Published online 20 April 2016

Nucleic Acids Research, 2016, Vol. 44, No. 11 ***e110***
doi: 10.1093/nar/gkw269

A new molecular signature method for prediction of driver cancer pathways from transcriptional data

Dmitry Rykunov, Noam D. Beckmann, Hui Li, Andrew Uzilov, Eric E. Schadt and Boris Reva*

Icahn Institute for Genomics and Multiscale Biology, New York, 10029, USA

Received August 14, 2015; Revised March 28, 2016; Accepted April 02, 2016

ABSTRACT

Assigning cancer patients to the most effective treatments requires an understanding of the molecular basis of their disease. While DNA-based molecular profiling approaches have flourished over the past several years to transform our understanding of driver pathways across a broad range of tumors, a systematic characterization of key driver pathways based on RNA data has not been undertaken. Here we introduce a new approach for predicting the status of driver cancer pathways based on signature functions derived from RNA sequencing data. To identify the driver cancer pathways of interest, we mined DNA variant data from TCGA and nominated driver alterations in seven major cancer pathways in breast, ovarian and colon cancer tumors. The activation status of these driver pathways were then characterized using RNA sequencing data by constructing classification signature functions in training datasets and then testing the accuracy of the signatures in test datasets. The signature functions differentiate well tumors with nominated pathway activation from tumors with no signs of activation: average AUC equals to 0.83. Our results confirm that driver genomic alterations are distinctively displayed at the transcriptional level and that the transcriptional signatures can generally provide an alternative to DNA sequencing methods in detecting specific driver pathways.

INTRODUCTION

To determine an optimal set of drugs for the treatment of a particular cancer patient, the molecular characterization of the activation status of cancer drivers in tumor cells of patients is critical (1). Cancer driver genomic alterations in tumors are most typically identified by sequencing tumor DNA, often comparing the tumor DNA to germline DNA (2). Increasingly, short-read next-generation sequencing (NGS) technologies are being used in both research and the clinic (3) to determine point mutations and small structural variants (deletions and insertions) via targeted sequencing panels as well as whole-exome and whole-genome sequencing (WES and WGS, respectively) assays. From WGS data and to some extent WES data (especially when complemented with genome-wide genotype data), large structural variants such as inversions and translocations may be identified as well. Microarray (4) or WES/WGS (5) data may also be used to determine somatic copy number variation with respect to a normal control specimen from the same individual. Genomic alterations in pathway driver genes are usually taken as primary markers of oncogenic activation (6–10).

Despite the impact DNA sequencing has had on the molecular characterization of the key driver genes and pathways in a given tumor, genomic alterations alone may not provide a complete picture of the activation status of driver cancer pathways. First, genomic DNA NGS typically explores a limited subset of the genome, with targeted panels generally covering <4 megabases and WES assays <65 megabases of an approximately 3 gigabase haploid human genome. Thus, important mutations in non-coding regulatory regions may not be detected, and fine-grained mapping of structural variation may not be possible. While WGS sequencing may cover the vast majority of the genome, cost considerations and reimbursement are still dominant factors preventing routine adoption, and beyond this, identifying and interpreting variants in regulatory regions are far from solved problems, making interpretation of non-coding variants very difficult. In addition, WGS still only effectively covers roughly 85% of the genome (short read data cannot effectively cover heterochromatin DNA comprising >10% of the genome and complex repeat regions and other such structural variants cannot be routinely assayed) (11). Second, pathway activation can be achieved by alterations in the expression of cancer driver genes that are not detectable by DNA sequencing. For example, perturbations in the signaling pathways of such genes, epigenetic changes or even protein state changes that feedback onto transcription can all result in changes in gene expression that would not be detected via today's standard WES/WGS assays. Third, our ability to accurately interpret primary genomic alterations is far from perfect. Mutations that do not change the amino acid sequence (e.g. synonymous, intronic, and non-

*To whom correspondence should be addressed: Tel: +1 212 824 9663; Fax: +1 212 824 9699; Email: boris.reva@mssm.edu

coding mutations) may have an effect on gene expression, but interpreting them is far more challenging than codon-changing mutations. For copy-number variation, the extent to which large chromosomal gains or losses (which could include tens or hundreds of genes) impact the expression of specific pathway-relevant genes in the affected chromosome segment is often unclear.

Therefore, it is both interesting and practically important to explore alternative, complementary approaches to determine driver cancer pathways. Gene expression data are of high interest for the prediction of activated pathways and driver genes (12–23) and drug response (1,13,24–26), given RNA reflects real time state information in the tumor.

The hypothesis underlying the relevance of RNA for identifying cancer driver genes and pathways is that activating genomic alterations in major cancer driver pathways are distinctively displayed at the transcriptional level. A typical tumor may have activated several driver cancer pathways. Therefore the direct interrogation of pathway activation in mouse models or cell lines using a limited number of pathway specific drugs may not be efficient to determine optimal treatment paths for any given combination of pathways. Further, the systematic exploration of all possible pathway combinations will be practically limited due to the large number of combinations that obtain as the number of relevant pathways grows (the growth can be exponential) as well as the cost of the experiments. Thus, currently, the accurate analysis of activated cancer pathways by experimental approaches cannot be scaled up to the population level. Developing computational approaches that can reduce the number of cancer driver hypotheses to consider for any particular tumor as a way of guiding subsequent experiments, validations and treatment paths is critical.

Nomination of activated pathways can be done using GSEA (15) or more complicated SPIA (16) methods. However, these methods are not practically accurate (18), in particular, because they do not take into account genomics information. Rapid accumulation of genomic and gene expression data motivated development of new approaches to study mutation impact on gene expression in cancer (18–23). These approaches are classified (17,23) by specifics tasks, computational methods and data used, and by applicability for interpretation of individual tumor profiles. All reviewed approaches (MOCA (19), CONEXIC (20), EPoc (21), DriverNet (22), xseq (23)) are aimed on prediction driver genes or driver pathways (PARADIGM (18)). All approaches except of MOCA (19), use gene interaction networks or pathway information. However, neither of the approaches is specifically developed for prediction of driver genomic alterations in cancer pathways from an individual gene expression profile.

The development of computational approaches has its own set of difficulties and uncertainties. First of all, tumors have to be categorized by driver alterations in specific pathways. This is not trivial both practically and biologically, given the concept of 'activated pathway' is a useful simplification of the coherent activity of certain components of the cell's more complex gene interaction network. In the general case, there is no guarantee that any combination of genomic alterations will lead to distinct phenotypic signals. However, major driver alterations in cancer have a strong tendency to be mutually exclusive (6–10), which supports the hypothesis of distinct transcriptomic phenotypes associated with major types of driver alterations in cancer.

Activation of a particular oncogenic pathway in a particular type of cancer can be associated with a number of moderate or weak changes in gene expression levels, neither of which is a strong predictive marker on its own. Differently from the above-reviewed methods (18–23), we introduce a new statistical approach for the recognition of altered cancer pathways using a 'molecular signature'—a weighted sum of gene expression levels, where the biomarker genes and their weights are determined from a training set. Thus, formally we solve the same task as PARADIGM (18), however we do not limit a search for a pathway marker genes to a particular set of pathway genes as does PARADIGM (18). Our method takes into account all genes and determines those, which expression levels are associated with the activated status of a given pathway determined from both mutation and gene copy numbers alterations. The method does not use the gene network inference models (18,20,23) which are limited by not always accurately known conditional dependencies, and it does not predict the specific driver roles of individual genes (18,20–23). We compared our approach to ten well-established classification methods (27–36) and found that our method outperformed all of them.

