Pathway-Informed Classification System (PICS) for Cancer Analysis Using Gene Expression Data

Michael R Young[1,2] and David L Craft[1]

[1]Department of Radiation Oncology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [2]Department of Biomedical Engineering and Biotechnology, University of Massachusetts, Intercampus, MA, USA.

ABSTRACT: We introduce Pathway-Informed Classification System (PICS) for classifying cancers based on tumor sample gene expression levels. PICS is a computational method capable of expeditiously elucidating both known and novel biological pathway involvement specific to various cancers and uses that learned pathway information to separate patients into distinct classes. The method clearly separates a pan-cancer dataset by tissue of origin and also sub-classifies individual cancer datasets into distinct survival classes. Gene expression values are collapsed into pathway scores that reveal which biological activities are most useful for clustering cancer cohorts into subtypes. Variants of the method allow it to be used on datasets that do and do not contain non-cancerous samples. Activity levels of all types of pathways, broadly grouped into metabolic, cellular processes and signaling, and immune system, are useful for separating the pan-cancer cohort. In the clustering of specific cancer types, certain pathway types become more valuable depending on the site being studied. For lung cancer, signaling pathways dominate; for pancreatic cancer, signaling and metabolic pathways dominate; and for melanoma, immune system pathways are the most useful. This work suggests the utility of pathway-level genomic analysis and points in the direction of using pathway classification for predicting the efficacy and side effects of drugs and radiation.

KEYWORDS: oncogenomics, systems biology, genomics-based optimization, data-mining, biological pathways

Background

Cancer is a genetically heterogeneous disease, both across cancer types and within. Nevertheless, most patients are prescribed treatments without regard to any specific biological signatures of their disease. Exceptions to this generally take the form of drug selection based on a single genetic mutation (examples include vemurafenib for BRAFv600 mutations, erlotinib for EGFR mutations, and crizotinib for ALK mutations) or in limited research settings based on a signature derived from a small number of genes (anywhere from one to about 50).[1,2] Such genomic-based partitioning of patients fails to predict with appropriate levels of confidence whether a potentially highly toxic drug will be effective or what side effects, and their intensities, may occur. This issue is compounded by the fact that advanced therapies are designed to target specific pathways, meaning treatments dependent on pathway regulation are being prescribed with incomplete knowledge of the state of the pathway itself.

A typical scenario for developing a predictive model for cancer treatment involves a small cohort of patients, usually less than 100, and gene expression data for each of them with upward of 20,000 gene expression levels measured. In this setting, one must be careful to avoid correlations that appear due to chance alone.[3] One way to deal with this is to collapse gene-level data into more compact, functional pathway-level data. A biological pathway is a set of biochemical reactions that perform a specific nameable function. A classic example of a metabolic pathway is the production of ATP from glucose. Some researchers attempt to discover pathways, or more generally interaction networks, through gene expression and protein–protein interaction data, and use that information to classify patients.[4,5] We take a different approach: utilizing pathways that have already been been discovered through decades of laboratory research and are systematically curated in several publicly accessible resources, including KEGG,[6] BioCarta,[7] Reactome,[8] and WikiPathways.[9] Several research teams have developed techniques to use pathways to aid in the interpretation of gene expression data. Some of the methods output a score indicating how dysregulated each pathway is for each patient.[10–16] Of these, Pathifier[12] and Kim-DeLisi[16] additionally use the pathway scores to classify patients. However, these algorithms are encumbered by computational challenges and limited scope and thus have not yet been successfully applied to actually improve treatment response prediction. For example, the algorithm closest to ours is Pathifier. The nonlinear dimensionality reduction scheme used in Pathifier can take up to a day to run, depending on the size of the pathways and the number of pathways. Our method runs in minutes. The utility of the purely linear approach (PCA)

is suggested in the appendix of Ref. 12; our work here verifies that computationally simple scoring algorithms are adequate. A method that runs in minutes as opposed to hours is very useful for data exploration.

We developed the Pathway-Informed Classification System (PICS) to extend the utility of pathway-based approaches to cancer classification with a method that is fast, robust, and scalable to large patient cohorts across all available cancer types. We validated the method on both pan-cancer cohorts, where we show separation of cancer types, and on individual cancer type cohorts, where we show that the patient clusters found split the patients into statistically distinct survival classes. This work serves as a stepping stone to a biologically informed statistical approach to predicting patient response to drugs and radiation therapy.
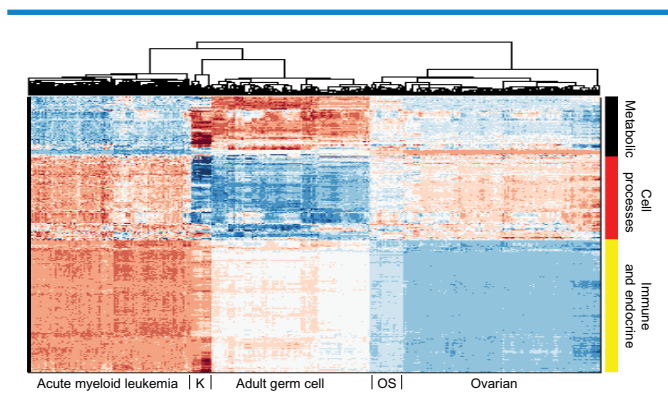

Figure 1. Hierarchical clustering across all pathways (scored by PCA method with two PCs per pathway) of five cancer cohorts combined into a single gene expression matrix. This figure serves as a proof of concept of pathway-based clustering by demonstrating the successful grouping of patients into their respective disease types via their gene expression levels. The hierarchical clustering results are shown in the upper tree dendrogram diagram, which has been suppressed for the pathway (row) sorting. The pathway list for this heatmap appears in Supplementary Information. We intentionally do not show a color key for the heatmaps when the particular values (magnitude, sign, etc.) of the matrix being plotted are not directly interpretable or valuable. In this case, since the values are from PCA, we do not include a color key. Note that the two ovarian sets intermingle and thus are not individually discernible, which validates that mRNA data from different studies can be analyzed together.
Abbreviations: K, kidney; OS, osteosarcoma.

Methods

PICS consists of two steps: the first step assigns to each patient and each biological pathway a score, which is either a single scalar value or a small set of numbers, and the second step takes these pathway scores for the entire cohort of patients and uses them to cluster the patients.

