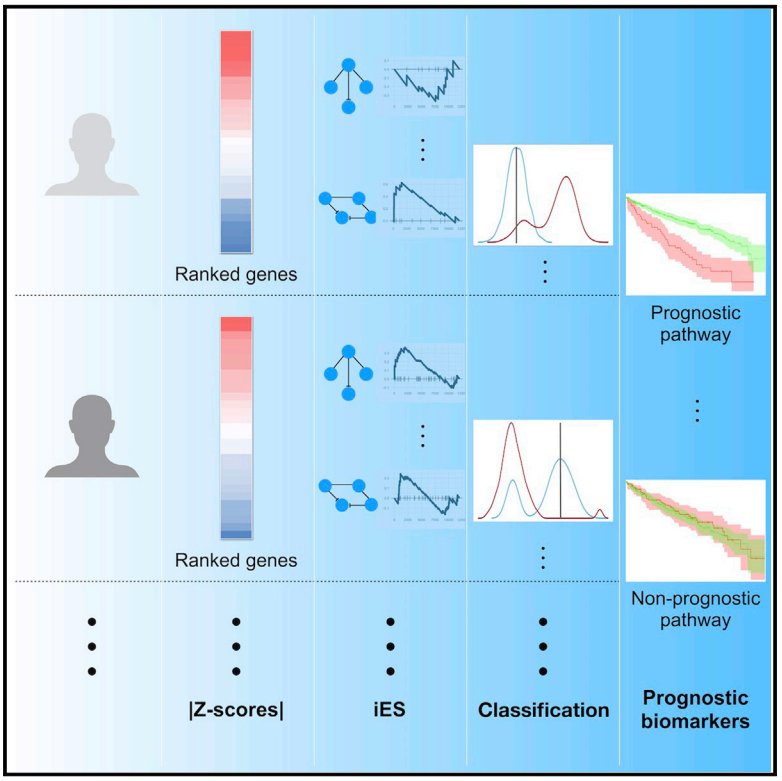
Cell Reports
Methods

Article

Pan-cancer analysis of pathway-based gene expression pattern at the individual level reveals biomarkers of clinical prognosis

Graphical abstract



Highlights

- iPath is a computational tool for identifying prognostic biomarker pathways in cancer
- Apply iPath in a pan-cancer analysis
- Pathway-based biomarkers are more effective than single-gene biomarkers
- iPath might be potentially applied in personalized cancer treatment

Authors

Kenong Su, Qi Yu, Ronglai Shen, Shi-Yong Sun, Carlos S. Moreno, Xiaoxian Li, Zhaohui S. Qin

Correspondence

zhaohui.qin@emory.edu

In brief

Detecting perturbed pathways at the patient level is crucial for personalized treatment. Su et al. create iPath, a computational method that enables identification of disrupted pathways for individual patients. They select and validate cancer-specific prognostic pathways and demonstrate that pathway-based biomarkers are more effective than single-gene biomarkers.

Su et al., 2021, Cell Reports Methods *1*, 100050
August 23, 2021
https://doi.org/10.1016/j.crmeth.2021.100050

CellPress

Cell Reports Methods

CellPress
OPEN ACCESS

Article

Pan-cancer analysis of pathway-based gene expression pattern at the individual level reveals biomarkers of clinical prognosis

Kenong Su,[1] Qi Yu,[2] Ronglai Shen,[3] Shi-Yong Sun,[4] Carlos S. Moreno,[5] Xiaoxian Li,[5] and Zhaohui S. Qin[1,2,6,7,*]
[1]Department of Computer Science, Emory University, Atlanta, GA 30322, USA
[2]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322, USA
[3]Department of Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY 10017, USA
[4]Department of Hematology & Medical Oncology, Emory University School of Medicine, Atlanta, GA 30322, USA
[5]Department of Pathology and Laboratory Medicine, Emory University School of Medicine, Atlanta, GA 30322, USA
[6]Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA 30322, USA
[7]Lead contact
*Correspondence: zhaohui.qin@emory.edu
https://doi.org/10.1016/j.crmeth.2021.100050

MOTIVATION Abundant single-gene biomarkers have been identified and used in clinics. However, hundreds of oncogenes or tumor-suppressor genes are involved during the process of tumorigenesis, and the efficacy of single-gene biomarkers might be hampered by the extensively variable expression levels measured by high-throughput assays. In this study, we devised a computational method named iPath to identify prognostic biomarker pathways, one sample at a time. To test its utility, we conducted a pan-cancer analysis across 14 cancer types from The Cancer Genome Atlas and demonstrated that iPath is capable of identifying highly predictive biomarkers for clinical outcomes, including overall survival, tumor subtypes, and tumor-stage classifications. We found that pathway-based biomarkers are more robust and effective than single genes.

SUMMARY

Identifying biomarkers to predict the clinical outcomes of individual patients is a fundamental problem in clinical oncology. Multiple single-gene biomarkers have already been identified and used in clinics. However, multiple oncogenes or tumor-suppressor genes are involved during the process of tumorigenesis. Additionally, the efficacy of single-gene biomarkers is limited by the extensively variable expression levels measured by high-throughput assays. In this study, we hypothesize that in individual tumor samples, the disruption of transcription homeostasis in key pathways or gene sets plays an important role in tumorigenesis and has profound implications for the patient's clinical outcome. We devised a computational method named iPath to identify, at the individual-sample level, which pathways or gene sets significantly deviate from their norms. We conducted a pan-cancer analysis and demonstrated that iPath is capable of identifying highly predictive biomarkers for clinical outcomes, including overall survival, tumor subtypes, and tumor-stage classifications.

INTRODUCTION

Cancer is a leading cause of morbidity and mortality worldwide, and its prevalence is rapidly increasing, primarily due to the aging of the population. Given this, there is an urgent need for understanding the molecular mechanisms of tumorigenesis to develop effective treatments. It has long been recognized that dramatic transcriptome alteration is a hallmark of cancer (Hanahan and Weinberg, 2011). Detecting gene signatures in transcriptome profiling data have been an essential step for many cancer studies (Cantini et al., 2018; Dang et al., 2019; Xu et al., 2016; Zuo et al., 2019). Using microarray or RNA sequencing (RNA-seq), many important discoveries have been made by using differential expression (DE) detection techniques (Rapaport et al., 2013; Soneson and Delorenzi, 2013; Zhao et al., 2014). For example, important biomarker genes in breast cancer have been identified by using high-throughput technologies (van de Vijver et al., 2002) (Joe and Nam, 2016).

Despite the successes and importance of DE gene detection, significant challenges limit its utility. First, the expression level of many genes is rather dynamic and is affected by many factors that might or might not relate to the disease. Second, most

high-throughput technologies produce data with substantial uncertainties: a long list of DE genes is usually produced, and many of them are potentially false positives. The low reproducibility of high-throughput technologies has long been acknowledged (Li et al., 2011). To overcome this challenge, scientists have developed gene set enrichment analysis (GSEA) (Subramanian et al., 2005). Instead of individual genes, GSEA focuses on pre-defined gene sets and uses rankings instead of actual expression levels to determine whether a given gene set shows concordant and statistically significant changes between two conditions. GSEA is specifically designed to analyze inherently noisy data produced from high-throughput assays, such as microarray and RNA-seq. Operationally, GSEA first ranks all genes in the genome on the basis the level of expression changes between two conditions (e.g., treatment and control). It then focuses on whether the genes from pre-defined functional gene sets locate toward the top or bottom of the sorted list by calculating a Kolmogorov-Smirnov version of enrichment score (ES). GSEA has been shown to be a powerful method, especially for cancer research. Besides GSEA, a dozen or so methods designed for pathway analysis were developed around the same time, and Butte's group systematically reviewed these pathway analytic approaches during the previous 10 years in 2012 (Khatri et al., 2012). Recent studies demonstrate that alterations in multiple genes tend to accumulate in pathways central to the control of cell growth and cell-fate determination (Arya and White, 2015; Schmelzle and Hall, 2000; Zhang and Liu, 2002).

However, cancer is characterized by tremendous phenotype heterogeneity, which is also reflected at the molecular level. The new precision-medicine philosophy advocates for a treatment plan that targets the unique characteristics of the tumor. Therefore, it is critically important that one focuses on the unique pattern shown in the individual tumor sample in order to identify the most promising treatment strategy for the patient. Despite its success, GSEA is predominantly carried out as a follow-up to DE analysis. GSEA looks for those gene sets that have gone through significant systematic changes between two groups of samples. Therefore, significant pathway changes that occur only in a small number of samples will likely be missed by GSEA.

Cancer is a disease of the genome. Multiple types of genomic or epigenomic alterations have been linked to human malignancies, including mutations and translocations, and changes in DNA copy number, gene expression, and CpG methylation patterns. Given the vast heterogeneity among disease prognoses, it is of great interest to identify biomarkers that can predict clinical progression and outcomes. In a recent study, Uhlen et al. (2017) comprehensively and systematically correlated gene expression differences with patient survival. Using data from The Cancer Genome Atlas (TCGA) (The International Cancer Genome Consortium, 2010), they identified multiple candidate prognostic genes whose expression level strongly correlated with the patients' overall survival.

Despite identifying many prognostic genes, the substantial variation and uncertainties that are ubiquitous in high-throughput technologies might raise concerns of robustness when using a single gene as the biomarker. Additionally, cancer is a complex disease: tens, or even hundreds, of genes are interactively involved and together play an important role in tumorigenesis and progression. Therefore, we hypothesize that gene sets—especially pathways and pre-defined, biologically meaningful gene sets—could serve as better biomarkers than individual genes to predict clinical outcomes for cancer patients in terms of robustness and interpretability. We acknowledge that a pathway is much more than just a gene set, given that how genes interact with each other is exceedingly important. However, in this work we only focus on the gene membership part of the pathway; for simplicity considerations, we use the two words interchangeably. Given that whole-transcriptome profiling has become increasingly affordable in the clinic, in this study we explored the feasibility and efficacy of using the expression profiles of pathways or pre-defined gene sets as biomarkers and compared them with individual-gene biomarkers.

Here, we introduce iPath, or individual-level pathway analysis, to quantify the magnitude of alteration occurring for a particular pathway at the individual-sample level. Our goal is to understand cancer one tumor sample at a time. Given that tens or hundreds of genes are required to work together harmoniously to achieve even a simple biological function, and because high-throughput assays are known to produce data with a substantial amount of noise and artifacts, we believe it is more effective and robust to study genes in a pathway or gene set collectively, as a group, rather than one by one. To achieve this, for each pathway we calculate a pathway-based individual-level enrichment score (iES) (see STAR Methods) to classify tumor samples into two groups—normal-like or perturbed—and then conduct a formal statistical test (reporting a log-rank p value) to check whether such grouping has any implication on clinical outcomes such as overall survival.