In our study, determining genes-biomarkers and transcriptional signatures of specific cancer pathways serves two goals: first, to provide an alternative and complementary approach to DNA sequence based approaches to nominate driver pathways in a given tumor; and, second, to expand our knowledge by identifying new genes that may act as potential participants in cancer pathways. Such genes can point to associated cellular processes as well as new therapeutic targets.

MATERIALS AND METHODS

The computational protocol we developed to derive predictive cancer pathway signatures consists of two parts: (i) a statistical algorithm for determining candidate gene expression biomarkers that are associated with pathway-specific genomic alterations; (ii) a machine learning algorithm for determining the optimal weights of combinations of candidate biomarkers in order to derive scoring functions—a signature for predicting key driver alterations in major cancer pathways.

Statistical framework for identifying candidate biomarkers

The algorithm used to identify candidate biomarkers is depicted in Figure 1. To map genomic alterations on cancer driver pathways, we started with the original gene-based profiles of mutations and DNA copy numbers and combined them into one profile of genomic alterations (Figure 1A). In the combined profile, homozygous gene deletions and genes affected by predicted functional mutations (37) were considered as 'inactivated'. Genes affected by amplification and known oncogenes affected by predicted functional mutations were considered as 'activated'; other genes were considered as 'normal'. For example, PTEN was considered as a driver gene if both copies of the gene were

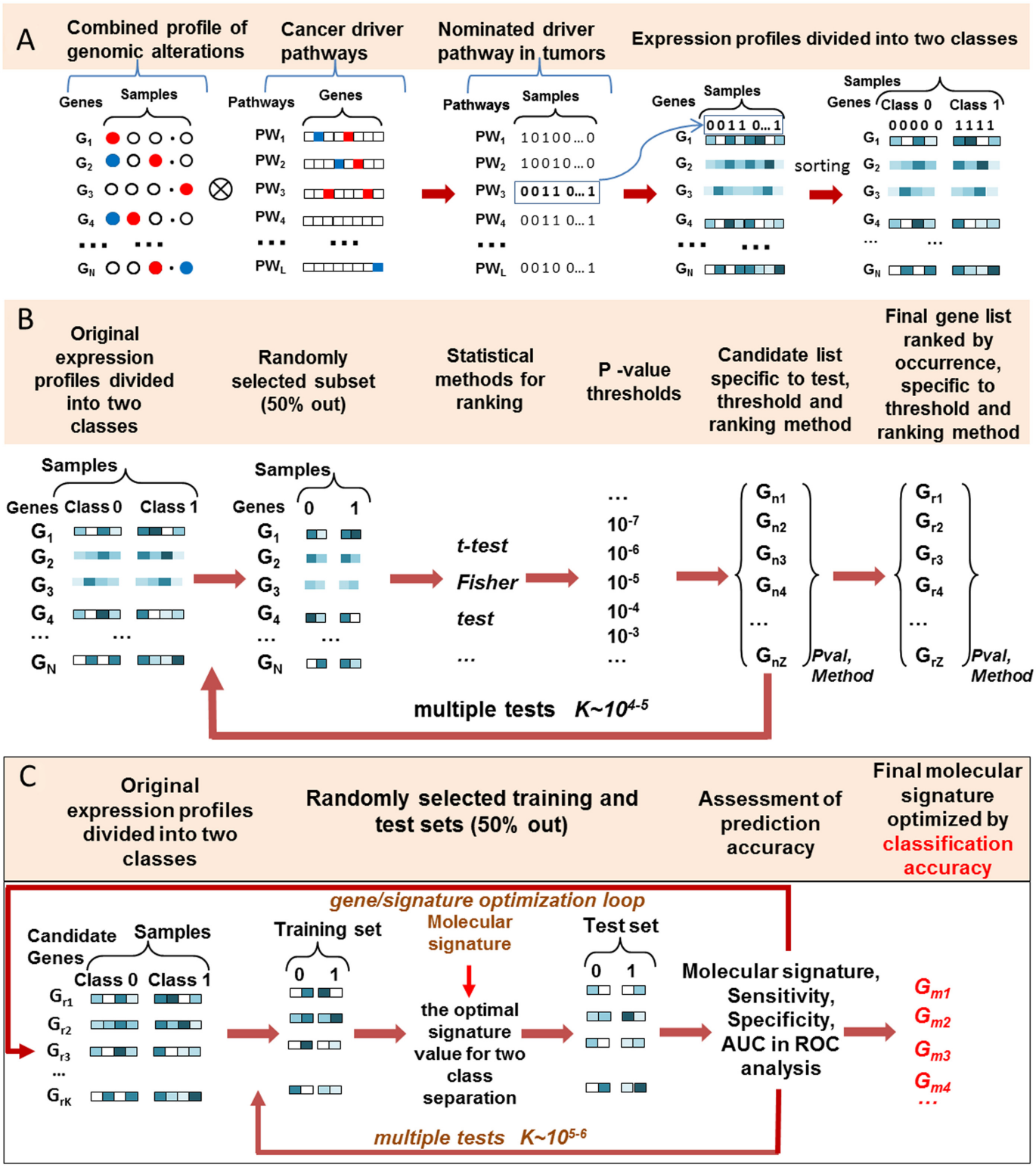


Figure 1. The general schema of the computational protocol. (**A**) Subdivision of tumor molecular profiles into classes based on nominated driver alterations. Tumors with nominated driver alterations in the key genes of a given pathway form a class of tumors with activated pathway (class 1); tumors with no alterations in pathways genes form a class of no pathway activation (class 0). (**B**) Ranking potential biomarkers by significance and occurrence. Tumors are divided by random into two sets of approximately equal sizes. For each of genes, the statistical significance (*P*-value) of association between expression values and tumor classes is determined. Genes frequently selected at the lowest *P*-value levels are the primary candidates for biomarkers. (**C**) Testing classification accuracy of molecular signatures. For each set of candidate signatures (built for 2, 3, 4,… genes) an original set of expression profiles is separated by random into training and test classes of approximately equal sizes. Based on a particular combination of tumors of classes 0 and 1, the molecular signature is computed for training set. Then, the tumors of the test set are classified into two classes based on the optimal separation threshold derived from the training set.

deleted or if it was affected by a functional mutation; EGFR was considered as a driver gene if it was amplified or affected by a functional mutation. Tumors with nominated driver alterations in key genes of a given pathway formed the class of activated tumors with respect to that pathway (class 1), while tumors with no alterations in genes of a given pathway formed the class of non-activated tumors (class 0) with respect to the pathway.

Given an original set of expression data separated into two classes, 0 and 1, tumors were divided by random into two sets of approximately equal sizes, a training set and a testing set (Figure 1B). Then, for each gene a statistical significance (*P*-value) of association between expression values and tumor classes was determined. Genes were then classified into one of several different bins based on *P*-value thresholds ($P = 10^{-14}$, 10^{-13},…,10^{-3}). By repeating this procedure multiple times, potential biomarkers were ranked by both significance and occurrence. Multiple tests based on randomly choosing half of the original dataset were run to reduce the effect of outliers and unevenness in the distributions. Selected candidate genes were also filtered by pairwise correlation thresholds ($r = 0.5, 0.6,\ldots, 0.999$) to ensure at least part of the variance component for each gene contributed independently. Genes frequently selected at the lowest *P*-value thresholds were the primary candidates for biomarkers.