The input to PICS is a matrix of messenger RNA (mRNA) expression values for a large number of genes for all the patients of a cohort. We require the matrix to be full and that all the patient samples, which may include patients from across disease types and noncancerous matched tissue samples as well, come from the same microarray technology so that the values are comparable. We will discuss the incorporation of other types of genomic data as well, but for this report, we use only mRNA expression-level data, which we obtain from publicly available resources.

We obtain pathway descriptions from the KEGG database API, which is also publicly available. The KEGG pathway database contains 301 pathways related to humans. The definition of a biological pathway, in particular the boundaries of the pathways, is not well agreed on, and indeed some KEGG pathways are combinations of more primitive KEGG pathways. Some pathways have been attributed to a specific cancer (eg, KEGG HSA:05216, thyroid cancer pathway) when really those pathways contain gene interactions that are implicated in cancer in general. The set of pathways used will have an effect on the results of the PICS algorithm, but as a first pass, we start by including all possible pathways from a given resource (KEGG in our case) without regard to the pathway's speculated involvement in cancer. We also use KEGG modules, which are descriptions of more basic reactions and therefore generally involve fewer genes. The KEGG database contains 187 modules for humans.

PICS step 1: scoring pathways. For our purposes, a pathway or module will be represented by the set of genes in that pathway. We ignore the network topological structure of the pathway and instead view the pathway as a set of genes. Since our goal is an improved classification and predictive framework, we believe this representation is an appropriate level of detail, where biology in terms of closely interacting gene

Table 1. Information for the individual disease site classifications. Note that the melanoma set (*) does not have normal tissue samples and so is analyzed slightly differently, see text.

| CANCER TYPE	SOURCE	DATASET	GENES IN EXPRESSION MATRIX	NUMBER OF NORMAL SAMPLES	NUMBER OF CANCER SAMPLES
Pancreatic ductal adenocarcinoma	PRECOG	GSE21501	20936	30	102
Non-small cell lung carcinoma	PRECOG	GSE19188	42423	65	82
Adult germ cell carcinoma and seminoma	PRECOG	GSE3218	20936	6	74
Ovarian	PRECOG	GSE26712	20965	10	185
Glioblastoma	PRECOG	GSE13041	42391	4	80
Sub-optimally debulked ovarian	PRECOG	GSE26712	20965	10	95
Melanoma*	TCGA	NA	20512	0	470
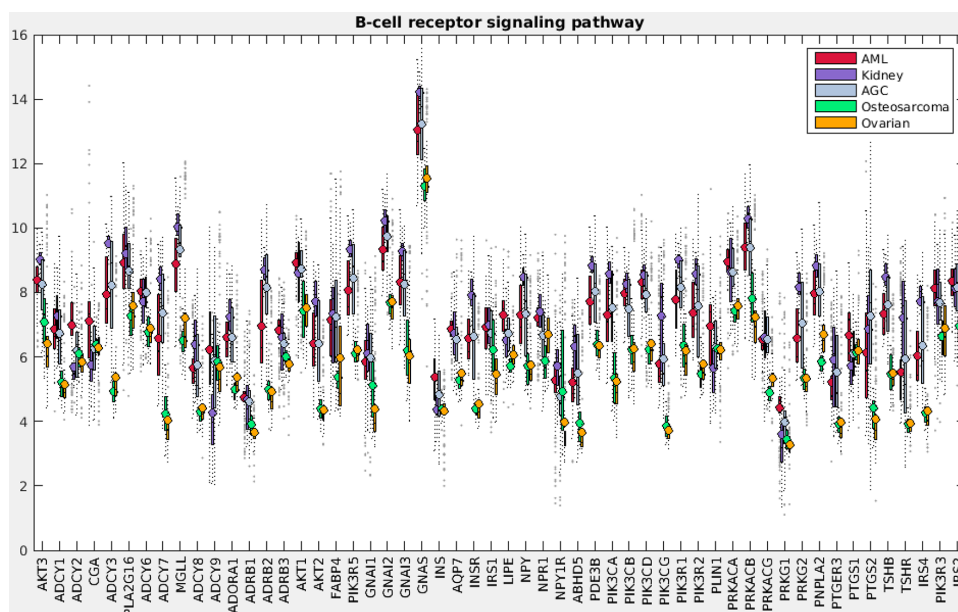
Figure 2. B-cell receptor signaling pathway gene expression levels for pan-cancer analysis. Levels are directly from the pan-cancer gene expression datasets, which have been $\log_2$ normalized.
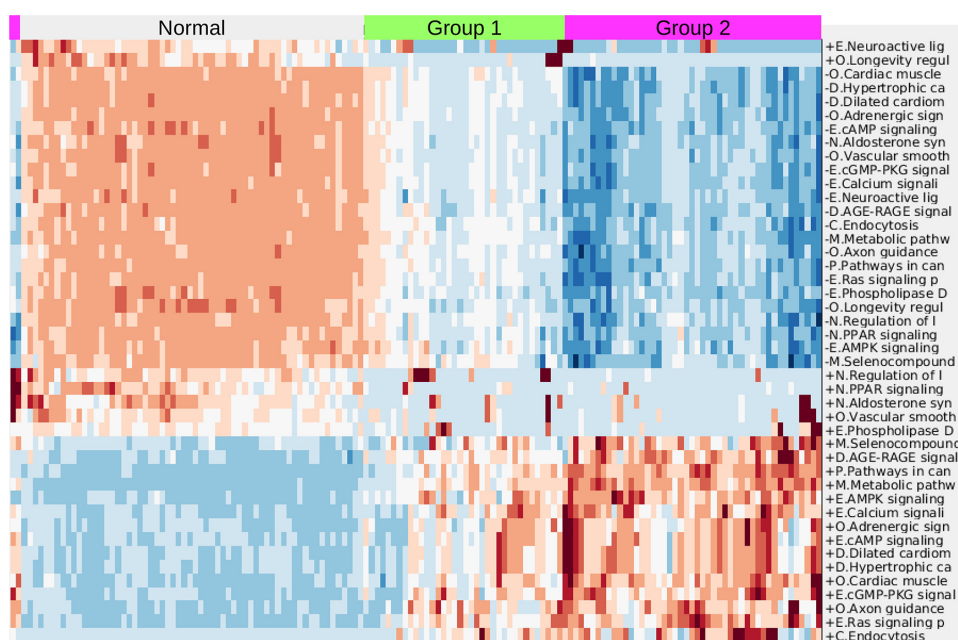

Figure 3. Clustergram for lung case including normal tissues. Colors are automatically scaled for optimal visual distinction; thus, the color bar is omitted. For the cancer samples (group 1 and group 2), red indicates increased expression levels of genes in the pathways and blue indicates decreased expression levels. The first letter of each pathway is an abbreviation for the KEGG pathway grouping. "+" stands for increased expression of the pathway genes in the cancer samples and "−" stands for decreased expression.
Abbreviations: E, environmental information processing; O, other organismal systems; D, diseases other than cancer; M, metabolic; C, cellular processes; P, pathways in cancer; N, endocrine.

products is taken into consideration, but detailed interactions are ignored.[12] It would not be hard to include the network structure of the pathways however, as has been done.[10,11,13] The only requirement for our pathway scoring is to take the gene expression data for pathway genes and condense this information into a single number or small set of numbers. For this, we introduce three methods: principal component analysis (PCA), normal tissue centroid (NTC), and gene expression deviation (GED). PCA can be used for any dataset, whereas NTC and GED methods require the dataset to have normal tissue samples.