The idea of conducting individual-level pathway analysis has appeared in the literature. For example, Barbie et al. (2009) introduced single-sample GSEA (ssGSEA), which internally integrates the calculation of GSEA with a modified weighting factor. Gundem and Lopez-Bigas (2012) introduced sample-level enrichment analysis (SLEA). Drier et al. (2013) developed a state-of-the-art representation method named Pathifier. These methods are all capable of producing a score for every pathway/sample combination. However, in ssGSEA, genes are ranked by their expression values and the ESs are based on their ranks. In SLEA, genes are randomly permuted, and a pathway is scored by comparing the expression levels of its member genes before and after permutation. The Pathifier algorithm computes the pathway deregulation score (PDS) over all pathways one by one, and hence is computationally intensive. Other sophisticated tools have also been developed for calculating the individual-level pathway scores. For example, gene set variation analysis (GSVA) (Hänzelmann et al., 2013) obtains the gene ranks by fitting the gene-specific kernel functions and computes a Kolmogorov-Smirnov statistic, similar to ES. Individual-level pathway score (iPS) (Fang et al., 2020) computes the perturbation of a pathway at the individual-sample level with reference to normal samples. In contrast, iPath ranks genes on the basis of the magnitude of their departure from the overall expression levels across the tumor and normal samples, which improves quantification of the changes induced by experimental condition or disease status. As a result, iPath

CellPress
OPEN ACCESS

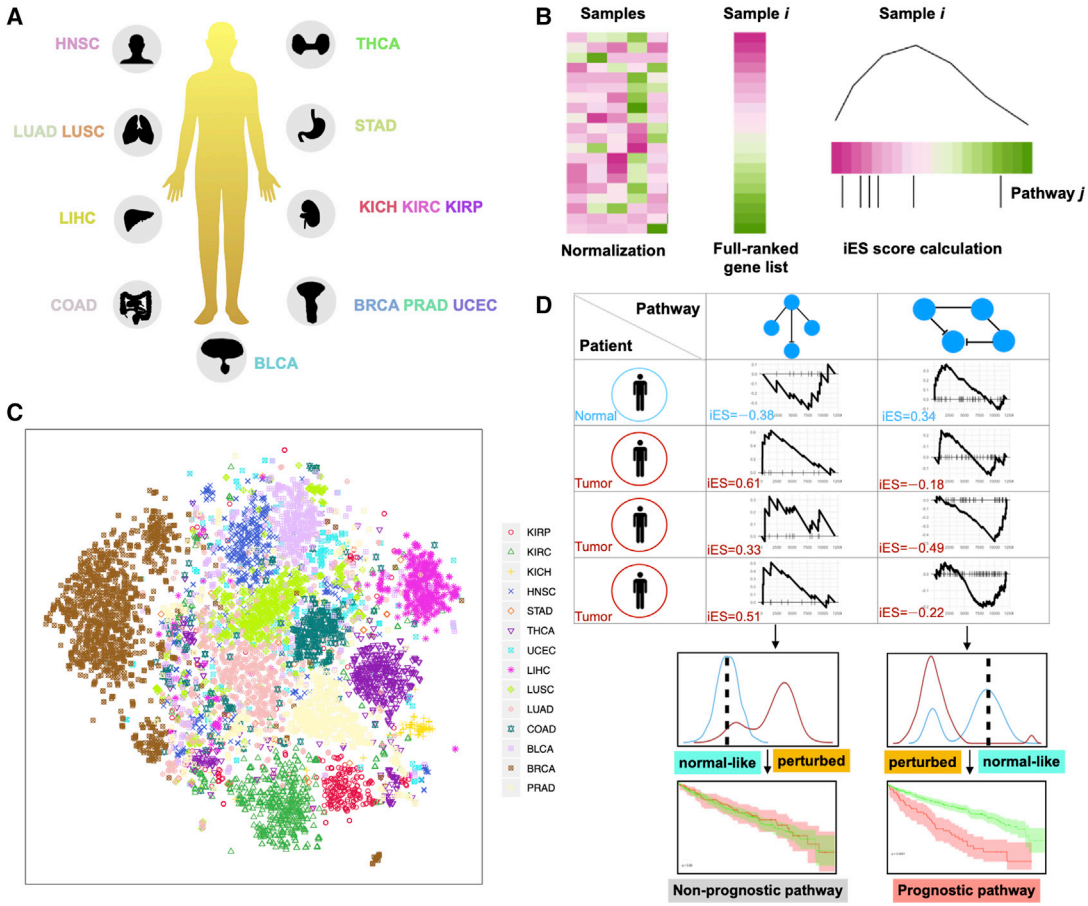


Figure 1. Overview of the iPath method

(A) The 14 cancer types analyzed in this pan-cancer study and the iPath workflow.

(B) Calculation of individual-level enrichment score (iES). iPath creates a full-ranked gene list for each sample. Given one pathway, it then projects the ranked gene list to the core of GSEA computation.

(C) t-SNE data visualization of the iES scores from all samples of the 14 cancer types. Abbreviations are as follows: KIRP, kidney renal papillary cell carcinoma; KIRC, kidney renal clear cell carcinoma; KICH, kidney chromophobe; HNSC, head and neck squamous cell carcinoma; STAD, stomach adenocarcinoma; THCA, thyroid carcinoma; UCEC, uterine corpus endometrial carcinoma; LIHC, liver hepatocellular carcinoma; LUSC, lung squamous cell carcinoma; LUAD, lung adenocarcinoma; COAD, colon adenocarcinoma; BLCA, bladder urothelial carcinoma; BRCA, breast invasive carcinoma; PRAD, prostate adenocarcinoma.

(D) Workflow of iPath, as demonstrated in the table with rows representing patients and columns representing pathways, iPath first calculates an iES score for each pathway and each sample (normal samples are indicated by blue circles and tumor samples by red circles). Then, for each pathway, iPath divides tumor samples as either normal-like or perturbed based on the iES scores. Finally, iPath performs survival analysis for the two tumor groups and identifies prognostic biomarker pathways based on survival analysis results.

is better at identifying disrupted pathways as prognostic biomarkers, which we demonstrate in the present study.

We applied iPath to perform a pan-cancer analysis by using well-established pathways and gene sets cataloged in the Molecular Signature Database (MSigDB) (Liberzon et al., 2011). Our results suggest that pathways are better options than single genes in terms of predicting clinical outcomes. Thus, we believe that prognostic pathways are promising and reliable biomarkers for precision oncology. Additional analyses further reveal that many of these prognostic biomarker pathways can be linked to frequently mutated cancer driver genes in a cancer-specific manner, illustrating the intricate interactions between somatic mutations, abnormal gene expression, and tumorigenesis.

RESULTS

Overview

We systematically explored the relationships between biological pathways or gene sets (referred simply as "pathways" hereafter for the sake of simplicity) and clinical outcomes in 14 solid cancer types (Figure 1A), using data available from TCGA (Table S1). These cancer types were selected because we require at least 20 matching normal samples in each cancer type. These normal samples are either normal or adjacent-normal tissues in the tumor patients.

We studied two major collections of pathways: C2 curated gene sets from MSigDB and Gene Ontology (GO) (Ashburner et al., 2000) (Table S1). There are 4,762, and 5,917 gene sets in these categories, respectively. Unlike most of the existing

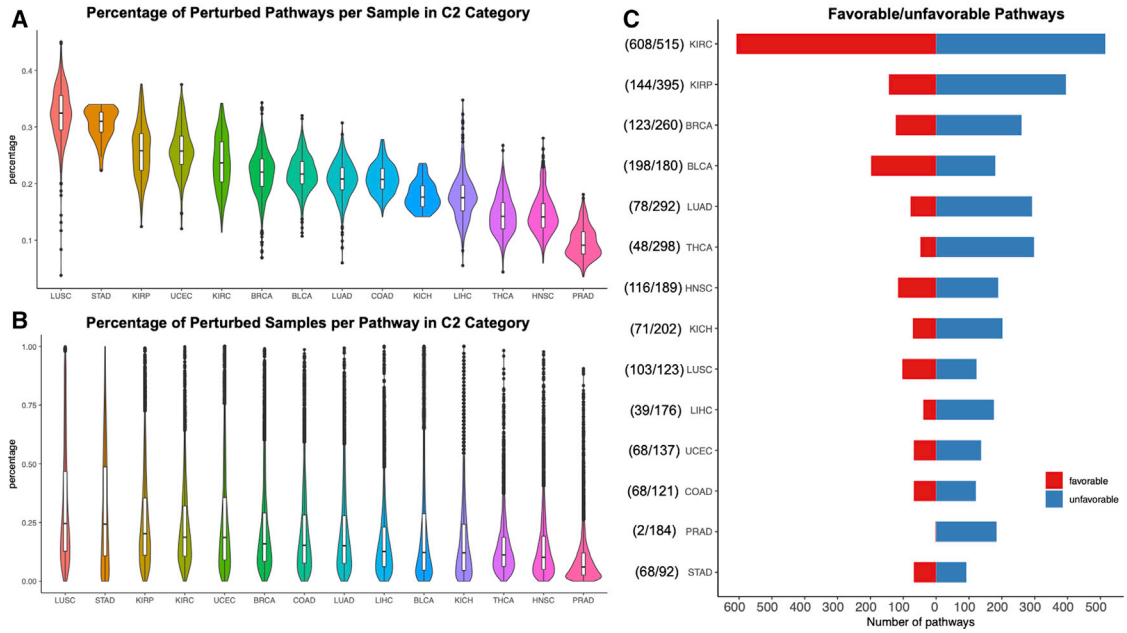


Figure 2. Survey of the proportions of perturbed pathways in the 14 cancer types

(A) Violin plots of parentage of perturbed pathways per tumor sample across 14 cancer types. The average proportions of the perturbed C2 category pathways among all tumor samples in the 14 cancer types are ranked from high to low.

(B) Violin plots of percentage of perturbed samples per pathway across 14 cancer types. The average proportions of perturbed tumor samples across all C2 category pathways in the 14 cancer types are ranked from high to low.

(C) Breakdown of favorable/unfavorable prognostic biomarker pathways in these 14 cancer types. Note that all analyses are performed by using the C2 category pathways, which includes 4,729 gene sets.

pathway-based studies (Li et al., 2019; Sanchez-Vega et al., 2018; Schubert et al., 2018; Wagle et al., 2018) that identify pathways with significant differences between the group of tumor samples and the group of normal samples, we intended to develop a method that focuses on pathway behavior at the individual patient level and to identify pathways in which departure from its norm has significant implication for patients' clinical outcomes. To achieve this, we developed a new computational approach named iPath. There are three major steps in iPath: First, for each individual patient and pathway, we calculate an individual-level ES (iES), analogous to the ES used in GSEA. Then, based on the iES, we dichotomize all tumor samples into two groups: normal-like and perturbed. Finally, we conduct survival analyses to compare whether the two groups of patients show differences in terms of their overall survival. Figure 1D illustrates the main workflow of iPath. We demonstrate that pathways identified by iPath have intimate connections with other biological and clinical properties, including somatic mutations, cancer subtypes, and pathology imaging features.

