SCIENTIFIC REPORTS

OPEN

Received: 11 December 2014
Accepted: 27 July 2015
Published: 21 August 2015

Large-scale RNA-Seq Transcriptome Analysis of 4043 Cancers and 548 Normal Tissue Controls across 12 TCGA Cancer Types

Li Peng[1], Xiu Wu Bian[2], Di Kang Li[3], Chuan Xu[2,4], Guang Ming Wang[5], Qing You Xia[1] & Qing Xiong[3]

The Cancer Genome Atlas (TCGA) has accrued RNA-Seq-based transcriptome data for more than 4000 cancer tissue samples across 12 cancer types, translating these data into biological insights remains a major challenge. We analyzed and compared the transcriptomes of 4043 cancer and 548 normal tissue samples from 21 TCGA cancer types, and created a comprehensive catalog of gene expression alterations for each cancer type. By clustering genes into co-regulated gene sets, we identified seven cross-cancer gene signatures altered across a diverse panel of primary human cancer samples. A 14-gene signature extracted from these seven cross-cancer gene signatures precisely differentiated between cancerous and normal samples, the predictive accuracy of leave-one-out cross-validation (LOOCV) were 92.04%, 96.23%, 91.76%, 90.05%, 88.17%, 94.29%, and 99.10% for BLCA, BRCA, COAD, HNSC, LIHC, LUAD, and LUSC, respectively. A lung cancer-specific gene signature, containing SFTPA1 and SFTPA2 genes, accurately distinguished lung cancer from other cancer samples, the predictive accuracy of LOOCV for TCGA and GSE5364 data were 95.68% and 100%, respectively. These gene signatures provide rich insights into the transcriptional programs that trigger tumorigenesis and metastasis, and many genes in the signature gene panels may be of significant value to the diagnosis and treatment of cancer.

Recent advances in cancer genomics have created a rich resource for studying the causes of cancer. The Cancer Genome Atlas (TCGA)[1] (http://cancergenome.nih.gov) has accrued more than 10,000 cases of human cancer including over 25 different cancer types. Datasets including RNA-Seq, miRNA-Seq, Exon-Seq, somatic mutations, methylation, CNV for each case are publically available via the TCGA Data Portal (https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp) and UCSC Cancer Genomics Hub (https://cghub.ucsc.edu). Translating these data into biological insights remains a major challenge. Currently several studies have analyzed genome-wide mutational patterns in different cancer types and identified genes harboring functional mutations implicated in cancerogenesis[2–5]. Cancer is thought to be driven by gene expression pattern changes due to the accumulation of mutations or epigenetic modifications; thus, a comprehensive characterization of alterations in gene expression will not only advance

[1]State Key Laboratory of Silkworm Genome Biology, Southwest University, Chongqing 400715, China. [2]Institute of Pathology and Southwest Cancer Center, Southwest Hospital, Third Military Medical University, Chongqing 400038, China. [3]Department of Computer Science and Technology, Department of Statistics, Southwest University, Chongqing 400715, China. [4]Department of Oncology, Chengdu Military General Hospital of PLA, Chengdu 610083, China. [5]Department of Pathology, Clinical School, Dali University, Dali 671000, China. Correspondence and requests for materials should be addressed to Q.X. (email: qingx@swu.edu.cn) or Q.Y.X. (email: xiaqy@swu.edu.cn)

	BLCA	BRCA	COAD	HNSC	LIHC	LUAD	LUSC	KICH	KIRC	KIRP	PRAD	THCA
BLCA	1.00	0.36	0.41	0.32	0.27	0.36	0.23	0.05	0.00	0.14	0.32	0.14
BRCA	0.25	1.00	0.41	0.36	0.36	0.45	0.32	0.05	0.00	0.18	0.18	0.09
COAD	0.29	0.17	1.00	0.41	0.45	0.45	0.27	0.00	0.00	0.18	0.23	0.05
HNSC	0.15	0.17	0.12	1.00	0.36	0.41	0.32	0.05	0.00	0.23	0.09	0.05
LIHC	0.06	0.12	0.07	0.04	1.00	0.36	0.27	0.05	0.00	0.23	0.09	0.00
LUAD	0.19	0.28	0.14	0.07	0.16	1.00	0.68	0.00	0.05	0.23	0.14	0.14
LUSC	0.18	0.27	0.12	0.09	0.26	0.55	1.00	0.00	0.18	0.18	0.09	0.05
KICH	0.04	0.06	0.07	0.06	0.03	0.03	0.03	1.00	0.05	0.14	0.00	0.05
KIRC	0.03	0.04	0.05	0.05	0.06	0.04	0.03	0.15	1.00	0.09	0.05	0.00
KIRP	0.09	0.11	0.09	0.06	0.07	0.12	0.12	0.18	0.36	1.00	0.05	0.05
PRAD	0.11	0.08	0.12	0.07	0.07	0.10	0.08	0.05	0.08	0.07	1.00	0.14
THCA	0.08	0.06	0.07	0.06	0.06	0.05	0.05	0.05	0.06	0.09	0.07	1.00

Table 1. The percentage of common genes and gene sets in top 3% most differentially expressed genes and gene sets between 12 cancer types*. *Results from DE gene comparisons lie below the diagonal line; results from DE gene set comparisons are above the diagonal line.

our understanding of cancer biology, it will also provide a large number of new potential diagnostic and therapeutic targets for cancer. Cheng *et al.*[6] introduced a method to identify cancer-associated attractors and revealed some interesting bimolecular events shared among multiple cancer types based on microarray gene expression data. However, genome-wide association analysis of RNA-Seq transcriptome data across various TCGA cancer types has rarely been reported. RNA-Seq, a revolutionary technology for genome-wide gene expression profiling, offers several key advantages compared to microarrays[7], it could better characterize the transcriptomic changes associated with human cancers.

In this study, we analyzed and compared the RNA-Seq transcriptomes of 4043 cancer and 548 solid tissue normal samples across 21 types of cancer from TCGA. We created a catalog of gene expression alterations for each cancer type, and our results show that the alterations in gene expression vary substantially between different tumor types. Studies have shown that cancer involves many different genes and a majority of these genes have a small to moderate effect[8], it is difficult to detect these effects by single gene analysis. By clustering genes into co-regulated gene sets, we are able to examine accumulative effects of a group of functionally related genes. We performed gene set association analysis for each cancer type; our results revealed several common gene signatures shared by multiple cancer types and a lung cancer-specific gene signature. We also validated these signatures using several non-TCGA data sets. These cross-cancer and cancer-specific transcriptional aberrations improve our understanding of the etiology of human cancers, and are of great importance for the diagnosis and treatment of cancer.