Finally, we assessed the classification accuracy of the selected biomarkers (Figure 1C). For each random splitting of the dataset into training and test sets, signature weights (Equation 2) were computed for 2, 3, 4,… combinations of candidate biomarkers sorted by occurrence at each statistical level determined by the given *P*-value thresholds (Figure 1B). The tumors of a training set were classified based on the values of the resulting molecular signatures (Equation 1). The classification accuracy (38) was assessed by the area under a receiver operating characteristic curve (AUC). To take into account the effect of unbalanced data, we computed the 'balanced accuracy' (BAC), which is defined in (39) (as the arithmetic mean of sensitivity and specificity: BAC=1/2*(Sensitivity + Specificity). The balanced accuracy is used to avoid inflated performance estimates on imbalanced datasets. If the classifier performs equally well on either class, the balanced accuracy reduces to the conventional accuracy (i.e., the number of correct predictions divided by the total number of predictions). If the conventional accuracy is above chance only because the classifier takes advantage of an imbalanced test set, then the balanced accuracy, as appropriate, will drop to chance (39). The threshold giving rise to the maximal value of the balanced accuracy was then determined. The same signature weights and threshold determined from the training set were used to classify tumors in the test set. The optimal set of genes-biomarkers was defined to maximize the value of the average AUC over all test sets. We also computed the probability for each tumor sample falling into class 0 or 1, and averaged the signature gene weights over all of the training sets. For the final assessment of the classification accuracy we took results produced for test sets with averaged signature weights.

We explored two methods for assessing statistical significance in selecting genes as candidate biomarkers: a traditional Student's *t*-test applied assuming continuously distributed gene expression values (40) and the Fisher's test (41) applied to discrete approximations of expression values (Supplementary Figure S1). In this approximation, all expression values that were smaller than the mean expression were classified as downregulated and those values that were larger than the mean value were classified as up regulated (Supplementary Figure S1). We rationalized that the coarse two-level discrete approximation would result in the selection of more robust biomarkers, given it imposes a stringent restriction on statistical associations between gene expression values and tumor classes, reducing the detection of associations due to outliers, errors and unevenness in the distributions. Given the more stringent selection criteria and the loss of power that comes from discretizing the data, across all tests performed, the Fisher's test applied to discrete data produced significantly fewer biomarker candidates as compared to the Student's *t*-test. Both methods generally identified the same top candidate genes. Based on the average prediction accuracies obtained with candidate biomarker genes, neither of the methods was found to provide for a systematic advantage. In practice, we tested gene sets obtained by both methods and selected those that produced the highest accuracy in multiple testing.

A machine learning algorithm for constructing classification signature of gene expression

The molecular signature method is based on a theoretical physical model (42–44) developed to reduce energy errors in the calculation of native folds in protein chains. According to the theory, the energy gap between the native (minimal energy) fold and the average energy of competing folds measured in units of the standard deviation of energies (Z-score) can be dramatically increased (by absolute value) by averaging energies of natural homologs of a protein chain (42–44). The increase in the normalized energy gap (Z-score) is a consequence of the linear dependence of energy differences among a number of homologs and the square root dependence of the standard deviation of energies across a number of homologs. Those homologs, which energies are the least correlated across all competing folds, yield the largest contributions of the increase in the Z-score. The theory leads to an analytical method involving computation weights of individual homologs in the total energy function that optimizes the normalized energy gap.

In our application, the individual candidate biomarkers, which have the potential to discriminate between two classes, play the role of the protein homologs. The gene expression levels of the candidate biomarkers correspond to the energy levels of the homologs. By defining a signature function as the average of the expression levels of candidate biomarkers using the analytically determined optimal weights, the gap (Z-score) between the given data classes can be increased, which in turn results in an improvement in class recognition.

Thus, following this general approach (42–44), we introduce the signature function V_s computed for a tumor profile *s*:

$$V_s = \sum\nolimits_m A_m E_{ms}, \quad (1)$$

where E_{ms} is the expression of biomarker m in tumor s, and A_m is a weight for biomarker m, where, as shown in (43), these weights are computed as:

$$A_m \sim D_m^{-1} \sum\nolimits_k [C_{mk}]^{-1} Z_k, \quad (2)$$

with

$$C_{mk} = \frac{1}{S} \frac{\sum_s (E_{ms} - \langle E_m \rangle)(E_{ks} - \langle E_k \rangle)}{D_m D_k} \quad (3)$$

representing the pairwise correlation between biomarkers m and k; $[C_{mk}]^{-1}$ an element of the inverse matrix; $\langle E \rangle_m$ and D_m the average expression and standard deviation, respectively, of the expression for candidate biomarker m; S the total number of tumors in a data set. In Equation 2 we also use z-scores, Z_k, defined as $Z_k = \frac{\langle E_k \rangle_1 - \langle E_k \rangle_2}{D_k}$, where $\langle E_k \rangle_1$ and $\langle E_k \rangle_2$ are the average expression levels for biomarker k computed for data classes 1 (non-altered pathways) and 2 (altered pathways), respectively.

According to the theory (43), the optimal z-score, Z_M, for the signature V_s composed of M biomarker genes, gets the absolute value of $Z_M = \sqrt{\sum_m \sum_k Z_m [C_{mk}]^{-1} Z_k}$, $(m, k = 1, 2 \ldots M)$. When gene expressions are independent, $[C_{mk}]^{-1} = \delta_{mk}$ and $Z_M = \sqrt{\sum_m Z_m^2} = \sqrt{\langle Z^2 \rangle} \sqrt{M}$, where $\sqrt{\langle Z^2 \rangle} = \sqrt{\sum_m Z_m^2 / M}$ is a root mean square of z-scores ($\delta_{mk} = 1,\ if\ m = k$; $\delta_{mk} = 0,\ if\ m \neq k$).

Thus, two-class separation can be significantly improved by using the weighted sum of non-correlated gene expression traits. In practice, to determine an optimal list of biomarkers, we need to take into account the significance of the expression differences of the biomarker between the classes, the frequency with which the biomarker is found to discriminate between the classes across all subsets considered, and the pairwise correlations among the biomarkers.

Formally, two parts of the above approach are independent. The molecular signature algorithm can be used with any given set of genes, gene-based tumor profiling data and tumor classes, while genes biomarkers obtained by execution of the protocol of the '*statistical framework*' can be used with different classification algorithms (as was done in multiple method testing presented in Table 2). However, both computational protocols utilize two common ideas: (i) multiple random data sampling and (ii) ranking biomarkers and classification signatures by averaging results of multiple tests.

RESULTS

Identification of DNA-based biomarkers for use in training pathway classifiers

Here we used the original genomic data produced by The Cancer Genomic Atlas (TCGA) project as well as the classifications of seven major cancer driver pathways presented in marker publications from the TCGA research network (7,8,10): (i) RTK/RAS/MEK, (ii) cell cycle, (iii) PI3K, (iv) DNA repair, (v) WNT, (vi) NOTCH and (vii) the TGFB pathways (Table 1). Based on genomic alterations (mutations and gene copy number variations) in key genes of each of the major cancer pathways, TCGA tumors of one cancer type were divided into two classes: tumors with likely activating alterations in a given cancer pathway and tumors with no alterations in genes known or predicted to be a member of this pathway. The pathway genes identified for this study are given in Table 1. We note that one activating (or inactivating) alteration (e.g. predicted functional mutation (37), homozygous deletion or amplification) was taken as sufficient evidence for pathway activation. Given the tumor-specific driver pathways determined in this way, we annotated the transcriptional profiles available from TCGA for these tumors with respect to these pathways. In this way, the task of determining which of the cancer pathways is activated in a given tumor is reduced to a classification task: given the transcriptional profile of a given tumor along with the transcriptional profiles of two sets of tumors in which the pathways are known to be activated and not activated, determine the class to which the given tumor is most similar.