Furthermore, we investigated whether the expression pattern reflected in the pathway's iES values could illuminate the heterogeneity among different cancer types. Using the 4,762 gene sets from the C2 category, we plotted t-distributed stochastic neighbor embedding (t-SNE [Maaten and Hinton, 2008]) for all samples across 14 cancer types (Figure 1C). From the t-SNE plot, we observed that samples from the same tumor type (dots with the same color) tend to cluster together, indicating that iES values are highly informative in terms of the distinct pattern in their expression profiles. As expected, we found that three clusters of kidney cancer types—kidney renal papillary cell carcinoma (KIRP), kidney renal clear cell carcinoma (KIRC), and kidney chromophobe (KICH)—are located together, and two clusters of lung cancer types—lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD)—are located next to each other. Breast invasive carcinoma (BRCA) shows the greatest spread, and prostate adenocarcinoma (PRAD) shows multiple cluster formations indicating potential subtypes.

Identifying perturbed pathways

For a specific cancer type and a specific pathway, we classify each tumor sample as either normal-like or perturbed. The latter means the gene expression pattern of this pathway significantly deviates from that of a healthy, normal sample. We hypothesized that in any given tumor sample, multiple key pathways were perturbed. An important consideration is how many pathways are perturbed in a tumor sample and whether these numbers vary by tumor types. From our comprehensive survey on pathways belonging to the C2 category of MSigDB, we found that there was remarkable diversity among the 14 tumor types in terms of the average percentage of perturbed pathways per patient (Figure 2A). LUSC shows the highest proportions (32%) of perturbed pathways whereas PRAD shows the lowest proportions (9.6%). Interestingly, for the 14 tumor types, the proportions of tumor

samples showing perturbation averaged across pathways follow a similar order, but with much less variation among different tumor types (Figure 2B).

The MSigDB Hallmark gene set is a collection of 50 "refined" gene sets, curated from numerous "founder" sets, each representing a specific biological process or state and demonstrating coherent expression (Liberzon et al., 2015). The Hallmark set contains numerous well-known signaling pathways that have long been implicated in tumorigenesis and tumor progression, including the p53 pathway, Wnt, Notch, and PI3K pathways. It is of great interest to examine the expression pattern of these pathways at the individual tumor sample level. To achieve this, we applied iPath to the 50 pathways in the Hallmark category. For each of the 14 cancer types, we calculated the percentage of tumor samples that are perturbed for each Hallmark pathway (Figure S1A). As expected, we found that some pathways such as apoptosis and myogenesis are perturbed in more than half of the samples across multiple cancer types, whereas some other pathways, including PI3K, KRAS, and MTORC1, are perturbed in more than half of the samples in selected cancer types.

Identifying prognostic biomarker pathways

In this study, we applied iPath by using 10,679 gene sets to 6,198 tumor samples across 14 different cancer types. A pathway is named a prognostic biomarker pathway for a given cancer type if the Kaplan-Meier survival analysis yields a significant log-rank p value of less than 0.05. Here we used the same significance threshold used by Uhlen et al. (2017) to identify candidate prognostic genes. We later applied more stringent criteria to focus on the most promising prognostic biomarker pathways. Out of these 149,506 gene-set/cancer type combinations, 10,592 (7.1%) are deemed prognostic: 4,898 (7.3%) in the C2 category and 5,694 (6.9%) in the GO category. Tables S3 and S4 list the number of prognostic biomarker pathways by cancer type.

Among all the identified prognostic biomarker pathways, we further classified them by clinical outcomes into two subclasses: favorable prognostic biomarker pathways and unfavorable prognostic biomarker pathways. Favorable prognostic biomarker pathways imply that higher iES values than normal samples are correlated with better patient survival outcomes and vice versa. Unfavorable prognostic biomarker pathways designate the opposite. Among the 4,898 C2 pathway-cancer type combinations deemed significant in predicting patient outcome. 1,734 (35.4%) are favorable prognostic biomarker pathways and 3,164 (64.6%) are unfavorable prognostic biomarker pathways, respectively. The ratios of favorable to unfavorable prognostic biomarker pathways varied among the 14 different types of cancer. Figure 2C illustrates the number of prognostic biomarker pathways and the two subtypes for the 14 cancer types. The investigation of prognostic biomarkers for GO categories is summarized in Table S2.

To concentrate on the most promising results from this long list, we here present the most significant gene sets identified by iPath, using a combination of stringent criteria, including the q value (false discovery rate [FDR]) being less than 0.15 and the number of genes in the gene set being less than 100, in order to focus on more specific pathways. The breakdown of 1,473 (2.2%) and 1,541 (1.9%) significant prognostic biomarker pathways from C2 and GO categories in the 14 cancer types are summarized in Table S3. Excluding KIRC, which showed much more prognostic biomarker pathways than others, on average about 70 prognostic biomarker pathways (out of the total of 10,679 pathways, less than 1%) were found for each cancer type.

Given the diversity of clinical perspectives of various cancer types, in this pan-cancer study we choose not to use a very stringent threshold to ensure that we can select at least one prognostic pathway for every cancer type. To put this in perspective, in 6 out of the 14 cancer types we studied here, the p values corresponding to the 0.15 q value threshold are less than 0.001 (Table S3), which is the p-value threshold adopted in the study of prognostic marker genes by Uhlen et al. (2017).

Pan-cancer view on prognostic biomarker pathways identified

We examined the number of significant prognostic biomarker pathways identified among different cancer types (Table S3). We found that there was remarkable imbalance among these cancer types in terms of the number of such pathways identified. Most of the significant pathways were found in three kidney cancer types: KIRC, KIRP, and KICH. A few occurred for LUAD, PRAD, thyroid carcinoma, bladder urothelial carcinoma, and BRCA. Almost none were found in other cancer types. This could be because the clinical outcomes of different cancer types are quite diverse. It is also of interest to discover what proportions of the prognostic biomarker pathways overlap across cancer types. To this end, we calculated the Jaccard similarity between two lists of prognostic biomarker pathways for every pair of cancer types. We found that the similarity level is very low, except for the three kidney cancer types (KICH, KIRC, and KIRP), meaning that most cancer types have very few shared pathways (Figure S1B). In other words, the majority of prognostic biomarker pathways are cancer type specific. Our findings are consistent with the results presented in Uhlen et al. (2017) and highlight the extensive diversity in different types of human malignancy.

Compared with other cancer types, very few prognostic biomarker pathways were identified with breast cancer. This is somewhat surprising, given that multiple well-established pathways are known to play critical roles in the tumorigenesis and progression of breast cancer (Al-Hussaini et al., 2011; Criscitiello et al., 2015; Gasco et al., 2002; Johnson et al., 2016; King et al., 2012; Nagaraj and Ma, 2015; Witkiewicz and Knudsen, 2014). One possible reason for this is the substantial pathological differences among the four major subtypes of breast cancer: luminal A, luminal B, HER2$^+$, and basal-like. Supporting this hypothesis is the fact that the proportion of patients with such pathway alterations in these four breast cancer subtypes varies greatly (third column in Figures 3E and 3F). Given this observation, we were prompted to explore whether the disruption of a particular pathway preferentially occurs in a particular subtype of breast cancer. We then applied iPath to the four BRCA subtypes separately and identified 8, 10, 3, and 16 significant biomarker pathways (using FDR cutoff $q < 0.15$) in the four respective subtypes (Table S3).

Selected prognostic biomarker pathways identified

There were many interesting prognostic biomarker pathways identified by iPath. For example, in various kidney cancer types,

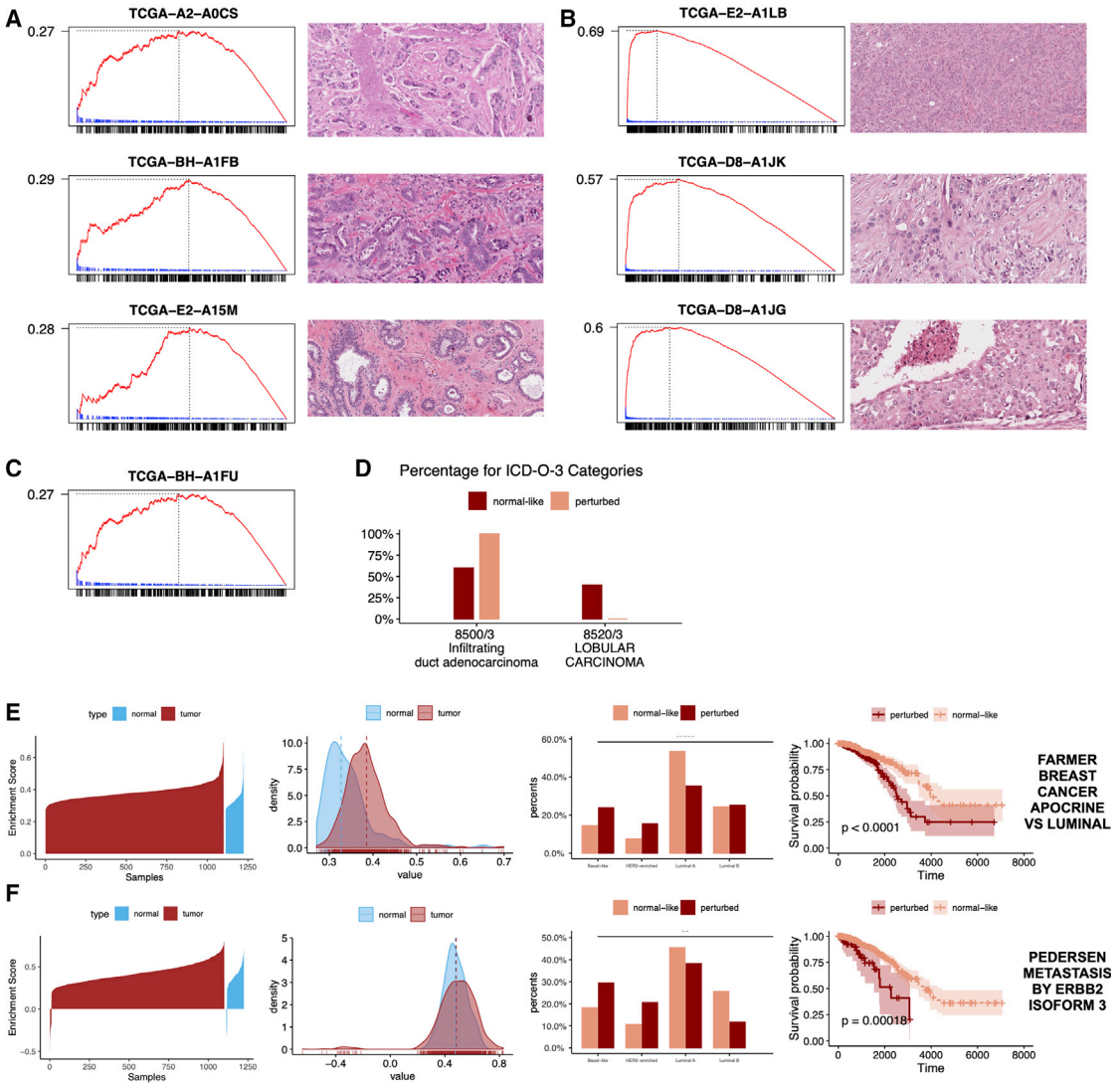


Figure 3. Demonstration of an example prognostic biomarker pathway (*FARMER BREAST CANCER APOCRINE VS LUMINAL*) in BRCA

(A) Enrichment plots of the pathway and corresponding pathology images of three samples labeled "normal-like."