Results

Gene-level differential expression analysis of transcriptomes. We conducted gene differential expression analysis and created a catalog of gene expression alterations for each of 12 cancer types; the results are shown in Supplementary Table S1. Our results show that a large number of genes were differentially expressed. Among a total of 20530 genes, the percentage of differentially expressed (DE) genes with FDR < 0.01 is 0.32, 0.72, 0.51, 0.52, 0.52, 0.65, 0.68, 0.54, 0.69, 0.46, 0.46, and 0.56 for BLCA, BRCA, COAD, HNSC, LIHC, LUAD, LUSC, KICH, KIRC, KIRP, PRAD, and THCA, respectively. To examine the similarity of DE genes between cancer types, we extracted the top 3% most differentially expressed genes from each cancer type. We then calculated the number of common DE genes between cancer types. As shown in Table 1, we found that DE genes vary substantially across cancer types, and there are less than 20% common DE genes between most cancer types. LUAD and LUSC, two forms of lung cancers, turn out to be most similar cancers since they share 55% of DE genes. Contrarily, the DE profiles of two kidney cancers, KICH and KIRC, are quite different from each other and others; the percentage of common DE genes is less than 10%. Additionally, THCA is also poorly overlapped with other cancers in terms of DE genes. The diversity in differential expression could be explained by several factors: (1) many of gene expression alterations may be cancer type-specific; (2) aberrations in different genes may have same phenotypic consequences; (3) single gene analysis may miss many subtle effects on causative genes.

Gene clustering. Prior to gene set association analysis, we clustered genes based on their expression profiles over all normal samples across 12 cancer types. We obtained a total of 3236 clusters (Supplementary Table S2). The expression changes of genes in a cluster are highly correlated under

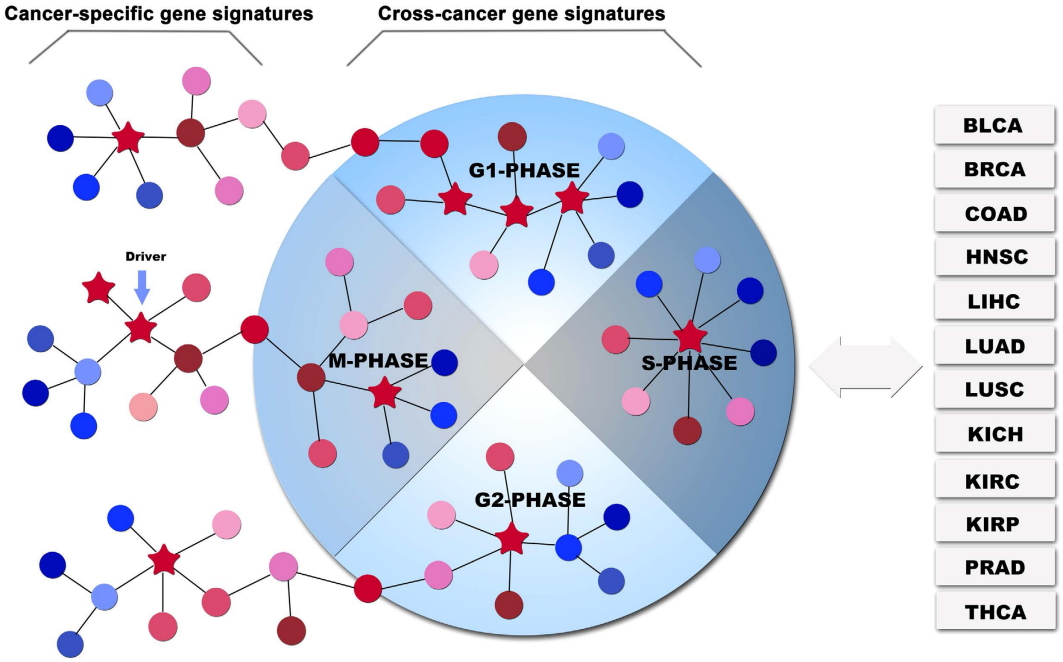

Figure 1. Two possible carcinogenic mechanisms. (1) gene expression aberrations in cell cycle-associated pathways can directly lead to carcinogenesis, these pathways are cross-cancer gene signatures altered across a range of cancer types; (2) gene expression aberrations in organ-specific pathways can indirectly lead to carcinogenesis by interacting with cell cycle-associated pathways, these pathways are cancer-specific gene signatures altered in a single cancer type. Stars represent driver mutations that can alter the expression levels of their target genes.

various conditions, thus, it is reasonable to assume that genes in the same cluster are co-regulated or belong to the same pathway.

Gene set association analysis of TCGA data. Cancers arise from the aberrations in multiple genes, many of which only have moderate or weak effect sizes that are difficult to detect by only analyzing individual genes, therefore, we adopted gene set association analysis to detect the accumulative effect of a group of functionally related genes and to reveal the transcriptional program accounting for the variability in phenotype.

Carcinogenesis is caused by the accumulation of mutations and epimutations in normal cells[9,10], which confer a growth and selective advantage upon these cells, resulting in uncontrolled cell division and the evolution of these cells by natural selection[11]. The mutations can be classified into two classes, driver mutation and passenger mutation, according to their phenotypic effects[12]. Driver mutations are causally implicated in carcinogenesis while passenger mutations don't contribute to the development of cancer. A driver mutation is expected to alter the gene expression of its target genes and/or genes that share the same biological pathway[13,14], and these changes in gene expression account for the phenotypic variance[15].

The cell cycle lies at the core of cancer[16,17]. In normal cells, the cell cycle is controlled by a series of signaling pathways by which a cell grows, replicates its DNA and divides. In cancers, as a result of mutations, this regulatory process malfunctions, resulting in uncontrolled cell proliferation that leads to carcinogenesis[18,19]. From the perspective of pathway, we hypothesize that there may be two potential carcinogenic mechanisms, as illustrated in Fig. 1: (1) one or more driver mutations are within a cell cycle-associated pathway, altering its expression pattern and consequently leading to cancer; (2) one or more driver mutations lie in an organ/tissue-specific pathway or other pathways not related to cell cycle, which interacts with a cell cycle-associated pathway, alters its expression pattern, and ultimately results in cancer. Since the deregulation of cell cycle is a common characteristic shared by multiple cancer types, we expected that the expression of cell cycle-associated pathways would be altered across a range of cancers. By analyzing and comparing the transcriptome data of 12 cancer types, we can test this hypothesis.

A gene signature denotes a set of genes that are significantly differentially expressed between cancer and normal samples. We call those pathways/gene sets significantly altered in multiple cancer types as cross-cancer gene signatures while those disrupted in just one cancer type as cancer-specific gene signatures. We performed gene set association analysis using all gene sets generated by gene clustering; the results are shown in Supplementary Table S3. We identified 20, 7, 7, 6, 7, 15, 30, and 1 significant gene sets for BLCA, BRCA, COAD, HNSC, LIHC, LUAD, LUSC, and KICH, respectively. No significant associations were found for KIRC, KIRP, PRAD, and THCA. Among 46 significant gene sets, seven are