Principal component analysis. The simplest way to reduce the dimensionality of a set of gene expression levels across a set of patients is using PCA.[17] Using PCA also provides

a straightforward means to condense the expression data to either a scalar (ie, the first principal component, that is, the projection of each patient's gene expression values onto the vector representing the first principal component) or a vector that consists of the first s principal component projections. The number of PCs used in a specific pathway scoring is a parameter. One can input either a fixed integer value (eg, 2) or a fraction f, which gets interpreted as round ($f \times I_D$) where I_D is the intrinsic dimensionality estimate[18] of the pathway data matrix. In this way, pathways with a larger intrinsic dimensionality will be represented by a larger number of principal components.

The PCA method makes no assumptions about the underlying datasets and is thus applicable to pan-cancer cohorts, datasets that include normal tissue samples, and single cancer-type datasets that contain just cancer samples. For sets that do have normal tissue samples, we provide two specialized methods, NTC and GED, which score pathways based on the differences of their gene expression values compared to the normal samples. Pan-cancer cohorts that do contain normal samples could in theory use the NTC or GED method, but for that one would have to combine all of the normal samples (further information below in the individual methods descriptions), or choose a set of them from only one of the cancers. Since neither of these makes sense from a biological perspective, we recommend and use the PCA method for pan-cancer analyses.

Normal tissue centroid. For a set of g genes (the genes of a particular pathway), each tissue sample can be viewed as a point in g-dimensional space. The NTC, or normal tissue centroid, is the location in this space computed by averaging the coordinates of each of the normal tissues. One can then score each sample by computing its distance in g-space from the NTC. This method is similar to the Pathifier method which, for each sample, computes the distance away from the normal samples along a curved line passing through the normal samples.[12]

Gene expression deviation. The resulting numerical data from the PCA and the NTC methods do not allow one to differentiate between gene over- and underexpression within a pathway, since this information gets lost in the algebra of those methods. We thus present a third scoring method called GED, or gene expression deviation. The GED score gives two numbers for each patient and each pathway: an aggregate overexpression score and an aggregate underexpression score. For each pathway, pathway genes are selected for inclusion in this score based on how the distribution of expression values for the cancer samples differs from that of the normal samples. Specifically, the Kolmogorov–Smirnov test for the difference of distributions is used. If a gene g passes the test, meaning its gene expression values are distributed significantly differently from the normal to the cancerous tissues, then for each patient p and that gene g, we form the differential-expression score Δ_{pg} as:

$$\Delta_{pg} = \frac{e_{pg} - m_{N_g}}{s_{N_g}}$$

where e_{pg} is the expression level for patient p of gene g, m_{N_g} is the mean expression level of gene g for the normal samples, and s_{N_g} is the standard deviation of the expression levels across the normal samples. Then for each sample we form two scores: one which adds up the positive values of Δ_{pg} across the pathway genes and one which adds up the negative values.

Deciding which genes to include. Many genes are involved in multiple pathways. For example, *MAPK*, *PI3K*, and *AKT* genes are all involved in over 60 KEGG pathways. Therefore, the gene expression levels of these genes may not be closely tied to the activity of any one of its particular pathways. On the other hand, most pathways contain at least a few genes that are unique to that pathway. For all scoring methods, we investigate which gene expressions should go into a pathway score based on how many other pathways the gene belongs to. In particular, we define a parameter called Mcut, for membership cutoff: a gene is included in a pathway score only if that gene belongs to Mcut or fewer pathways.

Deciding which pathways to include. Depending on the scoring technique, we can use various methods to decide whether or not to include a particular pathway in the matrix used for clustering.

For datasets containing normal tissue samples, we implement the PCA compression of the pathway as the first step in the selection process. Using the PCA representation, we compute silhouette scores for each sample[19] by using the groups, normal and cancerous. For each sample, the silhouette score is a measure of how well that sample fits into its defined group, given the data used for the clustering, which is the PCA compression. The maximum possible silhouette score for a sample point is 1, indicating that the sample perfectly belongs to the rest of the samples in that group. The average of all the silhouette scores gives a measure of how well the normal and cancerous samples cluster. Only pathways that have an average silhouette score above a given threshold are included in the final pathway score matrix.

For datasets that do not have normal samples, we rely on the patient cohort survival data to judge a pathway's usefulness. For these datasets, the PCA compressed pathway is clustered into two groups using either k-means or k-medoids clustering.[20] Given these two clusters, a pathway is accepted if the Kaplan–Meier survival curves for these groups separate at a P-value that is lower than a user-defined threshold.

