two published datasets (A and B), quantification of expression by nCounter assay using unrelated, FFPE material (C), and exon array expression from a fourth set of unrelated specimens (D). Lines dividing the two clusters were drawn by linear SVC.

fusion-negative samples (Fig. 1C and Supplementary Tables S7 and S8) with 68.3% of the variance accounted for by the first 2 principal components. We tested the best two- and three-gene classifiers from the Williamson signature and found complete separation by fusion status with pairs (*HOXC6*, *PIPOX*) and (*MAK16/RBM13*, *PIPOX*; Fig. 2C; Table 1). We then tested all 1,035 two-gene combinations in the 46-gene nCounter set and found 23 that yielded 100% separation of the samples by fusion status (Supplementary Table S9). In the Williamson dataset, these 23 dyads resulted in classification efficiencies ranging from 88.1% to 98.0%, with the (*HOXC6*, *PIPOX*) and (*MAK16/RBM13*, *PIPOX*) dyads performing best (98% separation of the Williamson samples).

To validate our minimal classifiers in another unrelated dataset, we tested the best dyads and triads in an unpublished set of Affymetrix exon array data from 81 rhabdomyosarcoma samples of known fusion status (J.R. Anderson and T.J. Triche, unpublished data). PCA conducted using the Williamson F1/F2 metagenes confirmed complete separation of the samples (Fig. 1D). As before, the best two- and three-gene classifiers were tested on the exon array data, yielding classification efficiencies of 97.5% to 100% (Fig. 2D; Table 1).

Finally, to show the generalizability of our algorithm to quickly find simple classifiers, we applied it to a previously published lung cancer gene expression dataset (12). We used our algorithm to test all the two-gene combinations in their ability to distinguish early death (<24 months) from long-term survival (>48 months) in patients who received chemotherapy. We found 21 two-gene combinations with classification efficiencies of at least 85% and a maximum efficiency of 88.1% (Table 2 and Supplementary Table S10). In addition, we used the most differentially expressed genes to test for the best three-, four-, and five-gene combinations and found a large number with the ability to distinguish the 2 groups with

**Table 2.** Top gene combinations from lung cancer dataset

| Symbol | Symbol | Symbol | Symbol | Symbol | Classification |
|---|---|---|---|---|---|
| SH3BP2 | MIR1292/NOP56 | KLHL21 | ZNF682 | BZRAP1 | 98.3 |
| SH3BP2 | MIR1292/NOP56 | N/A (216805_at) | EDN1 | ZNF682 | 98.3 |
| SH3BP2 | MIR1292/NOP56 | N/A (216805_at) | ZNF682 | FLJ22184 | 98.3 |
| KLHL21 | PCCA | EDN1 | ZNF682 | FLJ22184 | 98.3 |
| SH3BP2 | MIR1292/NOP56 | COBL | SAE1 | — | 96.6 |
| SH3BP2 | MIR1292/NOP56 | COBL | SYNRG | — | 96.6 |
| SH3BP2 | MIR1292/NOP56 | COBL | CERS4 | — | 96.6 |
| SH3BP2 | MIR1292/NOP56 | SYNRG | BCAM | — | 96.6 |
| SH3BP2 | KLHL21 | RBCK1 | IRF9/RNF31 | — | 96.6 |
| SH3BP2 | MIR1292/NOP56 | IRF9/RNF31 | — | — | 91.5 |
| SH3BP2 | MIR1292/NOP56 | RARA | — | — | 91.5 |
| SH3BP2 | N/A (216805_at) | MY06 | — | — | 91.5 |
| SH3BP2 | RBCK1 | IRF9/RNF31 | — | — | 91.5 |
| SH3BP2 | DDX28 | CD3EAP | — | — | 91.5 |
| SH3BP2 | ETV1 | — | — | — | 88.1 |
| BACE1 | GGT1/GGT2 | — | — | — | 86.8 |
| ZNF813 | GGT1/GGT2 | — | — | — | 86.8 |
| NME1-NME2 | CERS4 | — | — | — | 86.8 |
| SORD | AP1M2 | — | — | — | 86.8 |

classification efficiencies of 91.5%, 96.6%, and 98.3%, respectively (Table 2 and Supplementary Table S10).