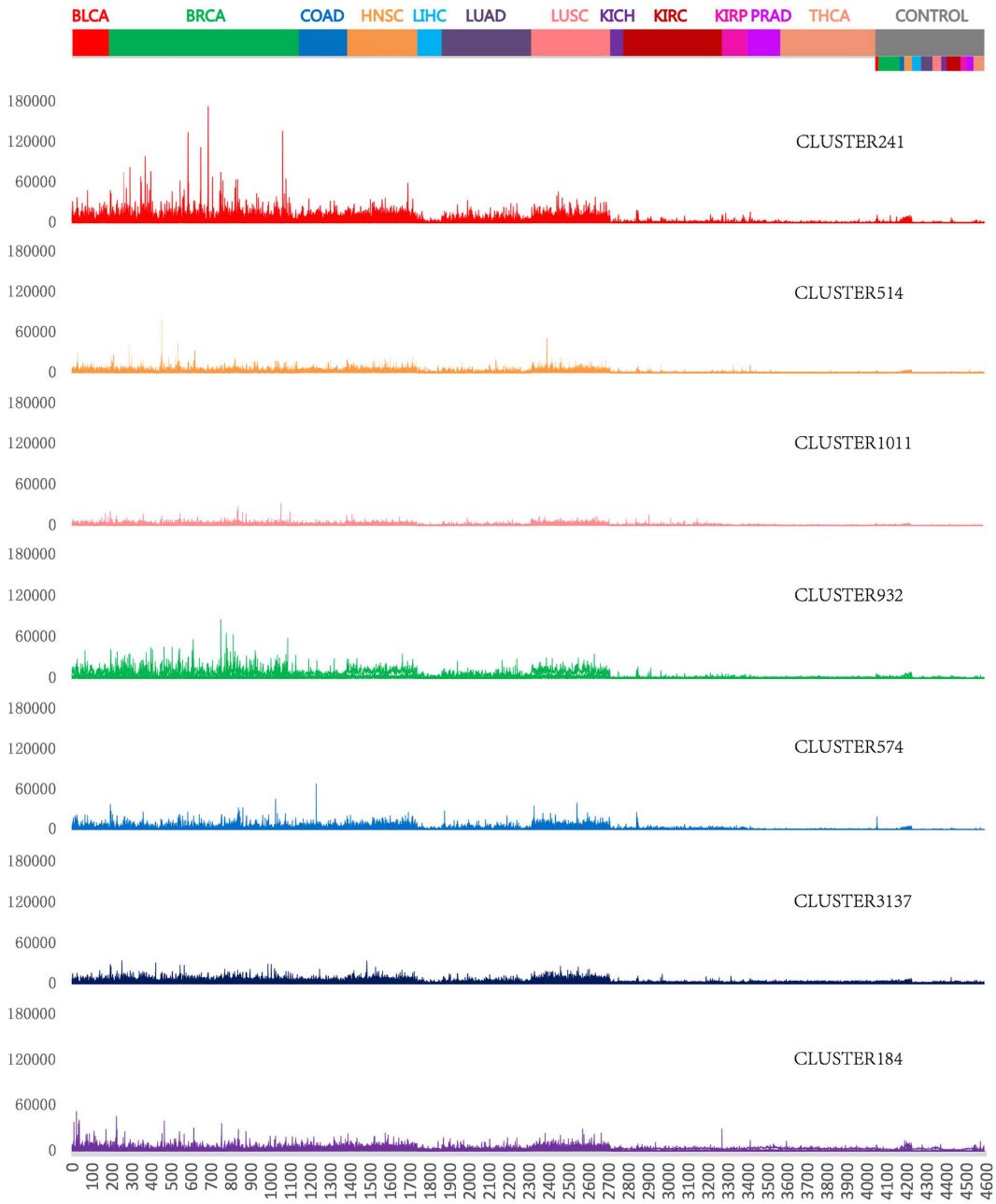

Figure 2. The normalized expression levels of seven cross-cancer gene signatures across 12 types of cancer and normal samples.

cross-cancer gene signatures whose expression levels were significantly altered in at least four cancer types (Fig. 2), the false discovery rates (FDRs) of these gene sets for each cancer type are shown in Table 2. In order to gain biological insights into these gene sets, we performed three types of pathway enrichment analyses, GO analysis, KEGG analysis, and Pathway Commons analysis, and disease association analysis for genes of each of these gene sets. The results of these analyses are shown in Supplementary Table S4. Interestingly, we found that these seven cross-cancer gene signatures are all closely related to cell cycle regulation, as we expected. Gene set CLUSTER2556 is significant in BLCA, COAD, and LUSC. There are 9 significant gene sets shared by two cancer types. Gene set CLUSTER242 is shared by LIHC and LUSC, and the remaining 8 gene sets are shared by LUAD and LUSC. LUAD and LUSC are more similar to one another than other cancer types possibly because they are both lung cancers.

Cross-cancer gene signatures. We identified seven cross-cancer gene signatures: CLUSTER241, CLUSTER514, CLUSTER1011, CLUSTER932, CLUSTER574, CLUSTER3137, and CLUSTER184, that were altered in at least four types of human cancers. All of these signatures are associated with cell cycle regulation.

CLUSTER	BLCA	BRCA	COAD	HNSC	LIHC	LUAD	LUSC	KICH	KIRC	KIRP	PRAD	THCA
CLUSTER241	0. 0134	0. 0066	0. 047	0. 005	0. 0122	0. 0048	0					
CLUSTER514	0. 0389	0. 0056	0. 1093	0. 0232	0. 0067	0. 0177	0. 004					
CLUSTER1011	0. 0755	0. 0838	0. 0878	0. 0458	0. 0209	0. 0431	0. 0201					
CLUSTER932	0. 0309	0. 0881	0. 1031	0. 0435	0. 1365	0. 0491	0. 0176					
CLUSTER574		0. 0181	0. 1089	0. 0321	0. 0385	0. 0412	0. 0199					
CLUSTER3137	0. 1385		0. 1108		0. 1303	0. 1052	0. 0221					
CLUSTER184	0. 124	0. 0724				0. 0983	0. 0719					
CLUSTER2556	0. 1398		0. 0915				0. 116					
CLUSTER242					0. 1285		0. 0566					
CLUSTER2527						0. 0469	0. 021					
CLUSTER2212						0. 0497	0. 0038					
CLUSTER901						0. 0683	0. 0271					
CLUSTER909						0. 0899	0. 0721					
CLUSTER1057						0. 0941	0. 0538					
CLUSTER844						0. 1011	0. 0192					
CLUSTER2771						0. 1379	0. 0212					
CLUSTER132						0. 1441	0. 0226					
CLUSTER2990	0. 1251											
CLUSTER1860	0. 1292											
CLUSTER2452	0. 1306											
CLUSTER2323	0. 1318											
CLUSTER2208	0. 1353											
CLUSTER1945	0. 1358											
CLUSTER492	0. 1389											
CLUSTER1219	0. 1395											
CLUSTER2174	0. 1416											
CLUSTER2712	0. 1446											
CLUSTER3003	0. 1458											
CLUSTER2738	0. 1482											
CLUSTER1403	0. 1499											
CLUSTER891		0. 1498										
CLUSTER2318				0. 1473								
CLUSTER1520							0. 0202					
CLUSTER2242							0. 0233					
CLUSTER2637							0. 0233					
CLUSTER831							0. 0234					
CLUSTER149							0. 0238					
CLUSTER55							0. 0264					
CLUSTER2822							0. 0357					
CLUSTER535							0. 0412					
CLUSTER1893							0. 0426					
CLUSTER533							0. 0545					
CLUSTER2916							0. 0738					
CLUSTER2752							0. 1162					
CLUSTER790							0. 1444					
CLUSTER2240								0. 0004				

Table 2. Significant gene sets at FDR < 0.15 from 12 types of cancers.