(B) Enrichment plots and corresponding pathology images of three samples labeled "perturbed."

(C) Enrichment plot of the pathway of a normal sample.

(D) Breakdown of the ICD-O-3 categories for the top ten perturbed (highest iES value) and bottom ten normal-like (lowest iES values) patient samples.

(E and F) Visual summary of two example pathways. The waterfall plot shows the iES in tumor samples marked in red and normal samples marked in blue; the density plot shows that overall tumor samples are upregulated, because the mean of the tumor sample GSEA scores is higher than normal sample iES. The distribution of perturbed and normal-like tumors across the four subtypes of breast cancer is listed in the third column. The Kaplan-Meier plot indicates a significant survival difference for the perturbed and normal-like tumor samples.

including KIRP, KIRC, and KICH, many prognostic biomarker pathways from the GO collection in MSigDB were found to be related to the cell cycle (Figure S2A). Recent studies have shown that cell cycle progression gene signatures are significant, independent predictors of long-term outcomes for patients with renal clear cell carcinoma (Morgan et al., 2018) or related biomarkers (Chen et al., 2018). Smaller studies on TCGA KIRC datasets have substantiated this (Askeland et al., 2015; Gu et al., 2017). Our findings are also consistent with reports of cell-cycle-related biomarkers for KIRP (He et al., 2017) and KICH (Yin et al., 2018).

In BRCA, multiple REACTOME pathways were identified by iPath as prognostic biomarker pathways. For the *REACTOME_P38MAPK_EVENTS* pathway (Figure S2B and Table S3), our results are consistent with studies showing that p38 MAPK signaling drives resistance to key breast cancer drugs including trastuzumab resistance in HER2^{+} breast cancer (Donnelly et al., 2014) and tamoxifen resistance in luminal breast cancer (Jia et al., 2018). Identification of the *REACTOME_RAF_MAP_KINASE_CASCADE* pathway (Figure S2C and Table S3) as a biomarker is supported by a recent study that found that a

transcriptional signature called the MAPK pathway activity score (MPAS) is associated with patient outcome in ERBB2-positive breast cancer (Wagle et al., 2018). The prognostic nature of the gene set *FARMER_BREAST_CANCER_APOCRINE_VS_LUMINAL* (Figure 3E and Table S3) is logical, given the fact that this signature discriminates between AR^+ basal breast cancers with poor outcomes and AR^+ luminal breast cancers with much better outcomes (Farmer et al., 2005).

Besides the C2 category gene set database, we also identified GO term *GO_CELLULAR_RESPONSE_TO_THYROID_HORMONE_STIMULUS* (Figure S2D and Table S3), which contains 13 genes, as a prognostic biomarker pathway for KIRP. Thyroid hormone has long been linked to the pathophysiology of various cancer types (Krashin et al., 2019). Although this pathway is not one of the top enriched pathways according to classical GSEA analysis ($p = 0.2112$), iPath determined that a small subset of 22 KIRP patients with much reduced expression in this pathway led to significantly poor clinical prognosis, suggesting that any intervention that increases the impression of this pathway might benefit this group of patients. Another GO term that has been identified as a prognostic biomarker pathway is *GO_ATP_DEPENDENT_MICROTUBULE_MOTOR_ACTIVITY* (in KICH, Figure S2E and Table S3). Cell proliferation is a hallmark of almost all tumors, and it is well known that microtubules play an important role (Chandrasekaran et al., 2015) in mitosis. Interestingly, for this pathway we found that individuals with reduced expression levels have much better clinical prognoses, thus it is an unfavorable prognostic biomarker pathway. Given this, it is likely that antimitotic therapies that impede mitosis-specific microtubule functions through inhibiting motor proteins (Salmela and Kallio, 2013) might benefit patients with high expression of this gene set.

Links to distinct patterns shown in pathology imaging

Pathology imaging has long been regarded as the gold-standard diagnostic tool in clinical oncology. We conjectured that individual-level expression profiles of a pathway could help to distinguish subtle tumor characteristics hidden in pathology imaging. To investigate this, we used the gene set *FARMER_BREAST_CANCER_APOCRINE_VS_LUMINAL*, one of the most significant prognostic biomarker pathways identified in BRCA, as an example. We selected three tumor samples from the far end of both the normal-like group and the perturbed group and obtained their corresponding pathology images from the cancer digital slide archive (Gutman et al., 2013). The image of the three normal-like samples and three perturbed samples are shown in the second column in Figures 3A and 3B, respectively. Among the six pathology images, the luminal type tumor shows well-differentiated morphology with well-formed tumor lumen, low to intermediate nuclear grade, and low mitotic features. The androgen type shows higher grade, with poorly formed tumor lumen, intermediate to high nuclear grade, and focal tumoral necrosis. To confirm this observation, we obtained the ICD-O-3 codes (8500/3 Infiltrating duct adenocarcinoma; 8520/3 Lobular carcinoma) of the top ten and bottom ten samples patients quantified by their iESs. The breakdown of these codes shows a distinct distribution between normal-like and perturbed samples (Figure 3D).

Comparison with GSEA

The core function of iPath is to identify perturbed pathways in every individual tumor sample. In contrast, the classical GSEA method identifies pathways that show differences when comparing two groups of samples, hence only one ES is calculated for each pathway no matter how many samples there are. Given their differences, a pathway identified by iPath might not have been picked up by GSEA and vice versa. This is possibly because a pathway is perturbed only in one individual sample, and is thus unlikely to display a significant difference when tested by GSEA. In other words, iPath is good at identifying perturbed pathways for a small minority of cancer patients. To illustrate the point, we used breast cancer (BRCA) as an example. We first calculated iES for each pathway in each individual. Using iESs, we applied a Wilcoxon signed-rank test (Wilcoxon, 1945) to each pathway, compared iES values between tumor and normal samples, and used the p values of the test to rank all pathways. For comparison, we also ran GSEA to obtain a different list of ranked pathways. The top ten pathways that differentiate the iES values of tumor and normal samples are listed in Table S4 along with their significance levels. The top ten differentiated pathways identified by GSEA are listed in Table S4 along with the corresponding ranking in the Wilcoxon signed-rank test comparing iES values. We found that two pathways (bold) in the two top ten lists are identical; for the remaining eight pathways, four pathways in the GSEA list are not cancer related (red), whereas only two pathways in the iPath list seem not immediately cancer related (red).

Comparison with other sample-level gene set analysis methods

We compared iPath against existing methods that are capable of measuring expression of a pathway at individual level, namely ssGSEA (Barbie et al., 2009), SLEA (Gundem and Lopez-Bigas, 2012), Pathifier (Drier et al., 2013), and GSVA (Hänzelmann et al., 2013). We adopted the performance comparison study design used in the GSVA study wherein the effectiveness of clustering a mixture of tumor and normal samples is compared. In such a study, sample-level ES scores were used to select the most differentiated pathways, which in turn were used in the clustering. The details of the performance comparison procedure are presented in STAR Methods.

The performance comparison results are shown in Figure 4A. We use adjusted Rand index (ARI) to measure the clustering performance. Higher ARI indicates better clustering, which can be attributed to better pathways selected by each individual method that calculates sample-level ES scores. Figure 4A indicates that the iPath approach results in the highest ARI among all methods tested. Pairwise comparison between iPath and the four competing methods by using a t test indicates that all the differences are statistically significant.

We then compared these methods in terms of their ability to consistently detect prognostic biomarker pathways. In brief, for each method we selected the most significant pathway in the training data and tested its ability to predict survival in the test data by reporting the c-index. Higher c-index indicates better correlation with the survival outcomes. The results demonstrate the consistency of iPath for identifying the most

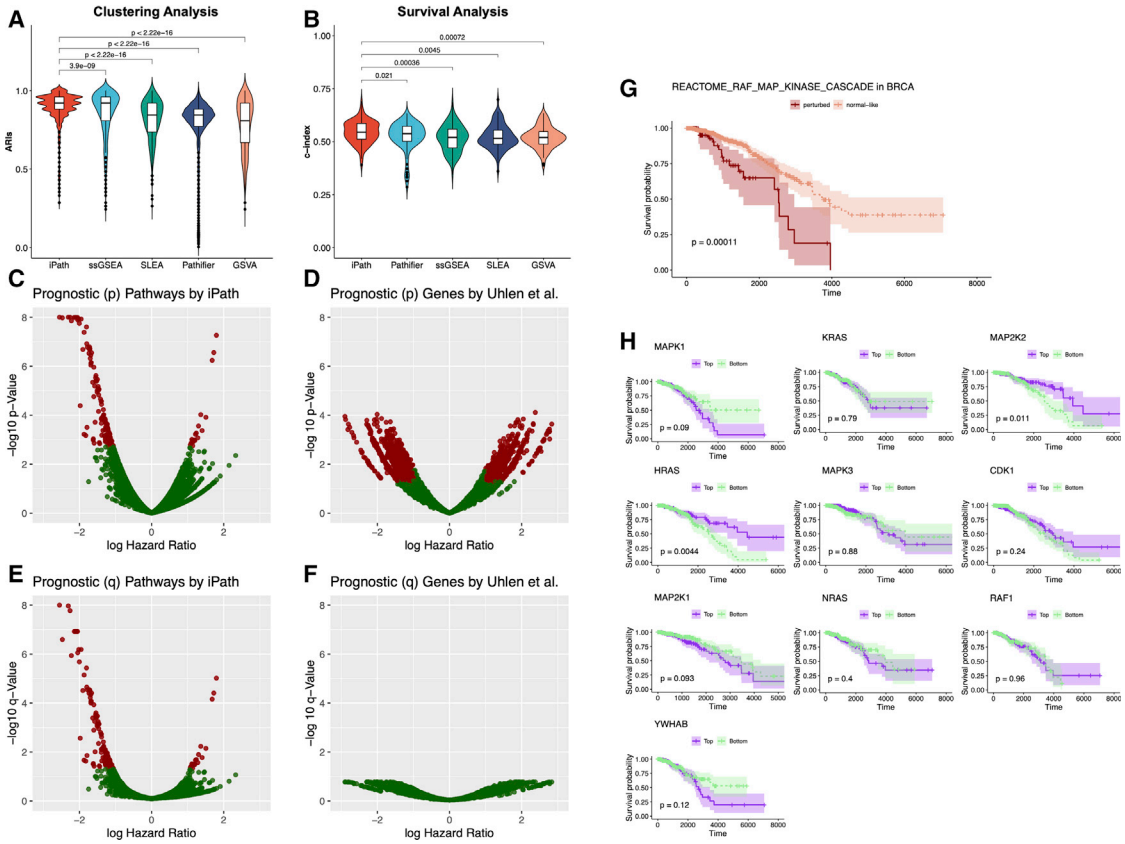


Figure 4. Comparisons between iPath and other sample-level gene set analysis methods including ssGSEA, SLEA, Pathifier, and GSVA, and comparisons between pathway biomarkers and individual-gene biomarkers

(A) Comparison of hierarchical clustering results in terms of separating tumor and normal samples from the ES matrix. The hierarchical clustering accuracy is measured by adjusted Rand index (ARI) value. As demonstrated in the violin plot, the clustering accuracy from iPath is significantly higher than that of the other methods.