PICS step 2: patient classification based on pathway scores. The above pathway selection and scoring results in the conversion of a gene expression-patient sample matrix, m, into a pathway score matrix, p. Matrix m is of size [number of genes]*[number of patients]. If each pathway is represented, for example, by two principal components, and h is the number of pathway that pass the silhouette score, then the size of the matrix p will be [2h]*[number of patients]. Note that in
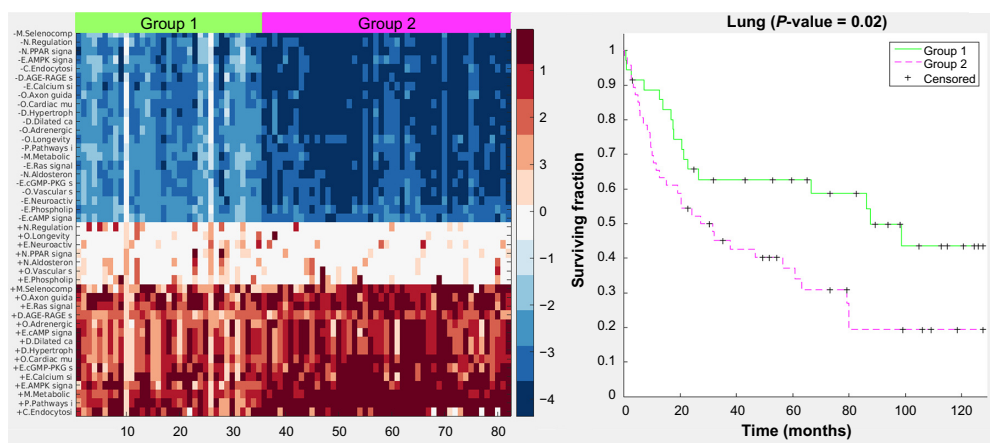
Figure 4. Heatmap and Kaplan–Meier curves (*P* value = 0.02) for the lung case. Color values are from the pathway scoring by the GED method. Values farther away from 0 mean the pathway gene expression levels are farther away from the gene expression levels of the normal samples.
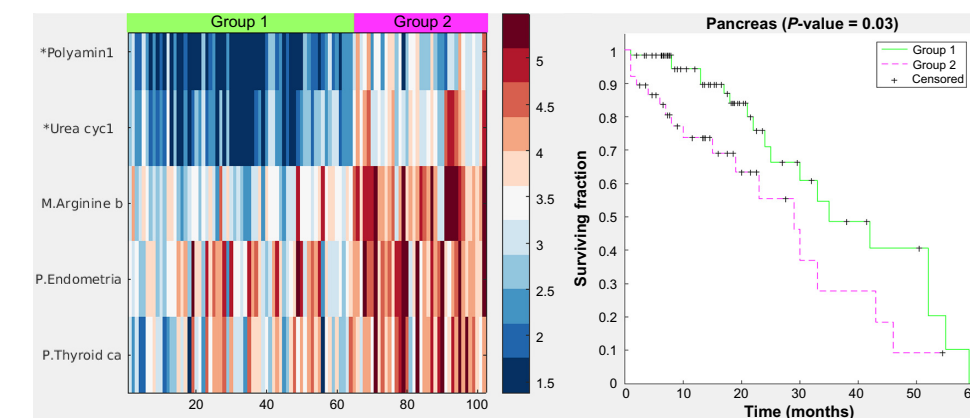

Figure 5. Heatmap and Kaplan–Meier curves (*P* value = 0.03) for the pancreas case. Color values are from the NTC pathway-scoring method, which yields higher values for the farther away a pathway's gene expression levels are from the gene levels of the normal samples. Group 2 is farther away in general than Group 1 and also has worse survivability, similar to the lung case.

general $2h \ll$ [number of genes], which means we have regularized or compressed the gene expression data.

We use three types of clustering to form distinct patient groups: k-means, k-medoids, and hierarchical. K-means and k-medoids are useful for fixing the number of clusters, for example, two, which makes for easier interpretability of separation of Kaplan–Meier curves. The k-medoids method, where the cluster centers are chosen as the "median" sample point of the cluster rather than the weighted average, that is, mean location, is more stable against outliers in the data. Hierarchical clustering is a more natural grouping of patients that does not involve prespecifying the number of groups. For hierarchical clustering, we use UPGMA (Unweighted Pair Group Method with Arithmetic mean) implemented in the Matlab (Version R2015b, Natick MA) function clustergram.

Data sources. The pan-cancer set contains six datasets consisting of 568 patients for five cancer types (acute myeloid leukemia [AML], kidney, adult germ cell [AGC], osteosarcoma, and ovarian) obtained from the PRECOG

database (Prediction of Clinical Outcomes from Genomic Profiles).[21] Two sets were used from different ovarian studies to verify that "like" cancer types from different studies cluster together using our method. All sets were profiled using the same microarray, the Affymetrix U133A, to minimize cross-set variability. The adult germ cell dataset included six normal tissue samples. None of the other sets included normal samples.

We also downloaded and normalized six datasets of cancers with normal tissue from the NCBI database via PRECOG: localized pancreatic duct adenocarcinoma (GSE21501),[22] non-small cell lung carcinoma (GSE19188),[23] adult germ cell carcinoma and seminoma (GSE3218),[24] ovarian (GSE26712),[25] glioblastoma (GSE13041),[26] and suboptimally debulked ovarian (GSE-26712). The adult germ cell, ovarian, glioblastoma (GBM), and suboptimally debulked ovarian datasets were profiled with the Affymetrix Human Genome U133A array. The pancreas set was profiled using the Agilent-014850 Whole Genome microarray. The lung set was profiled with the

Affymetrix U133Plus array. Datasets varied in the number of normal tissue samples, number of cancerous tissue samples, and number of genes reported (see Table 1). We also included one dataset, which did not have normal samples present – a melanoma cohort of 470 patient RNA-seq samples. This was a TCGA dataset obtained via the Broad Institute's GDAC firehose (gdac.broadinstitute.org, Skin Cutaneous Melanoma).

Analysis overview. For the pan-cancer analysis, we used all 301 pathways from KEGG and PCA scoring with 2 PCs per pathway. For the six cohorts with normal samples present, we looped through various scoring methods (PCA, NTC, GED), clustering types (k-means, k-medoids), pathway silhouette cutoffs, gene-pathway membership cutoff Mcut, and the number of principal components used in order to find the parameter set that best grouped the patients. After a coarse parameter search, we fine-tuned the parameters within that range to approach the "optimal" solution, which was determined by Kaplan–Meier curve separation via two-group clustering. The result was a reduced, cancer-specific pathway score matrix, which provides functional biological information and improved survival clustering.

We analyzed a melanoma dataset from TCGA consisting of 470 samples across disease stages. Since this patient cohort was large, and since the pathway selection criteria, due to the lack of normal samples, were based on the ability of the pathway to separate the KM curves, we split this dataset into a training and test set in order to gain more confidence in the robustness of the resulting classification. Similar to the classification of the individual cancer type datasets that had normal tissues, we looped through a wide range of parameter settings to find the best set of parameters to cluster the melanoma dataset. We additionally repeated the run for each parameter set 40 times, each time randomly splitting the cohort into a training set and a test set, and we finally chose the parameter set that consistently, across most of the 40 runs, effectively (regarding KM separation) clustered patients in both sets. We searched over the respective sizes of the training and test sets and found that a 50–50 split was optimal (this splitting was done for each of the runs by flipping a weighted coin for each patient, so the training and test group sizes are not always the same across the runs). Since there is a trade-off between how well you can separate the training set and how well you generalize to the testing set, we chose a point on the Pareto trade-off curve between these two objectives that balanced the two goals on average.