Cross-cancer gene signature 1 – CLUSTER241. CLUSTER241 is significantly altered in seven cancer types: BLCA, BRCA, COAD, HNSC, LIHC, LUAD, and LUSC. GO analysis, KEGG analysis, and Pathway Commons analysis indicate that genes in this cluster are enriched in pathways involved in the cell cycle. The top enriched GO biological process, KEGG pathway, and Pathway Commons pathway are M Phase, Cell Cycle, and Mitotic Prometaphase, respectively. The top associated disease is Aneuploidy. Aneuploidy, denoting cells with an abnormal number of chromosomes, is commonly observed in human cancer; it has been recognized as a key characteristic of cancer[20,21]. This cluster contains 33 genes, several of

which have reported roles in cancer. Kinesins have been reported to play critical roles in the initiation and development of human cancers[22,23]. Marker of proliferation Ki-67 (*MKI67*) is a prognostic marker for breast cancer[24,25]. Simultaneous aberration of topoisomerase (DNA) II alpha (*TOP2A*) and v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 2 (*ERBB2*/*HER2*) has been observed in multiple tumor types[26,27].

Cross-cancer gene signature 2 – CLUSTER514. CLUSTER514 is significantly altered in seven cancer types: BLCA, BRCA, COAD, HNSC, LIHC, LUAD, and LUSC. GO analysis, KEGG analysis, and Pathway Commons analysis indicate that genes in this cluster are enriched in pathways involved in the cell cycle. The top enriched GO biological process, KEGG pathway, and Pathway Commons pathway are Organelle Fission, Cell Cycle, and Cell Cycle, Mitotic, respectively. The top associated disease is Cancer or Viral Infections. This cluster contains 36 genes, of which a lot are prognostic markers for cancer. Enhancer of zeste 2 polycomb repressive complex 2 subunit (*EZH2*) has been linked to multiple cancers[28,29]. Aurora kinase A (*AURKA*) causes chromosome instability by inactivating p53 and contributes to tumorigenesis/carcinogenesis[30–32]. Baculoviral IAP repeat containing 5 (*BIRC5*) is over-expressed in most human cancers; the microRNA targeting BIRC5 suppresses cell proliferation in triple-negative breast cancer (TNBC) cells[33–35]. Thymidine Kinase 1 (*TK1*), which is elevated in the early stages of malignancies, is a universal marker for cancer[36,37]. Polo-like kinase 1 (*PLK1*) is overexpressed in many tumor types; it is a target for cancer therapy[38–40]. RAD51 recombinase (*RAD51*) plays a critical role in DNA Damage Repair and is a potential therapeutic target for cancer[41,42]. Hyaluronan-mediated motility receptor (*HMMR*) is correlated to the stemness and tumorigenicity of cancer stem cells[43,44]. Cyclin B1 (*CCNB1*), PDZ binding kinase (*PBK*), and cyclin-dependent kinase inhibitor 3 (*CDKN3*) are also prognostic biomarkers for various types of cancer[45–49].

Cross-cancer gene signature 3 – CLUSTER1011. CLUSTER1011 is significantly altered in seven cancer types: BLCA, BRCA, COAD, HNSC, LIHC, LUAD, and LUSC. GO analysis, KEGG analysis, and Pathway Commons analysis indicate that genes in this cluster are enriched in pathways involved in the cell cycle. The top enriched GO biological process, KEGG pathway, and Pathway Commons pathway are Cell Cycle, Cell Cycle, and DNA Replication, respectively. The top associated disease is Fanconi Anemia (FA). The FA proteins are involved in the cell-cycle checkpoint and DNA-repair pathways[50,51]. This cluster contains 19 genes, several of which have been linked to cancer. Mutations in BRCA1 interacting protein C-terminal helicase 1 (*BRIP1*) have been associated with ovarian cancer and breast cancer[52–54]. The overexpression of *KIAA1524*/*CIP2A* have been observed in multiple types of cancer[55–57]. Centromere protein H (*CENPH*) is a prognostic marker for cancer[58–61].

Cross-cancer gene signature 4 – CLUSTER932. CLUSTER932 is significantly altered in seven cancer types: BLCA, BRCA, COAD, HNSC, LIHC, LUAD, and LUSC. GO analysis, KEGG analysis, and Pathway Commons analysis indicate that genes in this cluster are enriched in pathways involved in the cell cycle. The top enriched GO biological process, KEGG pathway, and Pathway Commons pathway are Cell Cycle, Cell Cycle, and Cell Cycle, Mitotic, respectively. The top associated disease is Retinoblastoma. This cluster contains 19 genes, many of which have well-known roles in cancer. Cyclin E1 (*CCNE1*) and cyclin E2 (*CCNE2*) play critical roles in cell cycle regulation and are potential therapeutic targets in cancer[62–64]. The aberrant expression of cell division cycle 6 (*CDC6*) has been documented in multiple human cancers[65–67]. E2F transcription factor 7 (*E2F7*) is interacted with p53, it has been implicated as playing a role in tumorigenesis[68–70]. Ubiquitin-like with PHD and ring finger domains 1 (*UHRF1*) is an upstream regulator of the Tip60-p53 interaction and it has been linked to liver cancer[71].

Cross-cancer gene signature 5 – CLUSTER574. CLUSTER574 is significantly altered in six cancer types: BRCA, COAD, HNSC, LIHC, LUAD, and LUSC. GO analysis, KEGG analysis, and Pathway Commons analysis indicate that genes in this cluster are enriched in pathways involved in the cell cycle. The top enriched GO biological process, KEGG pathway, and Pathway Commons pathway are Mitotic Cell Cycle, Pyrimidine Metabolism, and Cell Cycle, Mitotic, respectively. The top associated disease is Pancreatic Diseases. This cluster contains 17 genes, several of which have been associated with cancer. Forkhead box M1 (*FOXM1*) is overexpressed in the majority of human cancers, it has well-known roles in cancer[72–74]. Thymidylate synthetase (*TYMS*) is considered a prognostic biomarker for cancer[75,76]. Ribonucleotide reductase M2 (*RRM2*) is associated with poor survival; it is also implicated in angiogenesis[77–79]. Spindle and kinetochore associated complex subunit 1 (*SKA1*) has been highlighted as a biomarker in several types of cancers[80].

Cross-cancer gene signature 6 – CLUSTER3137. CLUSTER3137 is significantly altered in five cancer types: BLCA, COAD, LIHC, LUAD, and LUSC. GO analysis, KEGG analysis, and Pathway Commons analysis indicate that genes in this cluster are enriched in pathways involved in the cell cycle. The top enriched GO biological process, KEGG pathway, and Pathway Commons pathway are Cell Cycle Process, Cell Cycle, and Mitotic M-M/G1 Phases, respectively. The top associated disease is Retinoblastoma. This cluster contains 16 genes, several of which have been linked to cancer. S-phase kinase-associated protein 2, E3 ubiquitin protein ligase (*SKP2*) is a protooncogene in human tumors and is a potential