(B) Comparison of survival analysis results using concordance index, showing that iPath can identify the most significant pathways that lead to the highest concordance in the violin plot.

(C) Volcano plots for the prognostic biomarker pathways. The significance threshold is set at a p value of 0.05 ($\log_{10}$(p value) = 1.4). The prognostic and non-prognostic biomarkers are marked by red and green dots, respectively.

(D) Volcano plots for the prognostic biomarker genes. The significance threshold is set at a p value of 0.05 ($\log_{10}$(p value) = 1.4).

(E) Volcano plots for the prognostic biomarker pathways. The significance threshold is set at a q value of 0.05 ($\log_{10}$(q value) = 1.4).

(F) Volcano plots for the prognostic biomarker genes. The significance threshold is set at a q value of 0.05 ($\log_{10}$(q value) = 1.4).

(G) Kaplan-Meier plot of prognostic biomarker pathway *REACTOME_RAF_MAP_KINASE_CASCADE* in BRCA.

(H) Kaplan-Meier plots of the member genes of the *REACTOME_RAF_MAP_KINASE_CASCADE* pathway in BRCA.

informative prognostic biomarker pathway across the training and test data. The details of the performance comparison procedures are presented in STAR Methods. The side-by-side box plots shown in Figure 4B again demonstrate the consistently good performance of iPath. Pairwise tests show that iPath produces significantly higher mean and median c-index values than competing methods.

Comparison with Pathifier

Among all the methods developed to calculate individual-level pathway scores, Pathifier is unique and has been widely applied to real studies, showing promising results (Liu et al., 2016; Mejía-Pedroza et al., 2018). Therefore, we conducted a direct comparison between iPath and Pathifier on a microarray dataset (Livshits et al., 2015) that had been previously analyzed by Pathifier. In this study, we analyzed 997 breast cancer tumor samples and 144 normal samples from the discovery set in the original study to determine which pathways were altered in the tumor samples. We first obtained iES scores from iPath and PDS scores from Pathifier for all samples and all 186 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. Given that these are microarray data, when calculating iES we only use normal samples to establish transcriptomic homeostasis. For each KEGG pathway, we then calculated a Wilcoxon signed-rank test comparing tumor samples with normal samples and ranked the pathways based on their p values. More details can be found in STAR Methods. We next manually curated the 186 KEGG pathways to select 52 of them (28.0%) that are considered breast cancer related (Table S5). Among these 52 pathways, we found 30 and 26 such pathways ranked among the top 50 by Pathifier and iPath,

respectively. Both enrichment results are statistically significant based on a binomial test (p values 2.2×10^{-6} and 2.8×10^{-4}, respectively). Pathifier identified four more cancer-related pathways than iPath but the difference is not statistically significant (Chi-squared test p value 0.546). From this result, we observed that on this microarray dataset, both methods successfully pick up breast cancer-related pathways and Pathifier performs moderately better than iPath.

Comparison with the Human Pathology Atlas

In a recent study, Uhlen et al. (2017) developed the Human Pathology Atlas (HPA), in which they adopted a system-level strategy to analyze 17 major cancer types with a focus on mining characteristic genes with respect to clinical outcomes. This method is based on genome-wide transcriptomic data and searches for prognostic genes whose top 20% or bottom 20% expression values, measured in fragments per kilobase of transcript per million mapped reads (FPKM), can stratify patient cohorts with significant survival differences ($p < 0.001$). Both HPA and iPath aim to identify prognostic biomarkers from transcriptome data. However, HPA relies on individual genes whereas iPath focuses on pathways. Hence, it is of great interest to compare their performance. Given the substantial noise that is ubiquitous in high-throughput technologies, we hypothesized that a pathway-based approach would be more robust and effective. To test our hypothesis, we applied both HPA and iPath to KIRP. First, we used the p-value threshold of 0.05 to determine whether a pathway or a gene would be considered prognostic by either approach (Figures 4C and 4D). Then, when using a more stringent threshold ($q = 0.05$), we found no significant prognostic biomarker genes (Figure 4F) but many significant prognostic biomarker pathways (Figure 4E). Tests conducted on KIRC gave similar results (Figure S3). These data indicate that the pathway-level biomarkers are more sensitive than the gene-level biomarkers.

A related question is whether member genes of a prognostic biomarker pathway are also prognostic biomarker genes. We found that this is not true in most cases. For some significant prognostic biomarker pathways identified by iPath, none of their member genes are prognostic genes according to HPA. In other words, at the individual-gene level, many genes are not prognostic biomarkers themselves, but their expression pattern as a whole can accurately predict a patient's clinical outcome. *REACTOME_RAF_MAP_KINASE_CASCADE*, for instance, is one of the significant biomarker pathways identified in BRCA (Figure 4G), but no gene inside this pathway correlates well with survival outcome (Figure 4H). This is reminiscent of the scenario in which a pathway is identified by GSEA as significant but none of its member genes show DE. Taking all of this together, we believe that pathway-based biomarkers are more robust and effective than single-gene-based biomarkers.

Connection with the mutations in cancer driver genes

Progressive accumulation of somatic mutations over time in crucial oncogenes or tumor-suppressor genes has been implicated in many cancer types (Martincorena and Campbell, 2015) (Kandoth et al., 2013; Leiserson et al., 2015; Zhang et al., 2018). Recently, the somatic mutation statuses of 127 genes have been shown to have significant effects on patient survival (Kandoth et al., 2013). With the identification of prognostic biomarker pathways using iPath, a natural question is whether the perturbed state of prognostic biomarker pathways is linked to somatic mutations occurring in cancer driver genes. To answer this, given a pathway and a cancer driver gene, we first constructed a contingency table dividing samples according to their normal-like/perturbed status for the pathway, and the mutation profile (present or absent) in the cancer gene. We then conducted a Fisher's exact test to identify the incidence of co-occurrence of the two events. A binary heatmap indicating whether a significant ($p < 0.05$, marked in the red block) connection between the top selected pathways and top mutated gene is shown in Figure 5. We found that indeed somatic mutation in key cancer driver genes and perturbed prognostic biomarker pathways are often co-occurring events. In breast cancer (BRCA), we observed that NOTCH1 and E-cadherin (CDH1) are associated with metastasis-related gene sets (Figure 5A), which is consistent with findings reported in the literature on NOTCH1 signaling (Leong et al., 2007) and CDH1 (Derksen et al., 2006; Ross et al., 2013). In LUAD (Figure 5B), we identified a couple of histone-lysine *N*-methyltransferase genes (MLL2 and MLL4) that are related to the top significant pathways found by iPath, and these genes are reportedly clustered in LUAD (Kandoth et al., 2013). We showed that PIK3CA is correlated with one early cell-cycle pathway, which demonstrates that PIK3CA deregulation serving as an early event precedes genome doubling in BRCA (Berenjeno et al., 2017) and colorectal adenocarcinoma (Carter et al., 2012).

Validation of top prognostic biomarker pathways

To verify that the prognostic biomarker pathways identified by iPath are indeed reliable biomarkers across studies, we use SurvExpress (Aguirre-Gamboa et al., 2013), an online biomarker validation tool, to check our top findings in multiple cancer types by using independent datasets (we used non-TCGA data whenever possible). SurvExpress employs the Cox model to estimate the prognostic index and draw Kaplan-Meier curves. Although using a different method and different datasets, we were able to confirm that all the top prognostic biomarker pathways are significant. The Kaplan-Meier curves, along with summary statistics of selected prognostic biomarker pathways, are presented in Figure S4.

Negative control experiments

To put the iPath results in context, we conducted negative control experiments to characterize whether and how often a random gene set demonstrates significant association between its status (normal-like or perturbed) and the patient's survival. We first randomly selected genes from the gene pool to form a hypothetical gene set. We then applied iPath to obtain an iES for each individual patient in this gene set, and test whether these scores can predict the patient's survival by using the same TCGA dataset. To ensure fair comparison, we let the sizes of the hypothetical gene sets match those of the established pathways. We repeated this process 1,000 times and reported the p values for the survival analysis. The reported p values are drawn on a histogram for each biomarker gene set in BRCA, KIRP, and KIRC in Figure S5. The results confirm that biomarker gene

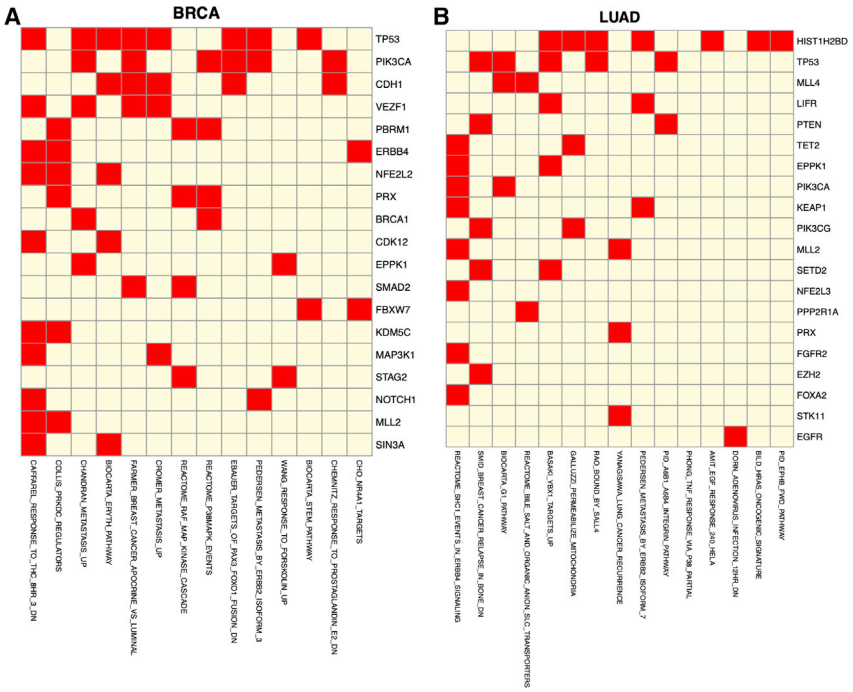


Figure 5. Association between prognostic biomarker pathways and somatic mutations of key cancer genes

(A) Matrix of gene/pathway association in BRCA. Red color indicates significant association.