Table 1. Genes of major cancer pathways used in the study

RTK/RAS/MEK	Cell Cycle	PI3K	DNA-repair	WNT	NOTCH	TGFB
EGFR	CCND1	PIK3CA	BRCA2	CTNNB1	JAG1	ACVR1B
ERBB2	CCNE1	AKT1	BRCA1	FZD10	JAG2	ACVR2A
ERBB3	CDK4	MTOR		APC	MAML1	ARID1A
FGFR1	CDK6	PTEN		AXIN2	MAML2	SMAD2
FGFR2	E2F3	STK11		DKK1	MAML3	SMAD3
FGFR3	AURKA	TSC1		DKK2	NOTCH3	SMAD4
PDGFRA	CDKN1B	TSC2		DKK3		TGFBR1
BRAF	CDKN2A	PIK3R1		DKK4		TGFBR2
HRAS	RB1			FAM123B		
KRAS				FBXW7		
NRAS				LRP5		
MAP2K1				TCF7L2		
MAP2K2				SOX9		
NF1						

The pathways and genes are annotated in *(6-9)*; genes activated by predicted functional mutations *(37)* or copy number amplifications are shown in red; genes inactivated by mutations or homozygous deletion are shown in blue.

RNAseq signatures for predicting key transcription-based driver alterations in breast cancer

The algorithm used to identify candidate biomarkers is depicted in Figure 1. In the first ('discovery') phase of our computational protocol we identified candidate RNA-based biomarkers by optimizing the area under the receiver-operating-characteristic curve (AUC (38)) for two-class separation between tumors with altered and non-altered key driver pathways identified in the TCGA data (Figure 1A). For candidate biomarkers identified at each of the significance thresholds considered, we further filtered the biomarkers based on pairwise correlation thresholds to ensure each gene selected provided some degree of independent information. The selected genes were then ranked by occurrence across the 1024 tests carried out (Figure 1B; Methods). In the second ('classifier construction') phase of our computational protocol, for each set of ranked genes, from a minimum of 2 to a maximum of 35 genes (when available), we generated 4096 random samples, splitting the dataset for each sample into training and test sets, constructed the classifier using the candidate biomarkers identified in phase 1 on the training set, and then assessed the accuracy of the classifier on the test set (Figure 1C). In the training component of these runs all gene combinations of 2, 3, …, on up to 35 genes that satisfied the imposed filters, were allowed to compete for inclusion in the classifier, and then in the testing component for each run only the best fitting set of ranked genes were used.

For the seven classifiers corresponding to the seven indicated pathways in breast cancer, the average AUCs ranged from 0.72 for the PI3K signaling pathway to 0.86 for the WNT and NOTCH pathways. The full results are provided in Table 2, but a representative example of the predictions is given for the cell-cycle pathway in Figure 2. As the number of genes incorporated into a classifier increases, the prediction accuracy steadily increases in the training sets (Figure 2A). However, in the test sets the maximum of the AUC is observed for four genes, demonstrating the necessity of the training/testing protocol to protect against overfitting. The biomarker genes, their 'signature weights' (averaged over 4096 tests) and brief UniProt (45) annotations are provided in Figure 2B for the top-performing classifier. These biomarker genes were selected at a *P*-value threshold of 10^{-6} using the Fisher's Exact test with appropriately discretized expression levels. The weighted sum of the RPKM gene expression values for the identified biomarkers results in a signature that can predict cell cycle activation status in breast cancer with an overall equally weighted accuracy of ∼77% (Figure 2C and D). From the TCGA data, we determined that two of the four biomarker genes, CTTN and NCAPD2, are statistically significantly amplified in breast and ovarian cancers, and that a third biomarker gene, ORAOV1, is significantly amplified in head and neck (HNSC) cancers. It should be noted that for ∼25% of the high-scoring tumors the false-positive rate is less than 5% (Figure 2C). Interestingly, the majority of tumors are classified in one of two pathway activation classes with probabilities that are close to either 0 or 1. There is a sharp transition between predicted classes, when a value of the signature function crosses the value of the equally weighted accuracy threshold (Figure 2D). However, the cumulative percentages of tumors that have (or do not have) driver alterations in cell cycle pathway are rather smooth (Figure 2D).

Predicting key transcription-based driver alterations in colon and ovarian cancer

Transcription-based classifiers were constructed for the seven pathways indicated in Table 1 in colon and ovarian cancer using the same procedure described above for breast cancer. The average AUCs ranged from 0.73 for the RTK pathway to 0.85 in the TGFB pathway in ovarian cancer, and from 0.83 in the Cell Cycle to 0.89 in the PI3K and WNT signaling pathways for colon cancer. The full results are summarized in Table 2 and Supplementary Table S4.

The ROC curves and individual tumor probabilities for representative examples in colon and ovarian cancer are provided in Figure 3 for the NOTCH (Figure 3A and B) and RTK (Figure 3C and D) pathways. The driver alterations in the NOTCH and RTK pathways can be correctly predicted for ∼50% of the high-scoring tumors with an accuracy of ∼90%. Similar to the individual probabilities of cell cycle activation in breast cancer (Figure 2D), the majority of ovarian and colon cancer tumors are classified (correctly or incorrectly) into one of two pathway activation classes with probabilities close to 0 or 1. The transition between predicted classes is again sharp, with the probability of class prediction depending non-linearly on deviation from the point of equally balanced accuracies (Figure 3B and D), while the cumulative distributions are smooth. Based on the value computed from these signature functions, one can assess both the probability of pathway activation and the corresponding error.

Characterizing the prediction tests for breast, ovarian and colon cancers

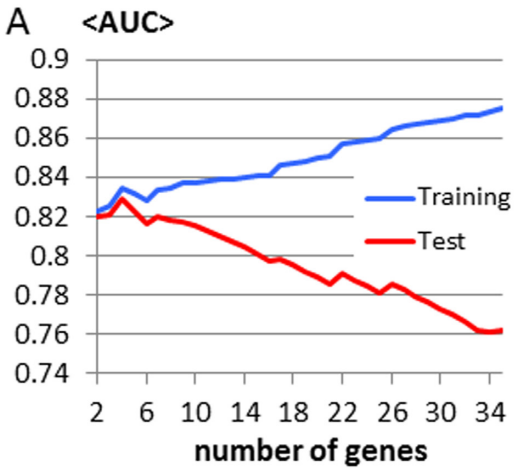
We transformed the RNA RPKM count data in two ways to assess the sensitivities of the classifiers based on the processing of the input data: (i) mean-centered and variance scaled z-score (40), and (ii) discrete approximations of expression values. The significance of the predictions was evaluated using the Fisher's Exact test (41). We found that both approaches gave rise to essentially equivalent results. For control and comparison, we applied ten other popular machine learning methods (27–36) using the same biomarkers and the same 2-fold validation protocol repeated 4096 times; the computations were performed with WEKA software (46). The results of the comparisons are reported in Table 2 and a Supplementary Table S4. In all tests conducted, the molecular signature method outperformed all other classification methods tested.

We further studied the class predictions by averaging the signature functions, as opposed to considering individual signature values. Because the gene expression levels, once computed, are fixed, averaging the signature functions is equivalent to averaging the signature gene weights. In testing the prediction accuracies with averaged signatures, it is critical that tumor expression profiles used for testing not have been used to compute the signature weights (training). This condition was easily satisfied in the general setting of our computational tests given the training sets were used to estimate the weights (Figure 1); that is, we simply summed and averaged the signature functions computed for a given tumor sample in all testing sets and used these averaged signatures for the class predictions. The accuracy characteristics produced with the averaged signatures were always higher than the average accuracy characteristics produced for individual signatures (Table 2).