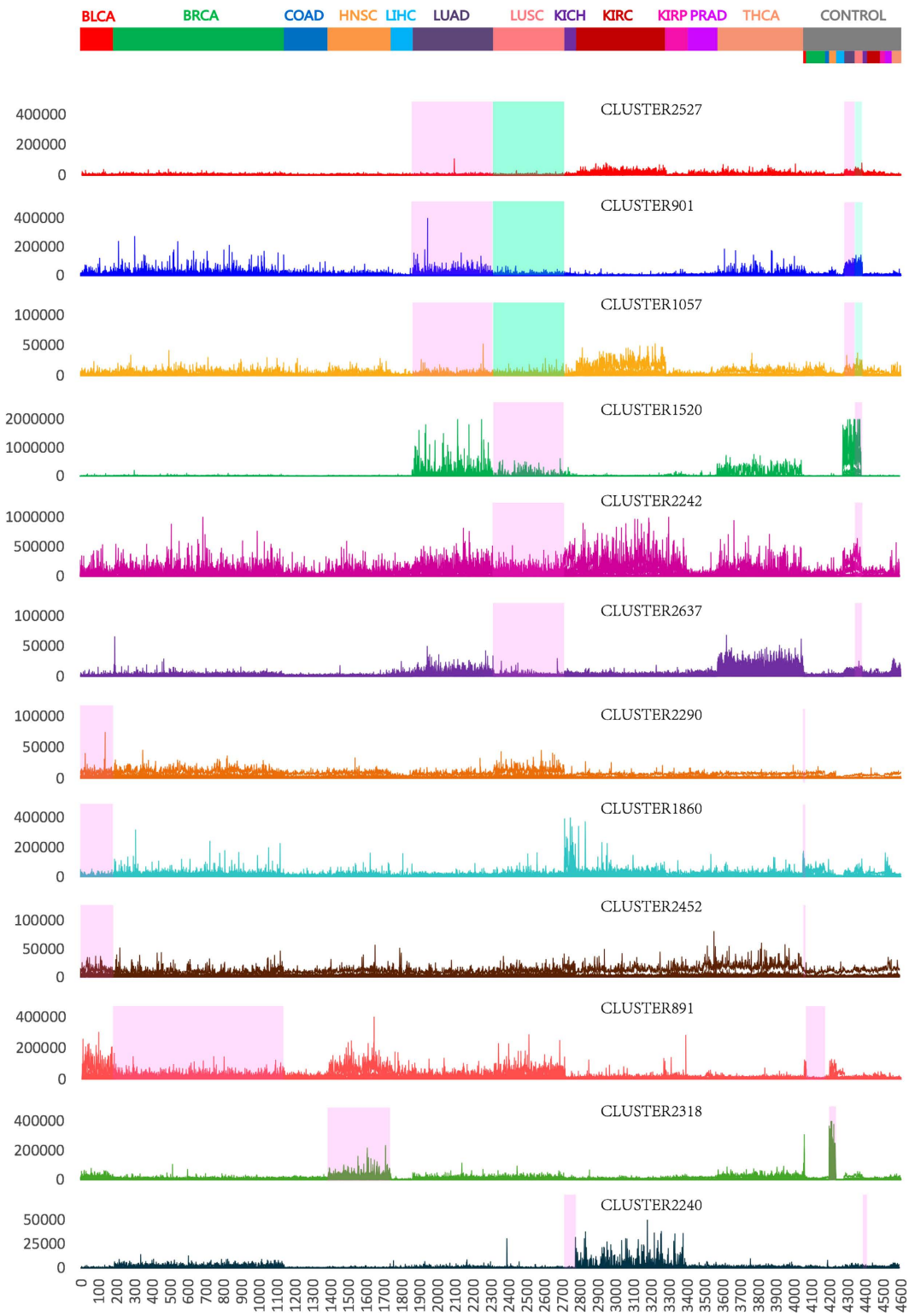

Figure 3. The normalized expression levels of gene signatures significantly altered in one type of cancer across 12 types of cancer and normal samples.

cancer drug target[81–83]. Ribonucleotide reductase M1 (*RRM1*) is a prognostic marker for cancer[84,85]. DNA (cytosine-5-)-methyltransferase 1 (*DNMT1*) is overexpressed in many cancers and is correlated to the aberrant methylation in human cancer cells[86]. The polymorphisms of *DNMT1* have been reported to increase breast cancer risk[87–89].

Symbol	Full name	Lung diseases
SFTPA2	surfactant, pulmonary-associated protein A2	lung cancer[101–104], acute and chronic lung disease[105,106]
SFTPA1	surfactant, pulmonary-associated protein A1	lung cancer[103], acute and chronic lung disease[107–109]
ROS1	ROS proto-oncogene 1, receptor tyrosine kinase	lung cancer[165,166],
ABCA3	ATP-binding cassette, sub-family A (ABC1), member 3	pulmonary fibrosis[167], respiratory disease[168–170]
AQP4	aquaporin 4	lung cancer[171,172]
HHIP	hedgehog interacting protein	lung cancer[173], chronic obstructive pulmonary disease[174], abnormal lung function[175]
SLC34A2	solute carrier family 34, member 2	pulmonary alveolar microlithiasis[176,177]
IL6R	interleukin 6 receptor	chronic obstructive pulmonary disease[178], abnormal lung function[179]

Table 3. List of lung disease-associated genes in CLUSTER1520.

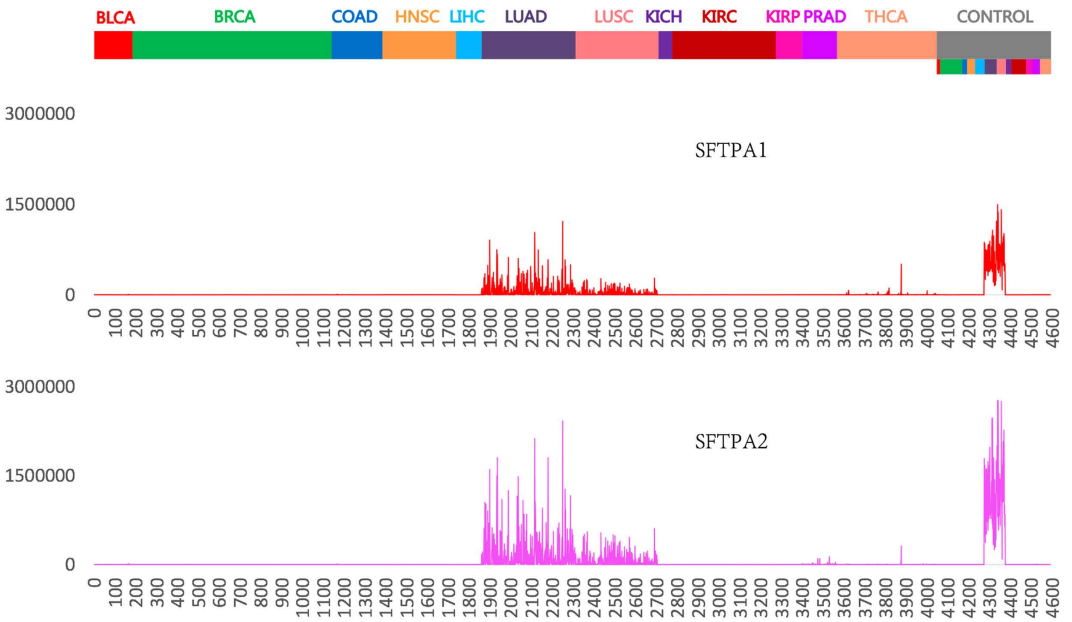

Figure 4. The normalized expression levels of SFTPA1 and SFTPA2 across 12 types of cancer and normal samples.