(B) Matrix of gene/pathway association in LUAD. For each pathway, we classified each tumor sample as either normal-like or perturbed. For each gene, we classified each tumor sample as either mutated or not mutated. A Fisher's exact test of association was then carried out on the two-by-two contingency table.

sets are more significantly related to survival outcomes than random gene sets with the same sizes.

The iPath R package and iPath explorer Shiny app

To make iPath readily accessible for researchers and clinicians so that they can identify prognostic biomarker pathways for other types of cancer, we implemented the iPath pipeline into an easy-to-use R package named iPath. It is freely available at https://github.com/suke18/iPath.

In this study, more than 450,000 plots (waterfall plots, distribution plots, Kaplan-Meier plots) have been produced, corresponding to all 10,724 pathways across 14 cancer types. To assist researchers in better visualizing and exploring pathways of interest, we developed a web-based interface using Shiny app to organize and present all iPath results and make them easily browsable through the interface. All plots and summary statistics are publicly accessible and available at https://suke18.shinyapps.io/iPath. On the left-hand panel, a researcher can select cancer types and pathways of interest, after which iPath will run on-the-fly and the results of the survival analysis will be displayed across four main panels (waterfall plot, density plot, Kaplan-Meier plot, and a table of detailed summary statistics).

DISCUSSION

Here, we describe iPath, a computational tool to identify perturbed pathways found in individual tumor samples. Unlike individual genes, the collection of functionally related genes in a pre-defined pathway provides a more robust assessment of the changes that affect key biological functions in tumor samples. The advantages of using pathways over individual genes have been well documented in the analysis of noisy high-throughput data (Reimand et al., 2019; Vaske et al., 2010) and as biomarkers (Zheng et al., 2020). What makes iPath unique is its ability to provide such an assessment one sample at a time. This is significant, because substantial heterogeneity among tumor genomes suggests that it is common for a critical pathway to be perturbed in only a few tumor samples. As a result, it is highly unlikely that these pathways will be identified by traditional GSEA. On the other hand, iPath can identify perturbed pathways, even if such disruption only occurs in a small subset of tumor samples. In short, iPath promises to improve patient care by enabling oncologists to develop more effective personalized treatment strategies with fewer side effects.

To demonstrate the effectiveness of iPath, we conducted a comprehensive pan-cancer study across 14 different cancer types with more than 6,000 tumor samples. For each cancer type, iPath identified about 70 prognostic biomarker pathways on average, many of them showing promising biological interpretations. We also validated the top prognostic biomarker pathways by using SurvExpress, an online biomarker validation tool. There are two types of prognostic biomarker pathways: favorable and unfavorable. Favorable pathways account for one-third of all the biomarker pathways. These pathways can be used to identify patients with better prognostics so that they can be spared unnecessary adjuvant therapies.

Our pan-cancer study using iPath yields two interesting results. First, we found that quite a few pathways or gene sets are potential prognostic biomarkers for most of the cancer types we studied. However, for the vast majority of these biomarker pathways, they are perturbed in only a small fraction of all the patients. Second, for any given pair of cancer types, there is little overlap among the two lists of prognostic biomarker pathways. Our findings highlight the fact that cancer is a highly heterogeneous disease and, therefore, a personalized treatment strategy is the key to effective care for cancer patients. Although the present study is conducted on RNA-seq gene expression profiles, iPath can be applied to other omics data such as microarray data.

The core of iPath is the iES, a single continuous value between -1 and 1, calculated for every pathway and every individual sample, which evaluates the spread of the member genes

belonging to a pathway among all genes in the genome ranked by its level of deviation from the norm. We believe this to be a powerful way to summarize the status of a pathway or provide a big-picture view of pathway changes at single-sample resolution. Our analysis has shown that iES is informative and sometimes predictive of patients' clinical features and prospects because it measures the level of the pathway's deviation from the norm in a holistic manner. When testing on TCGA RNA-seq data, we find iPath to be more sensitive than other methods including ssGSEA (Barbie et al., 2009), Pathifier (Drier et al., 2013), SLEA (Gundem and Lopez-Bigas, 2012), and GSVA (Hänzelmann et al., 2013). Although iPath is originally developed to analyze RNA-seq data, it can be applied to microarray data as well. Even though iPath performs slightly worse than Pathifier when tested on a microarray dataset (Livshits et al., 2015), the fact that it performs competitively with a state-of-the-art method on microarray datasets is still encouraging.

Given that scientists have already identified many biomarker genes for various cancer types—for example, thousands of prognostic genes have been identified in a recent study by Uhlen et al. (2017)—why it is important to identify prognostic biomarker pathways? In the present study, we found that compared with single-gene biomarkers, pathway-based biomarkers are more robust with better separation power, which gives clinicians more confidence in separating patients into different risk groups and to assign treatment strategies accordingly. Furthermore, given that they represent well-curated biological pathways they are easier to interpret, and hence more likely to be informative and meaningful to clinicians. Another key advantage of pathway-based biomarkers is that there are drugs that especially target specific pathways. For example, it is likely that MAPK perturbed patients will benefit more from MAPK inhibitor drugs. As this is beyond the scope of our current study, we plan to pursue this in future works.

iPath can be applied broadly to other types of cancer, for any given individual sample, as long as there are corresponding normal samples that can be used as controls. Thus, iPath might be a powerful resource for unraveling the paradigm shift that occurs in a small minority of samples. Therefore, it is an ideal tool for precision oncology. By way of illustrating its potential, some drugs have been developed to specifically target a kinase and its downstream genes (Cabanillas et al., 2019; Pal et al., 2010). Using iPath, we can group the drug target and its downstream genes together and identify patients with elevated expression in this gene set; such patients might benefit the most from this targeted therapy. We believe iPath can potentially provide fresh perspectives on patient selection and prognostic prediction.

Limitations of the study

In this study, we only examined individual pathways to try to establish whether a given pathway is predictive of a clinical outcome. For prediction purposes, we could consider multiple pathways jointly, which might produce better prediction performance. This represents one potential future research direction for the continuous development of iPath. Moreover, iPath has been developed for analysis of RNA-seq data. Although it can be applied to other types of data such as microarray data, other state-of-the-art methods such as Pathifier might be preferable.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Overview of the iPath approach
  - Calculation of iES
  - Definition of perturbed tumor samples
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Performance comparison among sample-level gene set analysis methods
  - Microarray data analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2021.100050.

ACKNOWLEDGMENTS

We thank Dr. Ya Wang for helpful input. We thank Noah Rawlings for careful editing of the manuscript. This research was supported by NIH/NCI grant (U01CA217875) to C.M., and NIH/NCI grants (R01CA223220, R01CA245386 and UG1CA233259) to S.S. Z.S.Q. is partially supported by NIH/NHLBI (R01AI145231).

AUTHOR CONTRIBUTIONS

This project was conceived by Z.S.Q. K.S. implemented and improved the method and ran all the analyses. Q.Y. performed method comparison analysis. C.S.M., X.L., R.S., and S.-Y.S. made key suggestions to improve the method and the overall design of the study. All authors discussed the results and contributed to the writing of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: December 16, 2020
Revised: May 7, 2021
Accepted: June 16, 2021
Published: July 23, 2021

REFERENCES

Aguirre-Gamboa, R., Gomez-Rueda, H., Martínez-Ledesma, E., Martínez-Torteya, A., Chacolla-Huaringa, R., Rodriguez-Barrientos, A., Tamez-Peña, J.G., and Treviño, V. (2013). SurvExpress: an online biomarker validation tool and database for cancer gene expression data using survival analysis. PLoS One *8*, e74250.

Al-Hussaini, H., Subramanyam, D., Reedijk, M., and Sridhar, S.S. (2011). Notch signaling pathway as a therapeutic target in breast cancer. Mol. Cancer Ther. *10*, 9–15.

Arya, R., and White, K. (2015). Cell death in development: signaling pathways and core mechanisms. Semin. Cell Dev. Biol. *39*, 12–19.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology:

tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. *25*, 25–29.

Askeland, E.J., Chehval, V.A., Askeland, R.W., Fosso, P.G., Sangale, Z., Xu, N., Rajamani, S., Stone, S., and Brown, J.A. (2015). Cell cycle progression score predicts metastatic progression of clear cell renal cell carcinoma after resection. Cancer Biomark. Sect. Dis. Mark. *15*, 861–867.

Barbie, D.A., Tamayo, P., Boehm, J.S., Kim, S.Y., Moody, S.E., Dunn, I.F., Schinzel, A.C., Sandy, P., Meylan, E., Scholl, C., et al. (2009). Systematic RNA interference reveals that oncogenic KRAS -driven cancers require TBK1. Nature *462*, 108–112.

Berenjeno, I.M., Piñeiro, R., Castillo, S.D., Pearce, W., McGranahan, N., Dewhurst, S.M., Meniel, V., Birkbak, N.J., Lau, E., Sansregret, L., et al. (2017). Oncogenic PIK3CA induces centrosome amplification and tolerance to genome doubling. Nat. Commun. *8*, 1173.

Cabanillas, M.E., Ryder, M., and Jimenez, C. (2019). Targeted therapy for advanced thyroid cancer: kinase inhibitors and beyond. Endocr. Rev. *40*, 1573–1604.

Cantini, L., Calzone, L., Martignetti, L., Rydenfelt, M., Blüthgen, N., Barillot, E., and Zinovyev, A. (2018). Classification of gene signatures for their information value and functional redundancy. NPJ Syst. Biol. Appl. *4*, 2.

Carter, S.L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P.W., Onofrio, R.C., Winckler, W., Weir, B.A., et al. (2012). Absolute quantification of somatic DNA alterations in human cancer. Nat. Biotechnol. *30*, 413–421.

Chandrasekaran, G., Tátrai, P., and Gergely, F. (2015). Hitting the brakes: targeting microtubule motors in cancer. Br. J. Cancer *113*, 693–698.

Chen, L., Yuan, L., Qian, K., Qian, G., Zhu, Y., Wu, C.-L., Dan, H.C., Xiao, Y., and Wang, X. (2018). Identification of biomarkers associated with pathological stage and prognosis of clear cell renal cell carcinoma by Co-expression network analysis. Front. Physiol. *9*, 399.

Criscitiello, C., Esposito, A., De Placido, S., and Curigliano, G. (2015). Targeting fibroblast growth factor receptor pathway in breast cancer. Curr. Opin. Oncol. *27*, 452–456.