Biomarker genes are associated with cancer-related processes

Because the expression levels of biomarker genes are associated with the activation of specific cancer pathways, we explored whether any of the biomarker genes were directly involved in carcinogenesis. To identify biomarker genes of interest, we rank ordered genes according to their (i) representation on the list of cancer pathway genes (Table 1), (ii) interaction with cancer pathway genes (47,48), (iii) exhibiting signs of positive selection in cancer ((49), Supplement S2, Supplementary Tables S1, S2), (iv) known role in cancer (50) and (v) functional annotations from UniProt (45) and NCBI (51). All of these data as well as the signature weights and gene expression fold changes are presented in Supplementary Table S1.

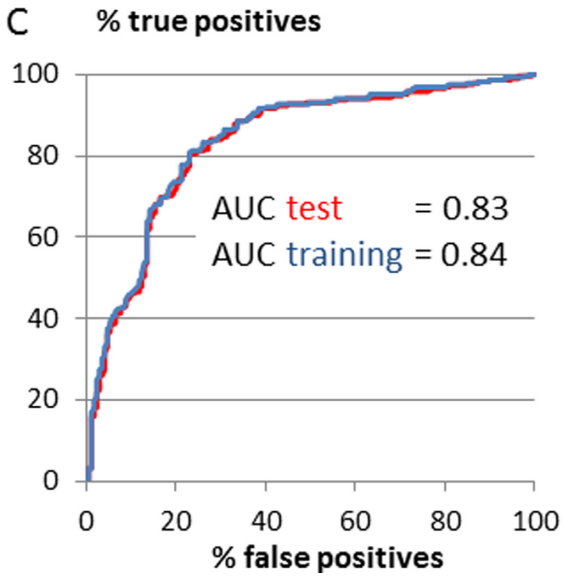
We found that among the 385 biomarker genes considered, 35 are known cancer genes and 6 are genes within the pathway being classified (e.g. ERBB2, PTEN, CDKN2A), with one of these genes a known cancer marker (ORAV01). About 20% of the 385 biomarker genes (74) interact with the



B **Cell Cycle pathway biomarkers for breast cancer**

Gene	Occurrence	Fold Change	Signature Weights	Interactors of Cell Cycle pathway
ORAOV1	0.990	1.7	-0.424	
CDT1	0.990	2.0	-0.862	CCNE1, CDK4
CTTN	0.990	1.5	-0.028	AURKA
NCAPD2	0.988	1.5	-0.167	AURKA

ORAOV1 Oral cancer-overexpressed protein 1
CDT1 DNA replication factor. Potential oncogene
CTTN Plays a role in the invasiveness of cancer cells, and the formation of metastases
NCAPD2 Regulatory subunit of the condensin complex



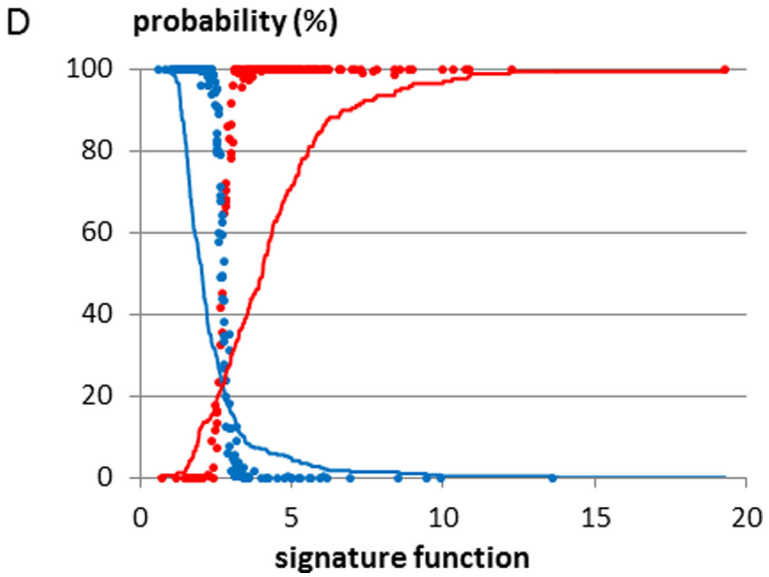


Figure 2. (**A**) Determining the optimal set of expression-based biomarkers. The classification accuracy is assessed by the area under a receiver operating characteristic curve (AUC). The optimal set of biomarkers is defined as the set maximizing the average AUC computed for classification of the test sets. The maximum average AUC value was observed for four genes. (**B**) Characteristics of the top biomarker genes. The obtained optimal biomarkers are either directly annotated as related to cancer or interact with key genes of cell cycle pathway. (**C**) The receiver operating characteristic (ROC) curves built from the classification of tumors in the training and test sets using the four biomarkers with the signature weights averaged over 4096 randomized tests. (**D**) Classifications of individual tumors: the test set derived probability that a tumor with (or without) driver alterations in the cell cycle pathway falls into the correct class are shown by red (blue) dots (each dot represents a tumor). The solid lines represent the cumulative percentage of tumors that have (red) or do not have (blue) driver alterations in the cell cycle pathway. The intersection of the solid lines defines the point where percentages of false positives in both classes are equal that correspond to equally weighted accuracy of ∼77%.

pathway genes, suggesting a likely role in cancer. We also assessed a percentage of known cancer genes (50) and genes that exhibit signs of positive selection among biomarker genes. To this end, we used the Fisher's Exact test to determine genes enriched with genomic alterations: predicted functional mutations (37) and truncating mutations, and also DNA amplifications or homozygous deletions (Supplementary Table S3 and Supplement S2.) We took into account only those genes where enrichments of genomic alterations were observed in at least two of eleven considered TCGA cancers. Out of the 6678 genes that are either known as cancer genes (50) or exhibit signs of positive selection, 167 fall in the set of 385 biomarker genes we identified, a ∼1.3-fold enrichment over what would be expected by chance ($P = 6.9 \cdot 10^{-5}$; Supplementary Tables S1 and S2). The similar analysis conducted on 291 'high confidence' driver genes reported in TCGA pan-cancer study (49) resulted in 15 common genes that would be expected by chance with ($P = 10^{-3}$).

We carried out gene set enrichment analysis using the Enrichr tool (52) and found that biomarker genes were significantly over represented among hub nodes in protein–protein interaction networks. Out of the 26 hub nodes in our set of biomarker genes, the overwhelming majority were well known in cancer, such as YWHAB, MAPK14, ESR1, EGFR and MDM2 (see Supplementary Table S3 for the full list). The biomarker genes were also enriched in the molecular function categories 'regulators of protein

Table 2. The averaged values of the AUC obtained in RNAseq based prediction of driver alterations in cancer pathwaysa