The optimal parameter set found was NumPCs = 2 and the cutoff value for the P-value of the KM curve separation for each pathway was 0.05. For displaying the results, we chose the median result (regarding the P-value for the Kaplan–Meier separation of the training set) over the 40 replicates.

Results

Pan-cancer clustering. Figure 1 demonstrates the successful hierarchical clustering by disease type. Pathways generally sort by type, which therefore allows us to add broad labels to the pathway groupings. See the Supplementary Information for the names of the pathways, from the top row downward. For example, all but one metabolic pathway is contained within the first grouping. This initial check validates that PCA can sort cancer types using pathway information via gene expression data. The two ovarian sets clustered together (mixed), and the normal tissues for the AGC normal samples clustered alongside the cancerous AGC samples (they are the six leftmost columns in that block). The most similar two cancer types from this viewpoint are kidney and AGC. AML and ovarian are similar with regard to metabolic and cell processes
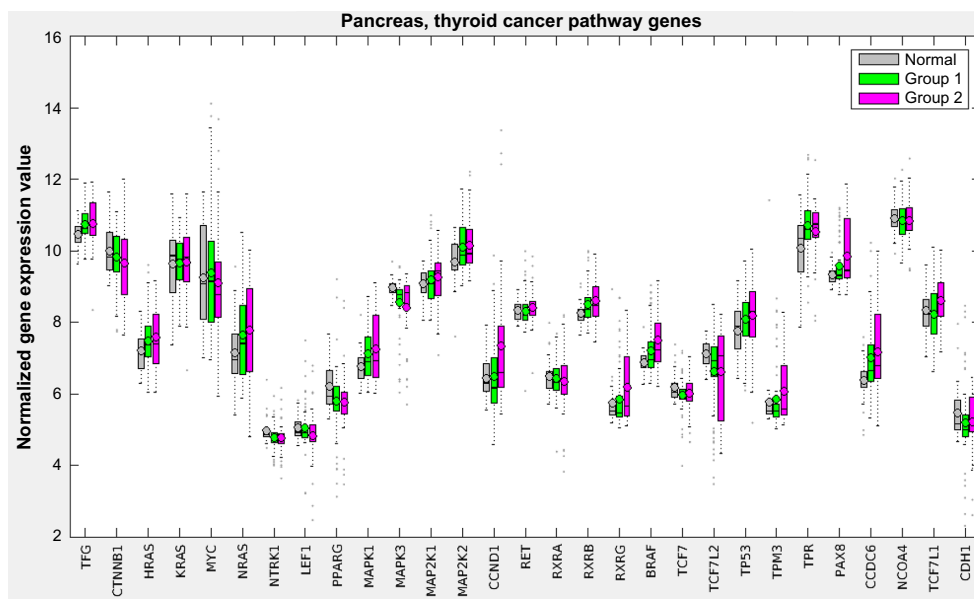

Figure 6. Gene expression levels for the pancreas cohort for the genes in the thyroid cancer pathway.

pathways but differ in their immune and endocrine signature. Since PCA decomposition was used to score these pathways, the absolute values of the pathway scores are not biologically meaningful (principal components are arbitrarily oriented; the magnitudes of the values are relevant but not the signs). In order to assess the differences across cancers regarding a particular pathway, we can plot the gene expression levels for the genes in that pathway across the five cancer types. As an example, gene expression levels for the B-cell receptor signaling pathway are shown in Figure 2. This confirms the data in the pan-cancer clustergram by showing that for this immune pathway, AML and kidney are similar, osteosarcoma and ovarian are similar, and AGC falls in between. Additionally, Figure 2 reveals that AML and kidney have the highest gene expression levels for this pathway overall and osteosarcoma and ovarian have the lowest.

Individual cancer site clustering with normal tissue present. Table 2 summarizes the runs and displays the best parameter sets/scoring types for each case. All six patient cohorts reached statistically significant solutions with *P* values <0.05. Pancreas and GBM were the most difficult to separate according to their *P* values of 0.03. The suboptimally debulked ovarian set had the best separation, per its low *P* value (0.0008). This is an improvement over an earlier gene expression–oriented study of the same dataset, which had a *P* value of 0.02 (the other studies did not produce patient groupings with Kaplan–Meier separation, so no comparisons can be made).[25] There is a general trend toward low *P* value (good separability) and low number of pathways and modules used in the classification; however, the pancreas dataset defies the trend being both relatively hard to separate and using very few pathways for optimal separation (discussed below).

For the lung case, hierarchical clustering of the final pathway matrix yields three distinct groups: the normal samples,

and Groups 1 and 2 (Fig. 3). This figure, and similar ones in the Supplementary Information, demonstrates that there is greater heterogeneity across tumor samples than across the noncancerous samples of the same tissue origin. The same Groups 1 and 2 were also uncovered via k-medoids clustering into two sets of just the cancer samples (Fig. 4). The clustergram colors (Fig. 3) are rescaled automatically for improved visualization of the groups, whereas in Figure 4, there is no scaling – the values are directly from the GED pathway scoring. Signaling pathways are the predominant type of pathways used in the lung cancer classification. The general trend, best observed in Figure 4, is that most pathways have both upregulation of the gene expression in the cancer groups (the bottom right red section) and downregulation of genes in those same pathways (the upper right blue section). The heatmap in Figure 4 shows a prominent center white strip. The color values in this heatmap correspond to the actual GED values of the pathway scoring (they are not rescaled) and thus we see that for these center strip pathways, the upregulation color values are close to zero, meaning those pathways, for the cancer samples, are almost exclusively downregulated. These downregulated pathways for the cancer samples are three endocrine system pathways, two organismal pathways, and two environmental information-processing pathways. We also see from both color maps that Group 2, which has less favorable survival statistics, is more distant from the normal gene levels than is Group 1.