Dang, H., Pomyen, Y., Martin, S.P., Dominguez, D.A., Yim, S.Y., Lee, J.-S., Budhu, A., Shah, A.P., Bodzin, A.S., and Wang, X.W. (2019). NELFE-dependent MYC signature identifies a unique cancer subtype in hepatocellular carcinoma. Sci. Rep. *9*, 3369.

Derksen, P.W.B., Liu, X., Saridin, F., van der Gulden, H., Zevenhoven, J., Evers, B., van Beijnum, J.R., Griffioen, A.W., Vink, J., Krimpenfort, P., et al. (2006). Somatic inactivation of E-cadherin and p53 in mice leads to metastatic lobular mammary carcinoma through induction of anoikis resistance and angiogenesis. Cancer Cell *10*, 437–449.

Donnelly, S.M., Paplomata, E., Peake, B.M., Sanabria, E., Chen, Z., and Nahta, R. (2014). P38 MAPK contributes to resistance and invasiveness of HER2-overexpressing breast cancer. Curr. Med. Chem. *21*, 501–510.

Drier, Y., Sheffer, M., and Domany, E. (2013). Pathway-based personalized analysis of cancer. Proc. Natl. Acad. Sci. U S A *110*, 6388–6393.

Fang, J., Pian, C., Xu, M., Kong, L., Li, Z., Ji, J., Chen, Y., and Zhang, L. (2020). Revealing prognosis-related pathways at the individual level by a comprehensive analysis of different cancer transcription data. Genes *11*, 1281.

Farmer, P., Bonnefoi, H., Becette, V., Tubiana-Hulin, M., Fumoleau, P., Larsimont, D., Macgrogan, G., Bergh, J., Cameron, D., Goldstein, D., et al. (2005). Identification of molecular apocrine breast tumours by microarray analysis. Oncogene *24*, 4660–4671.

Gasco, M., Shami, S., and Crook, T. (2002). The p53 pathway in breast cancer. Breast Cancer Res. *4*, 70–76.

Gu, Y., Lu, L., Wu, L., Chen, H., Zhu, W., and He, Y. (2017). Identification of prognostic genes in kidney renal clear cell carcinoma by RNA-seq data analysis. Mol. Med. Rep. *15*, 1661–1667.

Gundem, G., and Lopez-Bigas, N. (2012). Sample-level enrichment analysis unravels shared stress phenotypes among multiple cancer types. Genome Med. *4*, 28.

Gutman, D.A., Cobb, J., Somanna, D., Park, Y., Wang, F., Kurc, T., Saltz, J.H., Brat, D.J., and Cooper, L.A.D. (2013). Cancer Digital Slide Archive: an informatics resource to support integrated in silico analysis of TCGA pathology data. J. Am. Med. Inform. Assoc. *20*, 1091–1098.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646–674.

Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics *14*, 7.

He, Z., Sun, M., Ke, Y., Lin, R., Xiao, Y., Zhou, S., Zhao, H., Wang, Y., Zhou, F., and Zhou, Y. (2017). Identifying biomarkers of papillary renal cell carcinoma associated with pathological stage by weighted gene co-expression network analysis. Oncotarget *8*, 27904–27914.

Jia, Y., Zhou, J., Luo, X., Chen, M., Chen, Y., Wang, J., Xiong, H., Ying, X., Hu, W., Zhao, W., et al. (2018). KLF4 overcomes tamoxifen resistance by suppressing MAPK signaling pathway and predicts good prognosis in breast cancer. Cell. Signal. *42*, 165–175.

Joe, S., and Nam, H. (2016). Prognostic factor analysis for breast cancer using gene expression profiles. BMC Med. Inform. Decis. Mak. *16* (*Suppl. 1*), 56.

Johnson, J., Thijssen, B., McDermott, U., Garnett, M., Wessels, L.F.A., and Bernards, R. (2016). Targeting the RB-E2F pathway in breast cancer. Oncogene *35*, 4829–4835.

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., et al. (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333–339.

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. *28*, 27–30.

Khatri, P., Sirota, M., and Butte, A.J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput. Biol. *8*, e1002375.

King, T.D., Suto, M.J., and Li, Y. (2012). The Wnt/β-catenin signaling pathway: a potential therapeutic target in the treatment of triple negative breast cancer. J. Cell. Biochem. *113*, 13–18.

Kosinski, M., Biecek, P., and Chodor, W. (2016). The Family of R Packages Containing TCGA Data. https://rtcga.github.io/RTCGA/.

Krashin, E., Piekiełko-Witkowska, A., Ellis, M., and Ashur-Fabian, O. (2019). Thyroid hormones and cancer: a comprehensive review of preclinical and clinical studies. Front. Endocrinol. *10*, 59.

Leiserson, M.D.M., Vandin, F., Wu, H.-T., Dobson, J.R., Eldridge, J.V., Thomas, J.L., Papoutsaki, A., Kim, Y., Niu, B., McLellan, M., et al. (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat. Genet. *47*, 106–114.

Leong, K.G., Niessen, K., Kulic, I., Raouf, A., Eaves, C., Pollet, I., and Karsan, A. (2007). Jagged1-mediated Notch activation induces epithelial-to-mesenchymal transition through Slug-induced repression of E-cadherin. J. Exp. Med. *204*, 2935–2948.

Li, Q., Brown, J.B., Huang, H., and Bickel, P.J. (2011). Measuring reproducibility of high-throughput experiments. Ann. Appl. Stat. *5*, 1752–1779.

Li, S., Song, Y., Quach, C., Guo, H., Jang, G.-B., Maazi, H., Zhao, S., Sands, N.A., Liu, Q., In, G.K., et al. (2019). Transcriptional regulation of autophagy-lysosomal function in BRAF-driven melanoma progression and chemoresistance. Nat. Commun. *10*, 1693.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics *27*, 1739–1740.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. Cell Syst. *1*, 417–425.

Liu, C., Srihari, S., Lal, S., Gautier, B., Simpson, P.T., Khanna, K.K., Ragan, M.A., and Lê Cao, K.-A. (2016). Personalised pathway analysis reveals association between DNA repair pathway dysregulation and chromosomal instability in sporadic breast cancer. Mol. Oncol. *10*, 179–193.

Livshits, A., Git, A., Fuks, G., Caldas, C., and Domany, E. (2015). Pathway-based personalized analysis of breast cancer expression data. Mol. Oncol. *9*, 1471–1483.

Maaten, L.van der, and Hinton, G. (2008). Visualizing data using t-SNE. J. Mach. Learn. Res. *9*, 2579–2605.

Martincorena, I., and Campbell, P.J. (2015). Somatic mutation in cancer and normal cells. Science *349*, 1483–1489.

Mejía-Pedroza, R.A., Espinal-Enríquez, J., and Hernández-Lemus, E. (2018). Pathway-based drug repositioning for breast cancer molecular subtypes. Front. Pharmacol. *9*, 905.

Morgan, T.M., Mehra, R., Tiemeny, P., Wolf, J.S., Wu, S., Sangale, Z., Brawer, M., Stone, S., Wu, C.-L., and Feldman, A.S. (2018). A multigene signature based on cell cycle proliferation improves prediction of mortality within 5 Yr of radical nephrectomy for renal cell carcinoma. Eur. Urol. *73*, 763–769.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods *5*, 621–628.

Nagaraj, G., and Ma, C. (2015). Revisiting the estrogen receptor pathway and its role in endocrine therapy for postmenopausal women with estrogen receptor-positive metastatic breast cancer. Breast Cancer Res. Treat. *150*, 231–242.

Pal, S.K., Figlin, R.A., and Reckamp, K. (2010). Targeted therapies for non-small cell lung cancer: an evolving landscape. Mol. Cancer Ther. *9*, 1931–1944.

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C.E., Socci, N.D., and Betel, D. (2013). Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol. *14*, 3158.

Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., Wadi, L., Meyer, M., Wong, J., Xu, C., et al. (2019). Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nat. Protoc. *14*, 482–517.

Ross, J.S., Wang, K., Sheehan, C.E., Boguniewicz, A.B., Otto, G., Downing, S.R., Sun, J., He, J., Curran, J.A., Ali, S., et al. (2013). Relapsed classic E-cadherin (CDH1)-mutated invasive lobular breast cancer shows a high frequency of HER2 (ERBB2) gene mutations. Clin. Cancer Res. *19*, 2668–2676.

Salmela, A.-L., and Kallio, M.J. (2013). Mitosis as an anti-cancer drug target. Chromosoma *122*, 431–449.

Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W.K., Luna, A., La, K.C., Dimitriadoy, S., Liu, D.L., Kantheti, H.S., Saghafinia, S., et al. (2018). Oncogenic signaling pathways in the cancer genome atlas. Cell *173*, 321–337.e10.

Schmelzle, T., and Hall, M.N. (2000). TOR, a central controller of cell growth. Cell *103*, 253–262.

Schubert, M., Klinger, B., Klünemann, M., Sieber, A., Uhlitz, F., Sauer, S., Garnett, M.J., Blüthgen, N., and Saez-Rodriguez, J. (2018). Perturbation-response genes reveal signaling footprints in cancer gene expression. Nat. Commun. *9*, 20.

Scrucca, L., Fop, M., Murphy, T.B., and Raftery, A.E. (2016). mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. R. J. *8*, 289–317.

Soneson, C., and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics *14*, 91.

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.

The International Cancer Genome Consortium (2010). International network of cancer genome projects. Nature *464*, 993–998.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515.

Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., Benfeitas, R., Arif, M., Liu, Z., Edfors, F., et al. (2017). A pathology atlas of the human cancer transcriptome. Science *357*, eaan2507.

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.

van de Vijver, M.J., He, Y.D., van 't Veer, L.J., Dai, H., Hart, A.A.M., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J., et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. N. Engl. J. Med. *347*, 1999–2009.

Wagle, M.-C., Kirouac, D., Klijn, C., Liu, B., Mahajan, S., Junttila, M., Moffat, J., Merchant, M., Huw, L., Wongchenko, M., et al. (2018). A transcriptional MAPK Pathway Activity Score (MPAS) is a clinically relevant biomarker in multiple cancer types. NPJ Precis. Oncol. *2*, 7.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biom. Bull. *1*, 80–83.

Witkiewicz, A.K., and Knudsen, E.S. (2014). Retinoblastoma tumor suppressor pathway in breast cancer: prognosis, precision medicine, and therapeutic interventions. Breast Cancer Res. *16*, 207.

Xu, Q., Chen, J., Ni, S., Tan, C., Xu, M., Dong, L., Yuan, L., Wang, Q., and Du, X. (2016). Pan-cancer transcriptome analysis reveals a gene expression signature for the identification of tumor tissue origin. Mod. Pathol. *29*, 546–556.