Cross-cancer gene signature 7 – CLUSTER184. CLUSTER184 is significantly altered in four cancer types: BLCA, BRCA, LUAD, and LUSC. GO analysis, KEGG analysis, and Pathway Commons analysis indicate that genes in this cluster are enriched in pathways involved in the cell cycle. The top enriched GO biological process, KEGG pathway, and Pathway Commons pathway are M Phase, Oocyte Meiosis, and Cell Cycle, Mitotic, respectively. The top associated disease is Aneuploidy. This cluster contains 24 genes, several of which have been linked to cancer. The aurora kinase B (*AURKB*) was shown to be overexpressed in many types of cancer cells, and it has been implicated in the carcinogenesis and tumor development process[90–92]. Ubiquitin-conjugating enzyme E2C (*UBE2C*/*UBCH10*) has been reported to play a critical role in carcinogenesis and tumor development[93–95].

Although these seven cross-cancer gene signatures in a broad sense are involved in the cell cycle, they may manifest different cellular processes leading to the abnormal cell cycle regulation in malignancy. For example, DNMT1 in CLUSTER3137 is the major enzyme responsible for maintenance of the DNA methylation pattern[96–98]. *DNMT1* has been reported to be overexpressed in many cancers and to be involved in the epigenetic silencing of tumor suppressor genes in human tumor cells[86]. Therefore, the perturbation of CLUSTER3137 might be an epigenetic trigger of tumorigenesis. The deregulation of CLUSTER1011 may reveal the roles of components of the Fanconi anemia/BRCA pathway in human cancers. Increasing evidence shows that FA proteins are involved in the DNA damage response[50,51]. In this cluster, except for genes that have established roles in the DNA damage response, such as Fanconi anemia, complementation group D2 (*FANCD2*)[99], our study also suggests genes, e.g., downstream neighbor of SON

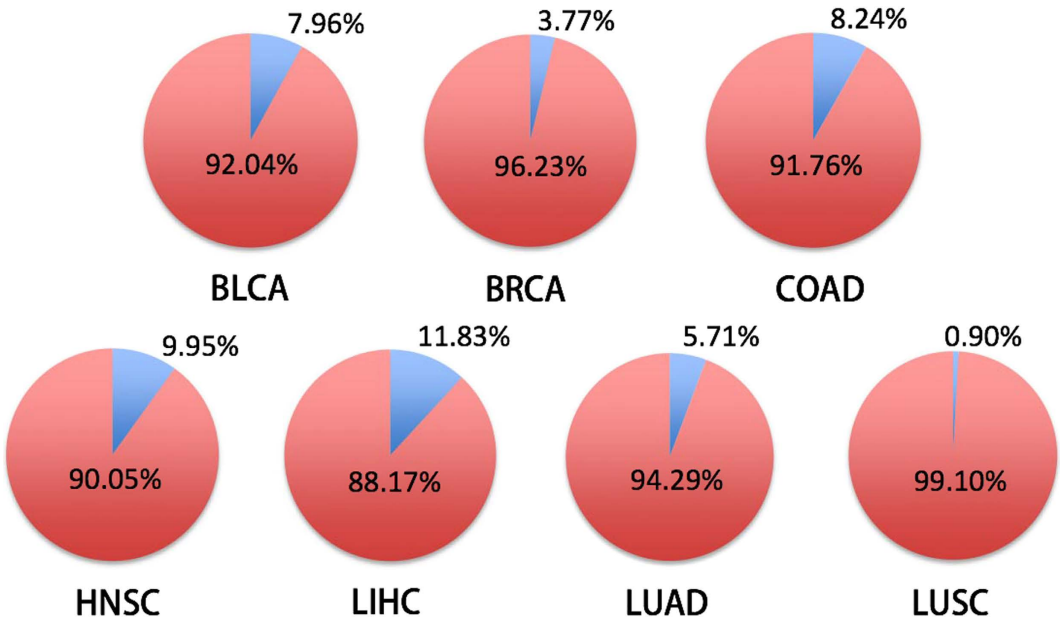

Figure 5. The predictive accuracy and error rates of LOOCV for each cancer type using the 14-gene signature. Red indicates the predictive accuracy; Blue represents error rates.

(DONSON) and proline/serine-rich coiled-coil 1 (PSRC1), that may have new unrevealed functions in DNA repair since the expression levels of theses genes were up-regulated in accordance with FA proteins and BRIP1 in cancer samples. Altogether, these seven cross-cancer gene signatures can not only deepen and broaden our understanding of the cellular events involving carcinogenesis related to the four phases of the cell cycle, they also reveal many potential novel therapeutic targets that have so far not been linked to cancers but may have unknown roles in cancer biology. Our study can be considered as a starting point, and further investigations (e.g., mutation analysis, survival analysis, and functional analysis) on these genes or clusters may lead to the discovery of novel cancer biomarkers and development of new anticancer therapies.

Gene signatures significantly altered in one type of cancer. Based on the TCGA cancer data sets we used, we identified 37 gene signatures significantly altered (FDR < 0.15) only in one type of cancer, of which 21 gene signatures are for lung cancers: LUAD and/or LUSC, 13 for BLCA, 1 for BRCA, 1 for HNSC, and 1 for KICH. Figure 3 lists the expression patterns of part of these gene signatures across cancer and normal samples. Among these signature gene sets, several were implicated in relative organ-specific diseases by disease association analysis, and may provide insights into transcriptional aberrations underlying the initiation and progression of a specific cancer type.

Gene signatures for lung cancers. 21 clusters were significantly altered only in one or both of lung cancers, three of which, CLUSTER1520, CLUSTER901 and CLUSTER1057, have been implicated in lung diseases.

CLUSTER1520 contains 39 genes. Some genes in this cluster have been reported to be associated with lung cancer or other lung diseases (see Table 3 for details). Among them, two genes, SFTPA1 and SFTPA2, encode surfactant protein A (SP-A) that plays a vital role in maintaining normal lung function[100] and have been implicated in various lung diseases[101–104,105–109]. The expression levels of SFTPA1 and SFTPA2 were much higher in lung tissue samples than in any other tissue samples, moreover, these two genes were strikingly down-regulated in lung tumor tissues as compared to the adjacent nontumor tissues (Fig. 4). We thus speculate that the expression changes in these two genes might be an important indicator for lung function abnormalities, and those 39 genes in CLUSTER1520 might form a network underlying the initiation and/or development of lung cancers. It could be valuable to elucidate the possible roles of these genes in lung cancer in an experimental setting.

The top associated disease for CLUSTER901 is Lung Neoplasms (adjP $= 0.0006$). This cluster contains 32 genes, several of which have been reported to play roles in lung diseases. G protein-coupled receptor, class C, group 5, member A (*Gprc5a*) protein is detected in the lungs more than in any other tissue; Gprc5a knockout promotes lung inflammation and tumorigenesis in mice[110–112]. Moreover, *GPRC5A* is down-regulated in the adjacent field and normal bronchial epithelia of patients with chronic obstructive pulmonary disease and non-small-cell lung cancer[113,114]. Wingless-type MMTV integration site family, member 7A (*WNT7A*) has been reported to be associated with lung cancer[115,116]. Claudin 18 (*CLDN18*)

deficiency is related to alveolar barrier dysfunction[117,118]. Adrenoceptor beta 2 (*ADRB2*) is associated with lung function and lung diseases[119,120].