Figure 5 displays a hybrid heatmap for the pancreas dataset. PICS reduced the 301 pathways and 187 modules from KEGG to a single metabolic pathway (arginine biosynthesis), two cancer pathways (thyroid and endometrial), and two modules (polyamine biosynthesis and urea cycle-1). The pathway signature produced by PICS clustered the patient cohort into three distinct groupings: normal tissue (not shown in heatmap), Group 1, and Group 2. Note that cancer pathways curated in KEGG, such as the thyroid cancer and the endometrial cancer,

Table 2. PICS parameter settings and results for individual disease site classifications. Silhouette cutoff for pathways based on their membership in other pathways. *P*-values are for the logrank statistic for Kaplan–Meier curve separation, and for the melanoma dataset, the two values are for the training and the test subsets.

| CANCER TYPE	PATHWAY SCORING METHOD	CLUSTERING TYPE	SILHOUETTE CUTOFF	MCUT	PATHWAYS & MODULES USED	GROUP 1 SIZE	GROUP 2 SIZE	*P*-VALUE* (<0.05)
Pancreatic ductal adenocarcinoma	NTC	K-means	0.4	2.5	5	64	38	0.03
Non-small cell lung carcinoma	GED	K-medoids	0.67	50	44	35	47	0.02
Adult germ cell carcinoma and seminoma	GED	K-means	0.375	15	224	15	59	0.02
Ovarian	GED	K-medoids	0.7	40	43	125	60	0.01
Giloblastoma	GED	K-medoids	0.475	25	363	48	32	0.03
Sub-optimally debulked ovarian	NTC	K-medoids	0.775	15	14	61	34	<0.01
Melanoma	PCA	K-medoids	NA	10	74	285	185	0.03, 0.04
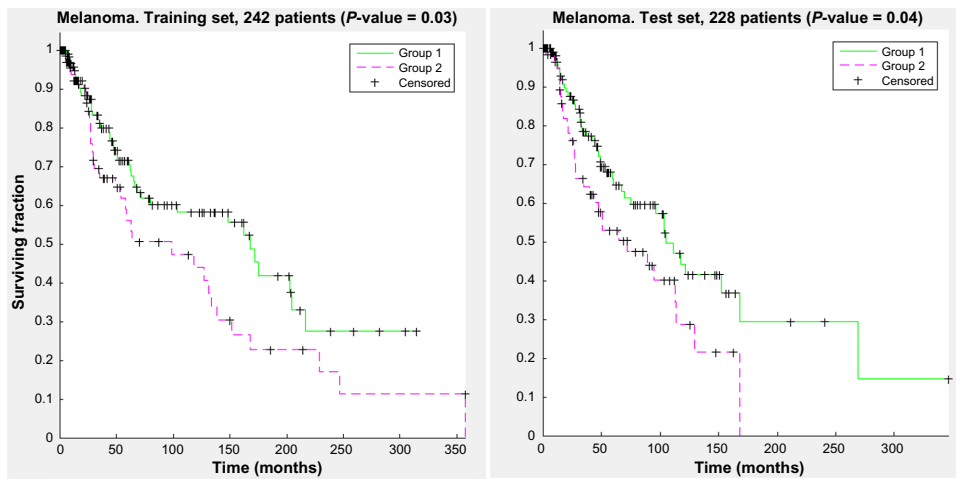
Figure 7. Kaplan–Meier curves for a representative run from the optimal parameter set found for the melanoma dataset. Both the training set and the test set show statistically significant separation between the two clustered groups.

are in fact relatively general cancer pathways, containing the signaling sub-pathways commonly implicated in cancer, including *MAPK*, *RAS*, *PI3K-AKT*, *Wnt*, *p53*, and *ErbB*. The heatmap shows that Group 2, which has worse survivability, is more dysregulated, as judged by NTC score, than Group 1. Since it is not possible to directly assess increased or decreased gene expression levels for pathways in the heatmap figure with the NTC method, Figure 6 displays the gene expression levels across the groups for one of the pathways used in the pancreas clustering, the thyroid cancer pathway. Here, we observe that most genes in this pathway are on average similar across the three groups, with a tendency toward more variability in

Figure 8. Hierarchical clustering for the pathways and k-medoids clustering for the combined training and testing samples. Pathway labels have been suppressed since 148 pathways and modules were used. The prominent red–blue band toward the top and blue–red band at the bottom are dominated by immune system–related pathways, and investigation into the gene expression levels from those pathways reveals heightened immune system expression for group 1, which showed higher overall median survival. The color scale values are the values from the PCA and are thus not directly interpretable; hence, no color key is shown.

Group 2, but some genes, such as *MAPK1*, *PAX8*, and *CCDC6*, are upregulated.

See the Supplementary Information for the heatmaps and Kaplan–Meier curves for the other individual disease sites.

Individual cancer site clustering without normal samples. Figure 7 shows the KM curve separation for this run for the training and the test sets.

Figure 8 displays the resulting pathway matrix for the melanoma run as a heatmap and contains the entire dataset (training and test). The most prominent pathway groupings are the thick red-blue band toward the top and the thinner blue-red band at the bottom. Investigation into the pathways involved in these bands (which are not shown due to the large number of pathways used, but see the Supplementary Information for the list of pathways in the same order as the heatmap rows) reveals that these are immune system pathways, which align with the general opinion of the relevance of the immune system in melanoma.[27] Group 1, which has generally higher levels of active immune system genes – as demonstrated in Figure 9, which shows the gene expression comparison for the T-cell receptor signaling pathway – also shows better survival characteristics.
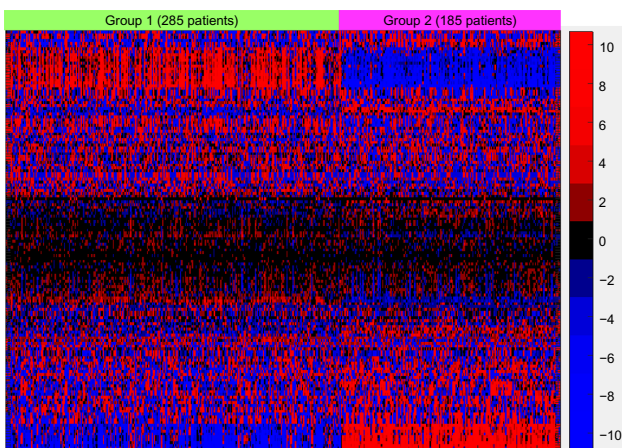
Discussion and Conclusions