## Discussion

Clinical and histologic features used to classify tumors are only surrogates for the underlying biologic pathways driven by the expressed genes or noncoding RNAs. Tumor classification schemes based on clinical and histologic factors are appropriately being supplanted by those derived from gene expression and proteomic studies (15, 16). In rhabdomyosarcoma, stratification by routine histology is being replaced by the detection of the *PAX–FOXO1* fusion transcript (17, 18). Such an approach would foster more personalized therapy if classification were based on molecular studies that robustly predict tumor behavior and clinical outcome.

Our findings indicate that as few as two genes can have high discriminatory power in defining molecular subsets within a type of cancer. We accomplished this by an exhaustive analysis of gene combinations from existing microarray-based gene expression data for rhabdomyosarcoma, and we validated our findings using 3 unrelated datasets, one of which represented substantially degraded RNA extracted from FFPE blocks. This last point highlights the fact that such an approach could be broadly applicable in the clinical setting. Because gene expression quantification using nCounter remains linear over a very large dynamic range (19), the ratio of the normalized expression of two genes could be applied to a single patient. Our data suggest that (*HOXC6*, *PIPOX*) or (*MAK16/RBM13*, *PIPOX*) would be particularly good candidates, and we will test them prospectively. Whether an RNA-based approach would outperform immunohistochemical detection is an obvious question that we have yet to address. Previously, several groups showed that antibody-based detection of a limited number of proteins by immunohistochemistry had the potential to discern rhabdomyosarcoma subtypes (20–22). We predict, though, that the marked specificity and highly quantitative nature of the nCounter assay will foster more robust classification.

We also recognize that our data show correlation between the expression of two or three genes and the expression of the *PAX–FOXO1* fusion, the detection of which is already an accepted clinical test. Yet, although such a two-gene surrogate may not supplant a single analyte test, it may provide insight into tumor biology. For example, that we and others (9) observed the *PAX3-NCOA1*–containing rhabdomyosarcoma clustering with the fusion-negative cases raises the interesting question of whether its biology will more closely mimic fusion-negative tumors. As such, a two-gene surrogate may outperform fusion gene testing as a clinical test, as it would provide additional insight into tumor biology.

We have described several simple classifiers that can distinguish the fusion-positive from the fusion-negative rhabdomyosarcoma samples with very high efficiency, and it seems that these signatures are valid across several disparate datasets. Nevertheless, the problem remains of what to do with the patients whose results are close to the line and the signatures are unable to classify with certainty. This is indeed a problem for gene signatures in general, as very few are capable of classifying samples with 100% accuracy and precision. In some cases, another two-gene signature could be applied, whereas biologic validation may be required in others. Our algorithm facilitates the process of finding a second-line signature, because it has the ability to generate a large number of distinct signatures with high classification efficiencies.

It is possible that either rhabdomyosarcoma as a tumor type or fusion status as a distinguishing characteristic is particularly amenable to this type of analysis, so to show that our approach is generalizable and this is not simply a rare example where our approach worked, we further tested our algorithm on a lung cancer dataset to distinguish survival. Although these gene signatures would need to be validated to prove their clinical usage, we were merely using it here to show that our algorithm can enhance discovery by quickly finding simple classifiers for any dataset.