The top associated diseases for CLUSTER1057 are Lung Diseases (adjP = 0.0037), Respiratory Tract Diseases (adjP = 0.0037), and Airway Obstruction (adjP = 0.0037). CLUSTER1057 contains many immunity-associated genes and might contribute to the immune reactions to lung cancers. Among them, interleukin 33 (*IL33*) has been linked to lung diseases[121,122]; interferon (alpha, beta and omega) receptor 2 (IFNAR2) is a prognostic biomarker for lung cancer[123]; GTPase, IMAP family member 6 (GIMAP6) and member 8 (GIMAP8) were significantly down-regulated in the non-small cell lung cancer[124].

Gene signatures for BLCA. Thirteen clusters were significantly altered only in BLCA, two of which, CLUSTER2174 and CLUSTER1860, have been implicated in bladder abnormalities. The top associated disease for CLUSTER2174 is Urogenital Abnormalities (adjP = 0.0008). This cluster contains 15 genes. Among them, fibroblast growth factor receptor 1 (*FGFR1*) is a well-known gene that plays a key role in the development of urothelial carcinomas[125,126]. The top associated disease for CLUSTER1860 is Cystitis (adjP = 3.08e-05). This cluster contains 18 genes. Gap junction protein, gamma 1 (*GJC1*/*CX45*) is one of the two most important gap junction proteins in bladder smooth muscle cells and suburothelial myofibroblasts that are essential for the coordination of normal bladder function[127]. SPARC-like 1 (*SPARCL1*) is down-regulated in bladder cancer and prostate cancer[128,129].

Gene signatures for BRCA. CLUSTER891 is significantly altered only in BRCA. The top associated disease is Adenocarcinoma (adjP = 0.0182). This cluster contains 16 genes, two of which have been linked to p53. p53 represses hepatoma-derived growth factor (*HDGF*), and loss of p53 function contributes to tumorigenesis by elevating *HDGF* expression[130,131]. p53 induces the expression of ferredoxin reductase (*FDXR*) which sensitizes cells to apoptosis[132,133]. Syndecan 1 (*SDC1*) promotes tumor angiogenesis and growth[134,135].

Gene signatures for KICH. CLUSTER2240 is significantly altered only in KICH. The top associated disease for CLUSTER2240 is Ciliary Motility Disorders (adjP = 1.85e-05), and Ciliary dysfunction is a risk factor for both syndromic and isolated kidney cystic disease[136]. This cluster contains 25 genes. Nephronophthisis 1 (*NPHP1*/*NPH1*) gene deletion is correlated with nephronophthisis[137–140].

We have shown that DE genes vary dramatically across 12 cancer types. To test whether there exists a more consistent DE pattern at the gene set level, we extracted the top 3% most differentially expressed gene sets from each cancer type, and calculated the number of common DE gene sets between cancer types. The results are shown in the upper triangular matrix in Table 1. We found that the percentage of common gene sets increase compared to the percentage of common genes for most of cancer pairs. This suggests there are common patterns shared by different tumor types, and these patterns can be detected more effectively at the gene set level. These similarities across cancer types shed light on biomarkers that can be used across a range of cancer types and thus have important implications for treatment.

Through genome-wide gene set association analysis of all co-regulated clusters, we identified both cross-cancer gene signatures, which regulate the cell cycle, and cancer-specific gene signatures, which are associated with relative organ/tissue-specific diseases. These partly verified our hypothesis that alterations in cell cycle-associated pathways directly contribute to the initiation and development of cancers, while some organ/tissue-specific pathways can lead to cancers possibly by altering the expression of cell-cycle associated pathways. More functional investigations are necessary for further validating this hypothesis.

Leave-one-out cross validation. Seven gene sets, CLUSTER241, CLUSTER514, CLUSTER1011, CLUSTER932, CLUSTER574, CLUSTER3137, and CLUSTER184, were differentially expressed in at least four of the seven cancer types: BLCA, BRCA, COAD, HNSC, LUAD, and LUSC. We extracted the top two most differentially expressed genes from these gene sets and created a 14-gene signature, including kinesin family member 4A (*KIF4A*), nucleolar and spindle associated protein 1 (*NUSAP1*), Holliday junction recognition protein (*HJURP*), NIMA-related kinase 2 (*NEK2*), Fanconi anemia, complementation group I (*FANCI*), denticleless E3 ubiquitin protein ligase homolog (Drosophila) (*DTL*), *UHRF1*, flap structure-specific endonuclease 1 (*FEN1*), IQ motif containing GTPase activating protein 3 (*IQGAP3*), kinesin family member 20A (*KIF20A*), tripartite motif containing 59 (*TRIM59*), centromere protein L (*CENPL*), chromosome 16 open reading frame 59 (*C16orf59*), and *UBE2C*. We employed leave-one-out cross-validation (LOOCV) to assess whether or not this 14-gene signature can be used to differentiate between the normal and cancerous tissue samples of those seven cancer types. Machine learning techniques, for example support vector machines, have been playing a vital role in sample classification[141–144]. LOOCV was performed using SVM-light[145] (http://svmlight.joachims.org/) that is an implementation of support vector machines. The predictive accuracy of LOOCV for each cancer type are shown in Fig. 5. The predictive accuracy is the proportion of the total number of predictions that were correct. We found that most of samples were correctly classified based on the expression levels of these 14 genes, the classification accuracy for BLCA, BRCA, COAD, HNSC, LIHC, LUAD, and LUSC were 92.04%, 96.23%, 91.76%, 90.05%, 88.17%, 94.29%, and 99.10%, respectively.

Validation of the 14-gene cross-cancer signature and a cancer-specific gene signature, CLUSTER1520, on non-TCGA data sets. We have shown that the cancerous and adjacent normal samples from BLCA, BRCA, COAD, HNSC, LIHC, LUAD and LUSC can be precisely classified using the 14-gene cross-cancer signature. To test if the same holds true for other non-TCGA data sources, we downloaded two RNA-Seq data sets, GSE40419[146] and GSE50760[147], and one microarray data set, GSE5364[148], from the Gene Expression Omnibus (GEO: http://www.ncbi.nlm.nih.gov/geo). GSE40419 includes the RNA-Seq expression values for 87 lung adenocarcinomas and 77 adjacent normal tissues, while GSE50760 contains the RNA-Seq expression values of 54 samples (18 primary colorectal cancer, 18 liver metastasis, and 18 normal colon) generated from 18 colorectal cancer patients. We performed LOOCV on these two data sets based on the expression values of the 14-gene cross-cancer signature. We found that the tumor and normal samples were accurately classified, the predictive accuracy for GSE40419 and GSE50760 were 97.14% and 93.33%, respectively. GSE5364 includes 341 samples from multiple solid cancers: 18 lung tumor samples, 12 lung normal samples, 183 breast tumor samples, 13 breast normal samples, 9 colon tumor samples, 9 colon normal samples, 9 liver tumor samples, 8 liver normal samples, 16 oesophagus tumor samples, 13 oesophagus normal samples, 35 thyroid tumor samples, and 16 thyroid normal samples. LOOCV was carried out for tumor and normal samples of each tumor type in this data set, the predictive accuracy for lung, breast, colon, liver, oesophagus, and thyroid samples were 100%, 93.37%, 100%, 100%, 94.12%, and 68.63%, respectively. These results show that our 14-gene cross-cancer signature precisely differentiated between tumor and normal samples for all tumor types in GSE5364 except for those from the thyroid. Interestingly, we here were not able to effectively distinguish tumors from normal samples from the thyroid using this 14-gene cross-cancer signature, and this is consistent with the results from the TCGA data.