The heterogeneity across cancers, even within the same type of cancer, has been discussed much in the literature over the past decade.[28] However, for the most part, it has not been known how to use this information to better inform treatment decisions for patients.[29] In this report, we have introduced a method called PICS that can classify cancer patients based on their gene expression levels by compressing this information from the gene level to the pathway level. We view this data regularization/dimension reduction technique and the accompanying classification system as the first step in building a computational tool to aid with patient treatment planning decision making. PICS is advantageous due to its
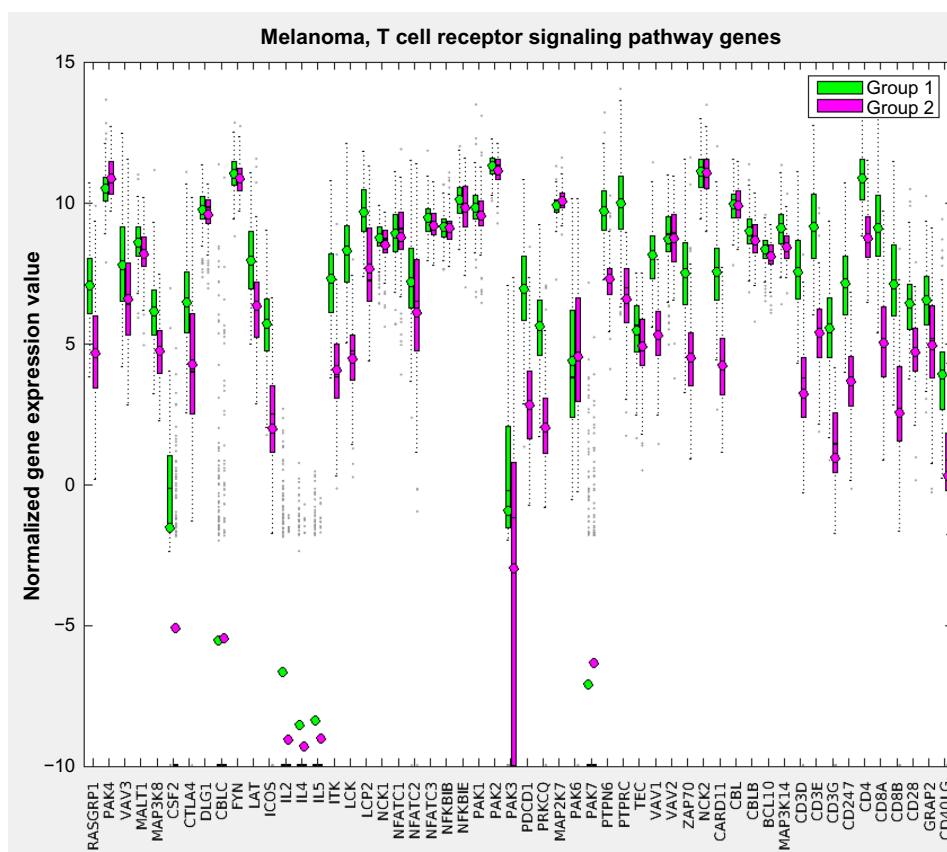
Figure 9. A sample gene expression value plot for the melanoma training set of the T-cell receptor signaling pathway showing that in general Group 1, the higher survivability class, has elevated expression levels for the genes involved in that pathway.

speed (a PICS run takes minutes compared to the related method Pathifier,[12] which takes hours), its interpretability, and its flexibility to handle a variety of types of datasets, as we have demonstrated. Although our main focus is broader and more statistical in nature, we welcome deeper biological analysis on the multitude of discovered pathway-cancer corollaries.

The ultimate vision is a system that takes into account a plethora of patient-level information (tumor site, stage, size, tumor and germline genomics data, patient age, etc.) and predicts what drugs and drug levels (including radiation) will likely be most beneficial and least toxic to the patient. With this goal in mind, we began with a method that classifies patients based only on their gene expression data, as measured by microarrays (RNAseq data would also be immediately possible to use). We have shown across a wide range of cancer types that classification based on biological pathway distillation of gene expression data allows for grouping patients, via clustering algorithms, into distinct survivability classes. The next steps are to obtain datasets containing treatment information as well as genomic and clinical variables, by mining currently existing sources, by performing clinical trials, or by gathering institutional data, in order to build a machine learning system for outcome prediction. We hypothesize that compressing the genetic information into pathway level scores will be useful in this broader context.

While we have explored several variants of the PICS algorithm for classifying the various cancer types studied, there is still room for refinement in the technique. For example, we relied solely on KEGG curated pathways. It remains to be investigated if other curated pathway databases might be better suited to the task. One shortcoming in KEGG is the redundancy across pathways since many pathways are built up of other smaller pathways. The MSigDB[30] database, which consolidates many pathway resources into a single repository, will likely be a useful resource. It also remains to be seen if using the network connectivity of the pathways rather than just the list of genes the pathways contain will be a useful adjustment to the algorithm. We hope that by building a method that allows for pan-cancer analysis, we can extend learning across cancer types and make cancer research less of an isolated, site-specific endeavor. If common pathway signatures across various cancer types end up being predictive of the success of drugs, the goals of biological knowledge building and improved patient care will both have been achieved.

There are pros and cons associated with each of the pathway-scoring methods introduced. The PCA method is the simplest method and is applicable to datasets that do not have normal tissues, but it was outperformed in the datasets that did contain normal tissues. This is likely because both the NTC and GED pathway-scoring meth-

ods are based on how different they are from the normal pathways, and this information, not surprisingly, is useful for classifying cancer patients. Our results echo the trend that has been observed earlier that the more dysregulation there is across pathways, the more malignant is the cancer.[12,31] It may prove useful to combine different scoring methods in the same analysis; for example, some pathways might better be scored by GED and some better by NTC, but we have not investigated that as of yet. We also used the global parameter Mcut to vary how many genes are used to score each of the pathways, but the scoring of each pathway should likely be done in a more pathway-dependent manner. Nevertheless, the results show good classification even without detailed modeling of pathway activity. We have intentionally tried to find the right balance between statistical modeling and biology in this effort in order not to be pulled too far in either direction.

The most important genomic addition to PICS will be the ability to include more molecular information, including post-transcriptional and post-translational data, into the pathway scoring. We chose pathways as our fundamental unit for classification based on the idea that they represent the right level of biological detail for the statistical prediction problem, and pathway activity can be informed by a large variety of genomic assays. Specifically, we expect gene mutations, chromosomal rearrangements and fusions, copy number variations, methylation, and proteomics to yield useful information regarding pathway activity. Our current pathway scoring and clustering approach will also benefit from insights into biological redundancies and more complex modeling of cellular and organismal functioning. For example, the use of logical statements such as AND, OR, and IF in modeling cellular systems will likely aid in classification, but this makes the search space for optimal classifiers combinatorial and therefore much harder. Informing these additions with known experimental biology will be necessary.