We have shown that a relatively simple but exhaustive algorithm can be applied to a large set of signature genes to identify a minimal set with great discriminatory power. Although it is tempting to merely ascribe "biomarker" status to these gene pairs, it is possible, in fact likely, that these gene combinations reveal underlying biologic connections that can reveal interconnected pathways important in the generation and maintenance of cancer. By finding small numbers of genes that can be used to classify samples, we provide a method for simplifying complex classifiers, enabling their use for clinical applications. We plan on developing this into open source software that would enable other groups to apply a similar approach to their data. Furthermore, analysis could be offered as a service on a high-performance computing web-based platform for general use. Software and related information will be available at http://exhaustive.msvalidator.org.

## Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

## Authors' Contributions

**Conception and design:** K.M. Bachmeyer, S.X. Skapek, S.L. Volchenboum
**Development of methodology:** R.A. Wilson, L. Teng, K.M. Bachmeyer, S.X. Skapek, S.L. Volchenboum
**Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.):** R.A. Wilson, K.M. Bachmeyer, M.L.Z. Bissonnette, A.N. Husain, D.M. Parham, T.J Triche, M.R. Wing, J.M. Gastier-Foster, F.G. Barr, J.R. Anderson, S.X. Skapek
**Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis):** R.A. Wilson, L. Teng, D.M. Parham, T.J Triche, F.G. Barr, S.X. Skapek, S.L. Volchenboum
**Writing, review, and/or revision of the manuscript:** R.A. Wilson, D.M. Parham, M.R. Wing, J.M. Gastier-Foster, F.G. Barr, D.S. Hawkins, J.R. Anderson, S.X. Skapek, S.L. Volchenboum
**Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases):** R.A. Wilson, K.M. Bachmeyer, T.J Triche, F.G. Barr, J.R. Anderson, S.L. Volchenboum
**Study supervision:** S.L. Volchenboum

## Acknowledgments

The authors thank Michael Wilde (University of Chicago) for generously conducting the Swift programming, Peter Houghton (St. Jude Children's Research Hospital) for kindly providing the cell lines, Children's Oncology Group for providing some rhabdomyosarcoma specimens, and Yang Xie (UT Southwestern Medical Center) for kindly providing the normalized expression data for the patients with lung cancer.

## Grant Support

This work was supported by funding from the St. Baldrick's Foundation (J.R. Anderson, D.S. Hawkins, S.X. Skapek, S.L. Volchenboum), NIH RC2 CA148216 (J.R. Anderson, J.M. Gastier-Foster, D.S. Hawkins, S.X. Skapek, T.J Triche), and the Intramural Research Program of the National Cancer Institute (F.G. Barr).

Received February 4, 2013; revised June 12, 2013; accepted July 9, 2013; published OnlineFirst August 2, 2013.