Cancer:	*Breast*							*Colon*				
Pathway:	Cell Cycle	RTK	PI3K	TGFB	WNT	NOTCH	BRCA1/ BRCA2	Cell Cycle	RTK	PI3K	TGFB	WNT
# tumors with altered PW	169	159	230	28	30	28	44	24	110	30	44	153
# tumors with non-altered PW	288	298	227	429	427	429	413	152	66	146	132	23
[b]MSM <Sgn>	**0.83**	**0.79**	**0.72**	**0.86**	**0.86**	**0.80**	**0.85**	**0.83**	**0.87**	**0.89**	**0.85**	**0.89**
[c]MSM	**0.83**	**0.77**	**0.72**	**0.84**	**0.85**	**0.79**	**0.84**	**0.83**	**0.85**	**0.88**	**0.84**	**0.88**
[e]Naïve Bayes (33)	0.70	0.64	0.65	0.72	0.78	0.71	0.75	0.81	0.79	0.86	0.78	0.78
SVM(34)	0.74	0.67	0.66	0.54	0.52	0.55	0.54	0.71	0.77	0.70	0.72	0.72
KNN (27)	0.75	0.65	0.64	0.50	0.51	0.51	0.53	0.65	0.77	0.78	0.70	0.76
Logit Boost (32)	0.74	0.69	0.62	0.53	0.54	0.53	0.55	0.65	0.70	0.62	0.67	0.60
Ada Boost (36)	0.73	0.68	0.63	0.52	0.53	0.53	0.54	0.65	0.71	0.62	0.67	0.58
Random Forest (28)	0.76	0.69	0.65	0.51	0.51	0.50	0.53	0.60	0.74	0.60	0.68	0.56
Class via Regression (31)	0.75	0.66	0.62	0.51	0.51	0.51	0.53	0.61	0.70	0.64	0.67	0.60
Decision Tree (35)	0.74	0.67	0.59	0.53	0.53	0.54	0.54	0.62	0.66	0.62	0.63	0.58
Ripper (29)	0.75	0.67	0.61	0.51	0.51	0.52	0.54	0.60	0.66	0.57	0.64	0.55
Rep Tree (30)	0.73	0.66	0.60	0.50	0.50	0.50	0.51	0.52	0.63	0.52	0.58	0.51

Cancer:	*Ovarian*						
Pathway:	Cell Cycle	RTK	PI3K	TGFB	WNT	NOTCH	BRCA1/ BRCA2
# tumors with altered PW	174	159	157	23	39	115	89
# tumors with non-altered PW	167	182	184	318	302	226	252
[b]MSM <Sgn>	**0.83**	**0.73**	**0.79**	**0.85**	**0.82**	**0.81**	**0.84**
[c]MSM	**0.80**	**0.72**	**0.75**	**0.81**	**0.79**	**0.80**	**0.82**
[e]Naïve Bayes (33)	0.74	0.66	0.69	0.66	0.73	0.67	0.75
SVM(34)	0.73	0.67	0.69	0.55	0.70	0.55	0.69
KNN (27)	0.69	0.62	0.68	0.55	0.65	0.51	0.67
Logit Boost (32)	0.66	0.62	0.64	0.56	0.65	0.58	0.60
Ada Boost (36)	0.65	0.62	0.64	0.54	0.64	0.59	0.55
Random Forest (28)	0.72	0.64	0.69	0.51	0.67	0.54	0.58
Class via Regression (31)	0.65	0.64	0.64	0.52	0.66	0.54	0.59
Decision Tree (35)	0.61	0.58	0.61	0.55	0.62	0.57	0.58
Ripper (29)	0.61	0.58	0.62	0.52	0.62	0.55	0.55
Rep Tree (30)	0.61	0.58	0.61	0.51	0.59	0.51	0.52

[a]The values of the area under curve (AUC) in receiver operating characteristic analysis (38) are averaged over 4096 randomized tests, where one half of tumor samples was used as a training set and another half as tumors was used as test set. [b,c]The AUC values for Molecular Signature Method are presented for both predictions based on averaged signature function (MSM<Sgn>) and for predictions based on individual signatures (MSM). Note that the characteristics of class recognition computed for averaged signatures are systematically higher than the average characteristics of class recognition obtained for individual signatures. [e]For control and comparison, ten classification methods were applied for recognition of tumor classes using the same sets of candidate genes and the same validation protocol (the references on the methods used are given in brackets). Only results obtained for test sets are reported.

serine/threonine kinase activity' and 'receptor protein tyrosine kinase activity' (Supplementary Table S3). We found significant enrichments of the biomarker genes as well in signature gene sets from functional genomics datasets in the Gene Expression Omnibus (GEO) repository (53,54). Of particular note were 271 drug perturbation signatures that were enriched for the cancer biomarker genes (Supplementary Table S3).

We note that only eight genes (ABAT, ATP11B, FBXO2, FXR1, POFUT1, PPP1R3D, SERINC2, ZNF415) were selected twice as biomarkers for multiple cancer pathways.

Class recognition tests on randomly generated data

We also compared accuracies of recognition of nominated cancer pathways with recognition accuracies obtained on randomly generated datasets. Toward this end, the original data were split randomly into two classes to represent (by size) the largest activated pathway. The tests were performed ten times for each cancer type. The results of these tests are reported in Supplementary Table S5. The average AUC values obtained in 4096 test on randomized data were equal to 0.71, and the average number of genes used in the classification signatures was ∼11.6. The corresponding values obtained in recognition of the cancer pathway were 0.84 and 20.3, respectively (Supplementary Table S5). Thus, based on the obtained results, we can conclude that separation of tumors based on activation of cancer pathways is transcriptionally more distinct, compared to randomly generated separations.

We found it biologically reasonable that a number of genes associated with the separation of tumors based on genomic markers of pathway activation is larger than the number of genes identified from the random separations. A larger set of genes-biomarkers improves the accuracy of class recognition between biologically distinct tumors. This result supports our hypothesis that 'activating genomic al-

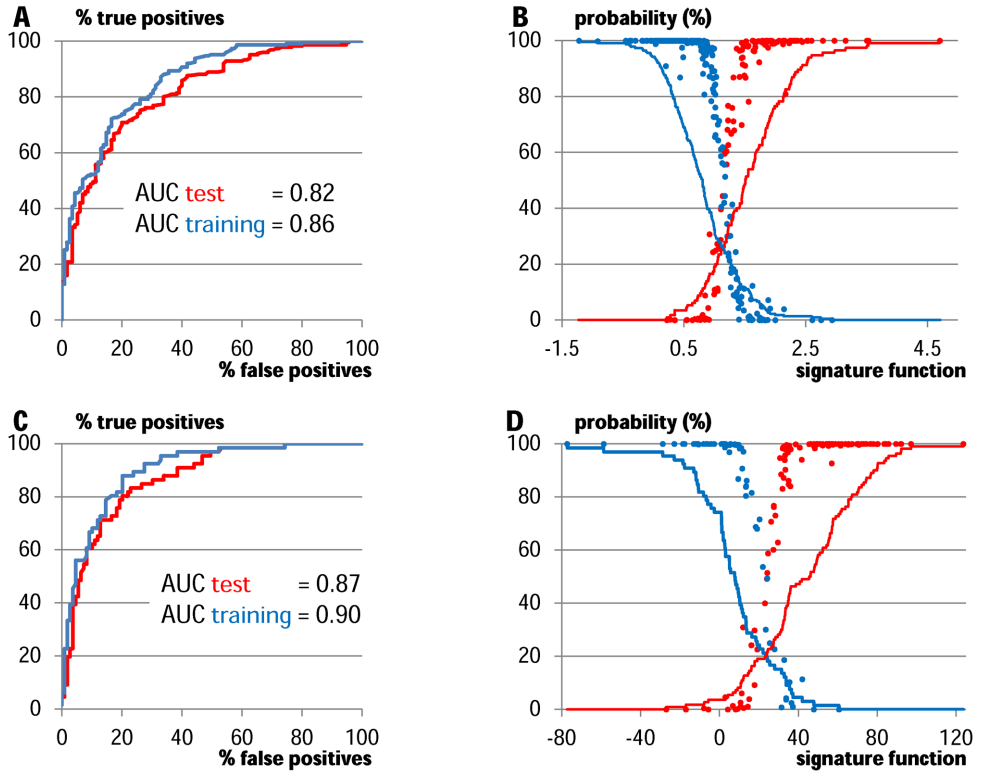


Figure 3. Predictions of driver alterations in the NOTCH pathway of ovarian cancer (**A and B**) and RTK pathway of colon cancer (**C and D**). The ROC curves (A and C) were computed for two-class separation for both the training (blue) and test (red) sets; the biomarker weights were averaged over 4096 randomized tests, where one half of tumor samples was used as a training set and another half as tumors was used as test set. The test set derived probabilities of classifications of individual tumors (dots) and the cumulative percentages of tumors (solid lines) are computed for tumors with (red) and without (blue) driver alterations for ovarian (panel B) and colon (panel D) cancers. The intersection of solid lines (panels B and D) defines a point of equally weighted accuracy: 75% for the NOTCH pathway in ovarian cancer and 82% for the RTK pathway in colon cancer.

terations in major cancer driver pathways are distinctively displayed at the transcriptional level'.