In this report, although we focused on classifying cancers, we also plan to apply PICS to classify normal tissues and the relevant organ systems of patients in order to build a system to predict toxicities and side effects as well, which is equally important for optimal cancer management. In the last five years, 70 cancer drugs were approved by the FDA. At this rate of oncology drug approval, sophisticated clinical decision tools are increasingly needed to make the best therapeutic choices.

Author Contributions

Conceived and designed the experiments: MRY, DLC. Analyzed the data: MRY, DLC. Wrote the first draft of the manuscript: MRY, DLC. Contributed to the writing of the manuscript: MRY, DLC. Agree with manuscript results and conclusions: MRY, DLC. Jointly developed the structure and arguments for the paper: MRY, DLC. Made critical revisions and approved final version: MRY, DLC. Both authors reviewed and approved of the final manuscript.

Supplementary Material

i. A pan-cancer pathway and principle component list corresponding to the hierarchical clustering across all pathways for five cancer cohorts.

ii. A pathway/module list, clustergram, heatmap, and Kaplan-Meier curve for all results not explicitly shown in the main text.

REFERENCES

1. Scott DW, Chan FC, Hong F, et al. Gene expression–based model using formalin-fixed paraffin-embedded biopsies predicts overall survival in advanced-stage classical Hodgkin lymphoma. *J Clin Oncol*. 2013;31(6):692–700.
2. Der SD, Sykes J, Pintilie M, et al. Validation of a histology-independent prognostic gene signature for early-stage, non–small-cell lung cancer including stage in patients. *J Thorac Oncol*. 2014;9(1):59–64.
3. Chalkidou A, ODoherty MJ, Marsden PK. False discovery rates in PET and CT studies with texture features: a systematic review. *PLoS One*. 2015;10(5): e0124165.
4. Song Q, Wang H, Bao J, et al. Systems biology approach to studying proliferation-dependent prognostic subnetworks in breast cancer. *Sci Rep*. 2015;5:12981.
5. Chuang H-Y, Lee E, Liu Y-T, Lee D, Ideker T. Network-based classification of breast cancer metastasis. *Mol Syst Biol*. 2007;3(1):140.
6. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res*. 2016;44(D1):D457–62.
7. Nishimura D. Biocarta. *Biotech Software Internet Rep*. 2001;2(3):117–20.
8. Joshi-Tope G, Gillespie M, Vastrik I, et al. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res*. 2005;33(suppl 1):D428–32.
9. Pico AR, Kelder T, Iersel MP, Hanspers K, Conklin BR, Evelo C. Wikipathways: pathway editing for the people. A Bradford Book; The MIT Press, Cambridge, MA, London, England. *PLoS Biol*. 2008;6(7).
10. Vaske C, Benz S, Sanborn J, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*. 2010;26(12):i237–45.
11. Ibrahim M, Jassim S, Cawthorne MA, Langlands K. A matlab tool for pathway enrichment using a topology-based pathway regulation score. *BMC Bioinformatics*. 2014;15(1):358.
12. Drier Y, Sheffer M, Domany E. Pathway-based personalized analysis of cancer. *Proc Natl Acad Sci USA*. 2013;110(16):6388–93.
13. Tarca A, Draghici S, Khatri P, et al. A novel signaling pathway impact analysis. *Bioinformatics*. 2009;25(1):75–82.
14. Margolin A, Nemenman I, Basso K, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*. 2006;7(suppl 1):S7.
15. Duarte N, Becker S, Jamshidi N, et al. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci U S A*. 2007;104(6):1777–82.
16. Kim S, Kon M, DeLisi C. Pathway-based classification of cancer subtypes. *Biol Direct*. 2012;7(1):21.
17. Van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. 2008;9(2579–2605):85.
18. Levina E, Bickel P. Maximum likelihood estimation of intrinsic dimension. *Advances in Neural Information Processing Systems*. 2004:777–84.
19. Kaufman L, Rousseeuw PJ. *Finding Groups in Data: An Introduction to Cluster Analysis*. Vol 344. Hoboken, NJ: John Wiley & Sons; 2009.
20. Nandi AK, Fa R, Abu-Jamous B. *Integrative Cluster Analysis in Bioinformatics*. Hoboken, NJ: John Wiley & Sons; 2015.
21. Gentles A, Newman A, Liu C, et al. The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat Med*. 2015;21(8):938–45.
22. Stratford JK, Bentrem DJ, Anderson JM, et al. A six-gene signature predicts survival of patients with localized pancreatic ductal adenocarcinoma. *PLoS Med*. 2010;7(7):e1000307.
23. Konstantinopoulos PA, Spentzos D, Karlan BY, et al. Gene expression profile of brcaness that correlates with responsiveness to chemotherapy and with outcome in patients with epithelial ovarian cancer. *J Clin Oncol*. 2010;28(22): 3555–61.
24. Korkola JE, Houldsworth J, Chadalavada RSV, et al. Down-regulation of stem cell genes, including those in a 200-kb gene cluster at 12p13. 31, is associated with in vivo differentiation of human male germ cell tumors. *Cancer Res*. 2006;66(2):820–7.
25. Bonome T, Levine DA, Shih J, et al. A gene signature predicting for survival in suboptimally debulked patients with ovarian cancer. *Cancer Res*. 2008;68(13): 5478–86.

26. Lee Y, Scheck AC, Cloughesy TF, et al. Gene expression analysis of glioblastomas identifies the major molecular basis for the prognostic benefit of younger age. *BMC Med Genomics*. 2008;1(1):52.

27. Cancer Genome Atlas Network. Genomic classification of cutaneous melanoma. *Cell*. 2015;161(7):1681–96.

28. Fisher R, Pusztai L, Swanton C. Cancer heterogeneity: implications for targeted therapeutics. *Br J Cancer*. 2013;108(3):479–85.

29. Manolio TA. Bringing genome-wide association findings into clinical use. *Nat Rev Genet*. 2013;14(8):549–58.

30. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*. 2011; 27(12):1739–40.

31. Rondeau S, Vacher S, De Koning L, et al. ATM has a major role in the double-strand break repair pathway dysregulation in sporadic breast carcinomas and is an independent prognostic marker at both mrna and protein levels. *Br J Cancer*. 2015;112(6):1059–66.