## References

1. Bridge JA, Cushman-Vokoun AM. Molecular diagnostics of soft tissue tumors. Arch Pathol Lab Med 2011;135:588–601.
2. Huh WW, Skapek SX. Childhood rhabdomyosarcoma: new insight on biology and treatment. Curr Oncol Rep 2010;12:402–10.
3. Meyer WH, Spunt SL. Soft tissue sarcomas of childhood. Cancer Treat Rev 2004;30:269–80.
4. Oberlin O, Rey A, Lyden E, Bisogno G, Stevens MC, Meyer WH, et al. Prognostic factors in metastatic rhabdomyosarcomas: results of a pooled analysis from United States and European cooperative groups. J Clin Oncol 2008;26:2384–9.
5. Malempati S, Hawkins DS. Rhabdomyosarcoma: review of the Children's Oncology Group (COG) Soft-Tissue Sarcoma Committee experience and rationale for current COG studies. Pediatr Blood Cancer 2012;59:5–10.
6. Davicioni E, Anderson JR, Buckley JD, Meyer WH, Triche TJ. Gene expression profiling for survival prediction in pediatric rhabdomyosarcomas: a report from the children's oncology group. J Clin Oncol 2010;28:1240–6.
7. Davicioni E, Finckenstein FG, Shahbazian V, Buckley JD, Triche TJ, Anderson MJ. Identification of a PAX-FKHR gene expression signature that defines molecular classes and determines the prognosis of alveolar rhabdomyosarcomas. Cancer Res 2006;66:6936–46.
8. Barr FG, Smith LM, Lynch JC, Strzelecki D, Parham DM, Qualman SJ, et al. Examination of gene fusion status in archival samples of alveolar rhabdomyosarcoma entered on the Intergroup Rhabdomyosarcoma Study-III trial: a report from the Children's Oncology Group. J Mol Diagn 2006;8:202–8.
9. Williamson D, Missiaglia E, de Reyniès A, Pierron G, Thuille B, Palenzuela G, et al. Fusion gene-negative alveolar rhabdomyosarcoma is clinically and molecularly indistinguishable from embryonal rhabdomyosarcoma. J Clin Oncol 2010;28:2151–8.
10. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12:2825–30.
11. Wilde M, Hategan M, Wozniak JM, Clifford B, Katz DS, Foster I. Swift: a language for distributed parallel scripting. Parallel Computing 2011;37:633–52.
12. Shedden K, Taylor JM, Enkemann SA, Tsao MS, Yeatman TJ, Gerald WL, et al. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nat Med 2008;8:822–7.
13. Wachtel M, Dettling M, Koscielniak E, Stegmaier S, Treuner J, Simon-Klingenstein K, et al. Gene expression signatures identify rhabdomyosarcoma subtypes and detect a novel t(2;2)(q35;p23) translocation fusing PAX3 to NCOA1. Cancer Res 2004;64:5539–45.
14. Sumegi J, Streblow R, Frayer RW, Dal Cin P, Rosenberg A, Meloni-Ehrig A, et al. Recurrent t(2;2) and t(2;8) translocations in rhabdomyosarcoma without the canonical PAX-FOXO1 fuse PAX3 to members of the nuclear receptor transcriptional coactivator family. Genes Chromosomes Cancer 2010;49:224–36.
15. Chin L, Andersen JN, Futreal PA. Cancer genomics: from discovery science to personalized medicine. Nat Med 2011;17:297–303.
16. Verweij J, Baker LH. Future treatment of soft tissue sarcomas will be driven by histological subtype and molecular aberrations. Eur J Cancer 2010;46:863–8.
17. Wexler LH, Ladanyi M. Diagnosing alveolar rhabdomyosarcoma: morphology must be coupled with fusion confirmation. J Clin Oncol 2010;28:2126–8.
18. Missiaglia E, Williamson D, Chisholm J, Wirapati P, Pierron G, Petel F, et al. PAX3/FOXO1 fusion gene status is the key prognostic molecular marker in rhabdomyosarcoma and significantly improves current risk stratification. J Clin Oncol 2012;30:1670–7.

19. Geiss GK, Bumgarner RE, Birditt B, Dahl T, Dowidar N, Dunaway DL, et al. Direct multiplexed measurement of gene expression with color-coded probe pairs. Nat Biotechnol 2008;26:317–25.
20. Heerema-McKenney A, Wijnaendts LC, Pulliam JF, Lopez-Terrada D, McKenney JK, Zhu S, et al. Diffuse myogenin expression by immunohistochemistry is an independent marker of poor survival in pediatric rhabdomyosarcoma: a tissue microarray study of 71 primary tumors including correlation with molecular phenotype. Am J Surg Pathol 2008;32:1513–22.
21. Davicioni E, Anderson MJ, Finckenstein FG, Lynch JC, Qualman SJ, Shimada H, et al. Molecular classification of rhabdomyosarcoma–genotypic and phenotypic determinants of diagnosis: a report from the Children's Oncology Group. Am J Pathol 2009;174:550–64.
22. Wachtel M, Runge T, Leuschner I, Stegmaier S, Koscielniak E, Treuner J, et al. Subtype and prognostic classification of rhabdomyosarcoma by immunohistochemistry. J Clin Oncol 2006;24:816–22.