DISCUSSION

We tested the hypothesis that driver genomic alterations in cancer pathways can be distinctively recognized in the functional molecular networks of the tumor. To recognize altered cancer pathways in tumors at the transcriptional level, we proposed a new machine learning algorithm for identifying candidate biomarkers and constructing classifiers based on multivariate data. Our algorithm determines biomarker genes and constructs transcription-based signature functions that are specific to pathways in a given cancer type. The signature functions are fit using training datasets comprised of transcriptional profiles of tumors with respect to nominated driver pathways. Based on the value of the signature function, tumors with driver alterations in a given pathway can be separated from tumors that have no alterations in that pathway. To reduce the inevitable over fitting that is well-known to occur with machine learning algorithms applied to high-dimensional datasets, we implemented a robust 2-fold validation process in which all statistical associations were derived using half of the available dataset (considered as the training dataset) and tested the accuracy of the resulting classifier on the remaining half of the dataset (considered as the testing dataset). We tested our approach on seven major cancer pathways in breast, ovarian and colon cancers. Overall, the transcriptional signature functions make it possible to predict genomic driver alterations with an average AUC of ~83%.

The approach we developed is easily implemented in practice given it uses a weighted sum of gene expression levels and a straightforward classification scale. The important practical feature of the approach is that the signature function values make it possible to differentiate tumors based on a probability measure on the activation status of a driver pathway. For example, one can select tumors with high probability (e.g. higher than 95%) that have a given pathway activated, and then test such predictions on tumor cell lines or mouse models using pathway specific drugs. This type of validated transcriptional signature for specific cancer pathways can be used as a practical alternative or in combination with DNA sequencing methods. RNA-based methods could be especially valuable when DNA data is incomplete or ambiguous in interpretation.

Beyond the ability to classify the activation status of key driver pathways in cancer, there remains significant interest in determining new genes involved in cancer driver pathways. While our biomarker selection procedure was geared toward the identification of biomarkers that could optimize discrimination of pathway activation status, roughly 40% (146) of all biomarker genes we identified had evidence of involvement in cancer, supported by interactions with

known key genes in cancer pathways, signs of positive selection and differential expression. Further, the majority of these biomarker genes were not annotated as having a well-known role in cancer. Many of these top ranked genes that interact with cancer pathways and that are significantly amplified in several cancers may be interesting to explore as potential drug targets. For example, RP1, serine/threonine-protein kinase 19, is a top-ranked gene for the Wnt pathway in colon cancer. LRG1 for the PI3K signaling pathway and CDT1 for the cell cycle pathway were top-ranked for these respective pathways in breast cancer. The biomarker genes that could be considered as potential cancer genes are provided in Supplementary Table S1.

The issues facing the prediction of functional genomic alterations based on transcriptional data are not unlike the issues encountered in determining genotype based on phenotype (55). The success of the predictions depends on real biological associations between genomic alterations and phenotype as well as on the quality of the available data. In this present study, we assumed that a single genomic alteration in any of the genes known to comprise a given pathway of interest could result in the 'activation' of the pathway. Thus, we did not take into account the context of other genomic alterations in the tumor, the biological differences between alterations, the unknown genes and alterations that may also activate pathways, and so on. Therefore, it is of particular note that in spite of these uncertainties, the proposed approach was able to predict with reasonable accuracy the specific activation states of genomic alterations from the transcriptional data. These results support that the major driver alterations in cancer genomes produce distinct molecular phenotypes and may have diagnostic utility for better targeting of cancer therapies to individual cancer cases.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Author contributions: B.R. and E.S. designed the project; B.R. developed the algorithm; B.R. and D.R. developed the software and performed data analysis; N.B. and H.L. performed classification analysis using published methods; B.R., D.R., A.U. and E.S. wrote the paper.

FUNDING

Funding for open access charge: Icahn School of Medicine at Mount Sinai.
Conflict of interest statement. None declared.

REFERENCES

1. Simon,R. and Roychowdhury,S. (2013) Implementing personalized cancer genomics in clinical trials. *Nat. Rev.*, **12**, 358–369.
2. Kris,M.G., Johnson,B.E., Berry,L.D., Kwiatkowski,D.J., Iafrate,A.J., Wistuba,II, Varella-Garcia,M., Franklin,W.A., Aronson,S.L., Su,P.F. *et al.* (2014) Using multiplexed assays of oncogenic drivers in lung cancers to select targeted drugs. *JAMA*, **311**, 1998–2006.
3. Ulahannan,D., Kovac,M.B., Mulholland,P.J., Cazier,J.B. and Tomlinson,I. (2013) Technical and implementation issues in using next-generation sequencing of cancers in clinical practice. *British J. Cancer*, **109**, 827–835.
4. Peiffer,D.A., Le,J.M., Steemers,F.J., Chang,W., Jenniges,T., Garcia,F., Haden,K., Li,J., Shaw,C.A., Belmont,J. *et al.* (2006) High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res.*, **16**, 1136–1148.
5. Duan,J., Zhang,J.G., Deng,H.W. and Wang,Y.P. (2013) Comparative studies of copy number variation detection methods for next-generation sequencing technologies. *PloS one*, **8**, e59128.
6. Ciriello,G., Miller,M.L., Aksoy,B.A., Senbabaoglu,Y., Schultz,N. and Sander,C. (2013) Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.*, **45**, 1127–1133.
7. Cancer Genome Atlas Research Network. (2011) Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**, 609–615.
8. Cancer Genome Atlas Research Network. (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, **487**, 330–337.
9. Cancer Genome Atlas Research Network. (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.
10. Cancer Genome Atlas Research Network. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.
11. Chaisson,M.J., Huddleston,J., Dennis,M.Y., Sudmant,P.H., Malig,M., Hormozdiari,F., Antonacci,F., Surti,U., Sandstrom,R., Boitano,M. *et al.* (2015) Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, **517**, 608–611.
12. Watters,J.W. and Roberts,C.J. (2006) Developing gene expression signatures of pathway deregulation in tumors. *Mol. Cancer Therapeut.*, **5**, 2444–2449.
13. Bild,A.H., Yao,G., Chang,J.T., Wang,Q., Potti,A., Chasse,D., Joshi,M.B., Harpole,D., Lancaster,J.M., Berchuck,A. *et al.* (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, **439**, 353–357.
14. Miller,L.D., Smeds,J., George,J., Vega,V.B., Vergara,L., Ploner,A., Pawitan,Y., Hall,P., Klaar,S., Liu,E.T. *et al.* (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.
15. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. U S A*, **102**, 15545–15550.
16. Tarca,A.L., Draghici,S., Khatri,P., Hassan,S.S., Mittal,P., Kim,J.S., Kim,C.J., Kusanovic,J.P. and Romero,R. (2009) A novel signaling pathway impact analysis. *Bioinformatics*, **25**, 75–82.
17. The Mutation Consequences and Pathway Analysis working group of the International Cancer Genome Consortium. (2015) Pathway and network analysis of cancer genomes. *Nat. Meth.*, **12**, 615–621.
18. Vaske,C.J., Benz,S.C., Sanborn,J.Z., Earl,D., Szeto,C., Zhu,J., Haussler,D. and Stuart,J.M. (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, **26**, i237–i245.
19. Masica,D.L. and Karchin,R. (2011) Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival. *Cancer Res.*, **71**, 4550–4561.
20. Akavia,U.D., Litvin,O., Kim,J., Sanchez-Garcia,F., Kotliar,D., Causton,H.C., Pochanard,P., Mozes,E., Garraway,L.A. and Pe'er,D. (2010) An integrated approach to uncover drivers of cancer. *Cell*, **143**, 1005–1017.
21. Jornsten,R., Abenius,T., Kling,T., Schmidt,L., Johansson,E., Nordling,T.E., Nordlander,B., Sander,C., Gennemark,P., Funa,K. *et al.* (2011) Network modeling of the transcriptional effects of copy number aberrations in glioblastoma. *Mol. Syst. Biol.*, **7**, 486.
22. Bashashati,A., Haffari,G., Ding,J., Ha,G., Lui,K., Rosner,J., Huntsman,D.G., Caldas,C., Aparicio,S.A. and Shah,S.P. (2012) DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biol.*, **13**, R124.
23. Ding,J., McConechy,M.K., Horlings,H.M., Ha,G., Chun Chan,F., Funnell,T., Mullaly,S.C., Reimand,J., Bashashati,A., Bader,G.D. *et al.* (2015) Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. *Nat. Commun.*, **6**, 8554.
24. Hess,K.R., Anderson,K., Symmans,W.F., Valero,V., Ibrahim,N., Mejia,J.A., Booser,D., Theriault,R.L., Buzdar,A.U., Dempsey,P.J. *et al.* (2006) Pharmacogenomic predictor of sensitivity to preoperative

chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *J. Clin. Oncol.*, **24**, 4236–4244.
25. Dry,J.R., Pavey,S., Pratilas,C.A., Harbron,C., Runswick,S., Hodgson,D., Chresta,C., McCormack,R., Byrne,N., Cockerill,M. *et al.* (2010) Transcriptional pathway signatures predict MEK addiction and response to selumetinib (AZD6244). *Cancer Res.*, **70**, 2264–2273.
26. Naoi,Y., Kishi,K., Tanei,T., Tsunashima,R., Tominaga,N., Baba,Y., Kim,S.J., Taguchi,T., Tamaki,Y. and Noguchi,S. (2011) Prediction of pathologic complete response to sequential paclitaxel and 5-fluorouracil/epirubicin/cyclophosphamide therapy using a 70-gene classifier for breast cancers. *Cancer*, **117**, 3682–3690.
27. Aha,D., Kibler,D. and Albert,M. (1991) Instance-based learning algorithms. *Mach. Learn.*, **6**, 37–66.
28. Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.
29. Cohen,W.W. (1995) *Machine Learning: Proceedings of the Twelfth International Conference*. Morgan Kaufmann Publishers Inc., Tahoe City, pp. 115–123.
30. Elomaa,T. and Kaariainen,M. (2001) An analysis of reduced error pruning. *J. Artif. Int. Res.*, **15**, 163–187.
31. Frank,E., Wang,Y., Inglis,S., Holmes,G. and Witten,I. (1998) Using model trees for classification. *Mach. Learn.*, **32**, 63–76.
32. Friedman,J., Hastie,T. and Tibshirani,R. (2000) Additive logistic regression: a statistical view of boosting. *Ann. Statist.*, **28**, 337–407.
33. John,G.H. and Langley,P. (1995) *Uncertainty in artificial intelligence: Proceedings of the Eleventh conference*. Morgan Kaufmann Publishers Inc., Montreal, pp. 338–345.
34. Platt,J.C. (1999) In: Bernhard,S, lkopf Christopher,JCB and Alexander,JS (eds). *Advances in kernel methods*. MIT Press, pp. 185–208.
35. Quinlan,J.R. (1993) *C4.5 : programs for machine learning*. Morgan Kaufmann Publishers., San Mateo.
36. Freund,Y. and Schapire,R.E. (1996), *Machine Learning: Proceedings of the Thirteenth International Conference*, San Francisco, pp. 148–156.
37. Reva,B., Antipin,Y. and Sander,C. (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.*, **39**, e118.
38. Fawcett,T. (2006) An introduction to ROC analysis. *Pattern Recogn. Lett.*, **27**, 861–874.
39. Brodersen,K.H., Soon Ong,C., Stephan,K.E. and Buhmann,J.M. (2010) The balanced accuracy and its posterior distribution. *20th International Conference on Pattern Recognition*. Istanbul.
40. Press,W.H. and Numerical Recipes Software (Firm). (1993) 2nd ed., v2.0. ed. Cambridge University Press, Cambridge, pp. 2.
41. Fisher,R.A., Bennett,J.H., Fisher,R.A., Fisher,R.A. and Fisher,R.A. (1990) *Statistical methods, experimental design, and scientific inference*. Oxford University Press, Oxford.
42. Finkelstein,A.V. (1998) 3D protein folds: Homologs against errors — a simple estimate based on the random energy model. *Phys. Rev. Lett.*, **80**, 4823–4825.
43. Reva,B.A., Skolnick,J. and Finkelstein,A.V. (1999) Averaging interaction energies over homologs improves protein fold recognition in gapless threading. *Proteins Struct. Funct. Bioinform.,* **35**, 353–359.
44. Finkel'shtein,A.V., Rykunov,D.S., Lobanov,M., Badretdinov,F., Reva,B.A., Skolnick,J., Mirnyi,L.A. and Shakhnovich,E.I. (1999) [When and how can homologs overcome errors in the energy estimates and make the 3D structure prediction possible]. *Biofizika*, **44**, 980–991.
45. Magrane,M. and Consortium,U. (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database J. Biological Databases Curation*, **2011**, bar009.
46. Hall,M., Frank,E., Holmes,G., Pfahringer,B., Reutemann,P. and Witten,I.H. (2009) The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, **11**, 10–18.
47. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A. *et al.* (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
48. Liu,T., Lin,Y., Wen,X., Jorissen,R.N. and Gilson,M.K. (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.*, **35**, D198–D201.
49. Tamborero,D., Gonzalez-Perez,A., Perez-Llamas,C., Deu-Pons,J., Kandoth,C., Reimand,J., Lawrence,M.S., Getz,G., Bader,G.D., Ding,L. *et al.* (2013) Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.*, **3**, 2650.
50. Higgins,M.E., Claremont,M., Major,J.E., Sander,C. and Lash,A.E. (2007) CancerGenes: a gene selection resource for cancer genome projects. *Nucleic Acids Res.*, **35**, D721–D726.
51. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **39**, D52–D57.
52. Chen,E.Y., Tan,C.M., Kou,Y., Duan,Q., Wang,Z., Meirelles,G.V., Clark,N.R. and Ma'ayan,A. (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinform.*, **14**, 128.
53. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
54. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
55. Schadt,E.E., Woo,S. and Hao,K. (2012) Bayesian method to predict individual SNP genotypes from gene expression data. *Nat. Genet.*, **44**, 603–608.