Yin, X., Wang, J., and Zhang, J. (2018). Identification of biomarkers of chromophobe renal cell carcinoma by weighted gene co-expression network analysis. Cancer Cell Int. *18*, 206.

Zhang, W., and Liu, H.T. (2002). MAPK signal pathways in the regulation of cell proliferation in mammalian cells. Cell Res. *12*, 9–18.

Zhang, Y., Yang, L., Kucherlapati, M., Chen, F., Hadjipanayis, A., Pantazi, A., Bristow, C.A., Lee, E.A., Mahadeshwar, H.S., Tang, J., et al. (2018). A pan-cancer compendium of genes deregulated by somatic genomic rearrangement across more than 1,400 cases. Cell Rep. *24*, 515–527.

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., and Liu, X. (2014). Comparison of RNA-seq and microarray in transcriptome profiling of activated T cells. PLoS One *9*, e78644.

Zheng, X., Amos, C.I., and Frost, H.R. (2020). Comparison of pathway and gene-level models for cancer prognosis prediction. BMC Bioinformatics *21*, 76.

Zuo, S., Zhang, X., and Wang, L. (2019). A RNA sequencing-based six-gene signature for survival prediction in patients with glioblastoma. Sci. Rep. *9*, 2615.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Level 3 RNA-Seq data	TCGA data maintained by Broad Institute GDAC Firehose.	https://gdac.broadinstitute.org/
Clinical data	TCGA data maintained by Broad Institute GDAC Firehose.	https://gdac.broadinstitute.org/
C2 collection of pathways	MSigDB	https://www.gsea-msigdb.org/gsea/msigdb/
GO collection of pathways	MSigDB	https://www.gsea-msigdb.org/gsea/msigdb/
Analytical results	This paper	https://suke18.shinyapps.io/iPath
Software and algorithms		
RTCGA	Kosinski et al. (2016) Bioconductor package	https://www.bioconductor.org/packages/release/bioc/html/RTCGA.html
Mclust	CRAN R package	https://CRAN.R-project.org/package=mclust
iPath pipeline	This paper	https://github.com/suke18/iPath

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Zhaohui S. Qin (zhaohui.qin@emory.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Code and shiny app

The iPath pipeline implemented in an R package is freely available at https://github.com/suke18/iPath. All the analytical results about the MSigDB C2 collection and GO terms corresponding to 14 cancer types are publicly accessible and available at https://suke18.shinyapps.io/iPath.

METHOD DETAILS

Overview of the iPath approach

The goal of iPath is to identify pathways that show unusual patterns at single sample-level. To achieve this, we defined a novel statistics, namely iES. For each pathway and each given patient, iPath first computes iES, a single value that reflect the overall expression behavior of this pathway in this sample relative to the population. Such a method allows us to quantify the level of irregularity for a set of genes in a single sample. Next, for each pathway, using normal samples' distribution of the iESs (Figure 1D), we come up with an iES threshold which we use to classify all tumor samples into either the category of normal-like or perturbed. Last, we compare the survival difference between these two groups, and designate a pathway as a prognostic biomarker pathway if the two groups of patients show significant difference in overall survival (Figure 1D).

Calculation of iES

For each cancer type, we denote the RNA-seq expression matrix as $Y = \{y_{ij}\}$, with rows corresponding to the samples and columns corresponding to the genes, and $i = 1,\ldots,M$, and $j = 1,\ldots,N$, M is the total number of samples and N is the total number of genes in the genome. Here we consider both tumor and normal samples. For the TCGA data used in this study, there are matched tumor, normal sample pairs. Because such normal samples are scarce, we do not take the pairing information into consideration.

The expression levels Y are assumed to have already been normalized, for example, measured by FPKM or RPKM values (Mortazavi et al., 2008; Trapnell et al., 2010). We first use all the samples of this cancer type to construct a transcriptomic homeostasis, calculate the mean ($\bar{y}_j$) and standard deviation (s_j) of the expression level for every gene in the genome. Because the number of normal samples in TCGA is very limited and there are many zeros in RNA-seq data, so we decide against using just normal samples to define transcriptomic homeostasis in order to avoid numerical instability. For microarray data, or if the number of normal samples is abundant, we recommend using just the normal sample to construct the transcriptomic homeostasis.

Next, for each sample, for sample i, we calculate iES for every pathway (Figure 1B) as follows.

1. Calculate z-score $\boldsymbol{z_{ij}} = \frac{Y_{ij} - \bar{y}_j}{s_j}$ for every gene, here $\boldsymbol{z_{ij}}$ represents the level of deviation from the norm for gene $\boldsymbol{j}$ in the $\boldsymbol{i}$th sample, $\boldsymbol{i} = \mathbf{1},\ldots,\boldsymbol{M}$, and $\boldsymbol{j} = \mathbf{1},\ldots,\boldsymbol{N}$.
2. Next, sort the absolute value of $\boldsymbol{z_{ij}}$, denoted as $|\boldsymbol{z_{ij}}|$, in descending order to obtain the ranks of all genes in the genome, denoted as $\{\boldsymbol{g_{i1}}, \boldsymbol{g_{i2}}, \ldots, \boldsymbol{g_{iN}}\}$ such that $|\boldsymbol{z_{ig_{i1}}}| \geq |\boldsymbol{z_{ig_{i2}}}| \geq \ldots \geq |\boldsymbol{z_{ig_{iN}}}|$.
3. Subject the sorted gene list $\{\boldsymbol{g_{i1}}, \boldsymbol{g_{i2}}, \ldots, \boldsymbol{g_{iN}}\}$to the GSEA analysis: given one pathway ($\boldsymbol{S}$) including $\boldsymbol{R}$ genes, iPath loops through the sorted gene list $\{\boldsymbol{g_{i1}}, \boldsymbol{g_{i2}}, \ldots, \boldsymbol{g_{iN}}\}$and calculates a running sum (Kolmogorov–Smirnov) statistics $\boldsymbol{iES_i}$ for $\boldsymbol{i^{th}}$ sample in the following manner: if the $\boldsymbol{g_j}$ is not in $\boldsymbol{S}$, then subtract a penalty score $\frac{1}{N-R}$; If the $\boldsymbol{g_j}$ is in $\boldsymbol{S}$, then add a n incremental score $\frac{|z_{ij}|}{\sum_{j\in S}|z_{ij}|}$. By aggregating the scores from each position, it computes the $\boldsymbol{iES_{ip}}$ value at the $\boldsymbol{p^{th}}$ position in $\boldsymbol{L^i}$ as:

$$P_{increments}(S,\ p) = \sum_{\substack{g_j \in S \\ j \leq p}} \frac{|Z_{ij}|}{S_R},\ where\ S_R = \sum_{g_j \in S} |Z_{ij}|$$

$$P_{penalities}\ (S,\ p) = \sum_{\substack{g_j \notin S \\ j \leq p}} \frac{1}{N-R}$$

The iES score for i^{th} sample acquires the maximum deviation from zero of $P_{increments} - P_{penalities}$. It is worth noting that utilizing $|Z_{ij}|$ for the i^{th} sample allows for the estimation of the leading contribution of the most perturbed genes.

Definition of perturbed tumor samples

For each pathway, we classified each tumor sample as either normal-like or perturbed. Perturbed means a significant departure from the expression homeostasis observed for this group of genes in normal samples. To achieve this classification, we used the distribution of the normal samples' iESs as the benchmark (obtained their mean and standard deviation). Specifically, we labeled a tumor sample as ''perturbed'' if its iES was more than two standard deviations away from the normal samples' mean, in the direction along the normal samples' mean towards the tumor samples' mean. Otherwise, the sample is labeled ''normal-like''.

In cancer studies, especially for solid tumors, ''normal'' samples typically refer to tissues adjacent to the tumor site, hence the level of heterogeneity in the normal samples is usually quite high. This is evidenced by frequently observing more than one mode in the distribution of the iES values among the normal samples. In order to best estimate the mean and standard deviation of the *bona fide* normal samples, we fit a Gaussian mixture model for these iES values to account for heterogeneity, and selected the mean and the standard deviation for the subgroup of samples with the highest posterior probability. This can be achieved by specifying the modelName parameter to ''V'' inside the Mclust function (mclust R package, Scrucca et al., 2016), which is able to automatically determine the number of the modes and assign samples to clusters.

Using pathway ''*FARMER BREAST CANCER APOCRINE VS_LUMINAL*'' in BRCA as an example. In Figure 3E, from the density plots, we observed that the overall iESs for tumor samples were higher than the normal samples (first column: waterfall plot, and second column: density plot), so we used the mean + 2sd as the cutoff to determine whether a tumor sample was perturbed. Figure 3A shows enrichment plots of three normal-like samples in the first column. Figure 3B shows that of three perturbed samples. Figure 3C shows a random normal sample. After classifying all tumor samples into either normal-like or perturbed, survival analysis indicated that this was a prognostic biomarker pathway (see the Kaplan-Meier plot in the fourth column of Figure 3E). The same trend is found in another biomarker pathway *PEDERSEN METASTASIS BY ERBB2 ISOFORM 3*.

QUANTIFICATION AND STATISTICAL ANALYSIS

Performance comparison among sample-level gene set analysis methods

Clustering

We adopt the following steps: (1) randomly choose 50 normal and 50 tumor samples from the TCGA BRCA cohort; (2) for each method, using RNA-seq data of these samples, we calculate an ES matrix with rows corresponding to pathway/gene sets and columns corresponding to samples. (3) conduct DE analysis on the ES using limma (67). (4) select the top 10 gene sets according to the

adjusted p values and perform the hierarchical clustering. (5) bipartition the hierarchical tree into two classes and compare the clustering results with sample labels using ARI. (6) repeat the above process 1,000 times and summarize the average ARI for each method.

Survival analysis

We randomly select 70% of BRCA samples in TCGA as the training set and use the remaining 30% of BRCA samples as the test set. Using training data, we fit individual Cox proportional hazards model for each BIOCARTA pathway and select the pathway that best correlates with the survival. Then using the test data, we assess the predictive ability of the selected pathway by computing the concordance index (c-index). We repeat the random samplings for training and test data 1000 times. The distributions of c-indices are summarized using boxplots.

Microarray data analysis

We analyzed the METABRIC dataset which is under the accession number EGAS00000000083 at the European Genome-Phenome Archive (http://www.ebi.ac.uk/ega/). We downloaded the gene expression matrix for 144 normal samples (EGAF00000102978) and 997 tumor samples (EGAF00000102986) from the discovery set. The 48,803 Illumina probes in the gene expression data were mapped to 30,492 known gene symbols using Bioconductor package *illuminaHumanv4.db*. Next, we applied the iPath and Pathifier on these 1141 samples and all 186 KEGG pathways (Kanehisa and Goto, 2000) for calculating the iES and PDS individual-level enrichment matrices respectively. Next, we applied Wilcoxon signed-rank test (Wilcoxon, 1945) to rank the KEGG pathways based on their significance.