We found that CLUSTER1520 is a lung cancer-specific gene signature. In the 548 adjacent normal tissue samples of 12 TCGA cancer types, the expression level of CLUSTER1520 in the lung tissue samples was strikingly higher than any other tissue samples, and the same holds true for tumor samples if excluding THCA tumor samples from the analysis (Fig. 3). Moreover, CLUSTER1520 showed a substantially reduced level of expression in the lung tumor samples as compared to lung normal samples. In order to test if this signature can be used to differentiate lung tumors from other tumors, we divided all cancer samples from 12 TCGA cancer types into two classes: lung cancer samples (LUAD, LUSC) and non-lung cancer samples (BLCA, BRCA, COAD, HNSC, LIHC, KICH, KIRC, KIRP, PRAD, THCA), and performed LOOCV on these two classes of cancer samples using the expression values of CLUSTER1520. The predictive accuracy was 95.68%, namely we very effectively identified lung cancer samples out of a selection of 12 TCGA cancers based on the expression pattern of CLUSTER1520. We also validate that CLUSTER1520 is a lung cancer-specific gene signature on a non-TCGA microarray data set (GSE5364). GSE5364 includes 6 tumor types, and we divided those tumor samples into two classes: lung tumor samples and non-lung tumor samples (breast, colon, liver, oesophagus, thyroid). The predictive accuracy of LOOCV for these two classes of tumor samples was 100%, this demonstrated that lung tumor samples and non-lung tumor samples were accurately classified based on CLUSTER1520. These results show that CLUSTER1520 is a lung cancer-specific gene signature, and genes in this signature are potential targets for developing novel lung cancer therapies.

Gene set association analysis of two non-TCGA data sets. In order to test if the significant gene sets we identified from TCGA data can be validated on non-TCGA data sources, we performed gene set association analysis on two non-TCGA cohorts: a lung adenocarcinoma data set (GSE40419) and a colorectal cancer data set (GSE50760), the results are shown in Supplementary Table S5. For GSE40419, we identified 1 significant gene set (FDR < 0.25), CLUSTER514, which is one of the seven cross-cancer gene signatures. Moreover, we found that 9 out of 10 top gene sets with smallest FDRs were within the significant gene sets identified from TCGA LUAD and/or LUSC data. For GSE50760, we identified five significant gene sets (FDR < 0.25), two of them, CLUSTER2556 and CLUSTER514, were also identified as significant using TCGA COAD data. In the 10 top gene sets with smallest FDRs, 4 overlap with the significant gene sets from TCGA COAD data. These results show that at least part of significant gene sets identified from TCGA data can be validated on these two non-TCGA data sets.

Discussion

In this study, we comprehensively characterized the gene expression alterations of 12 types of cancers at the gene and gene set level. We identified DE genes and gene sets, some DE genes and gene sets are shared by different cancer types while others are only altered in one cancer type. We identified seven cross-cancer gene signatures that are differentially expressed in at least 4 cancer types. These signatures contain not only a number of genes which have established roles in cancers, but also genes that might be potential new biomarkers for cancers. These results reveal the aberrations in cancer transcriptomes and lead to a deeper understanding of the formation and development of human cancers.

Traditionally, we research one cancer type at a time, but there are patterns that can only be detected by making connections across different cancer types. Our results reveal that four gene sets, CLUSTER241, CLUSTER514, CLUSTER1011, and CLUSTER932, are significantly altered across seven cancer types: BLCA, BRCA, COAD, HNSC, LIHC, LUAD, and LUSC (Table 2). These similarities may indicate that there exist common mechanisms underlying the initiation and/or development of human cancers from

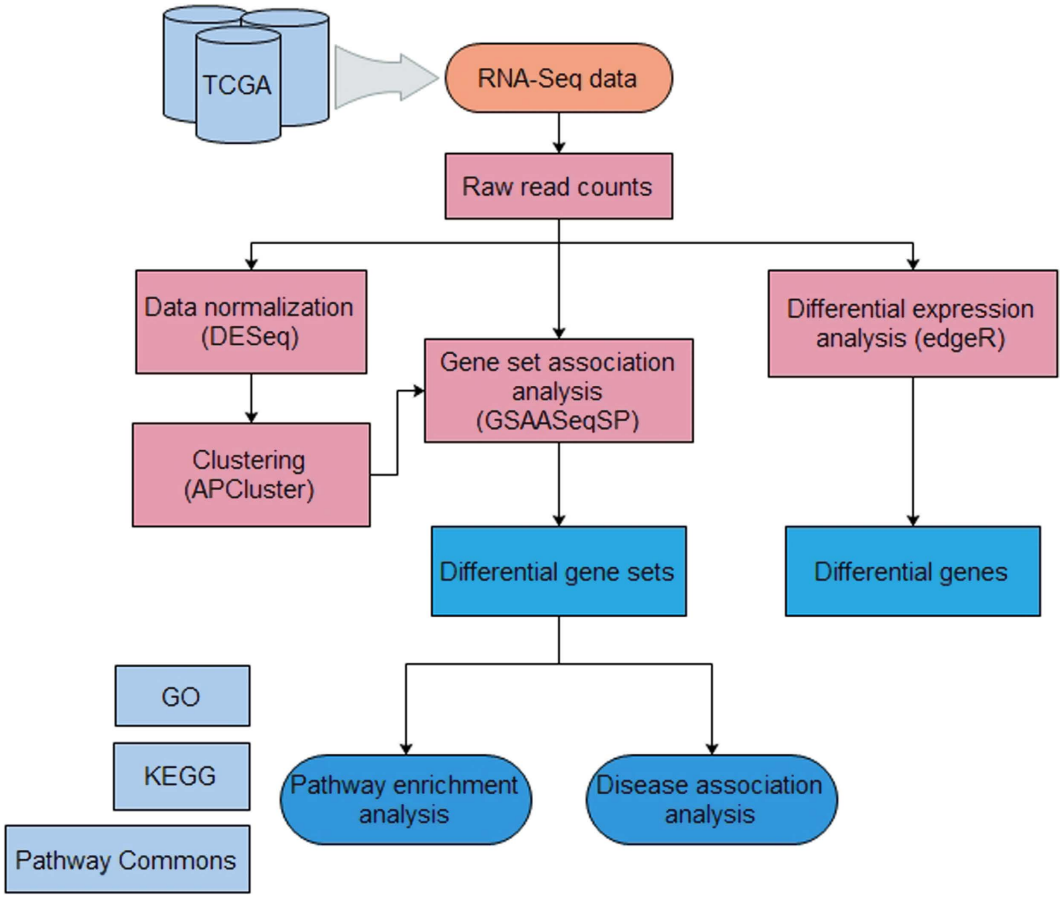

Figure 6. Pipeline of the analysis.

Abbreviation	Full Name	Number of Cancers	Number of Normals
BLCA	bladder urothelial carcinoma	185	16
BRCA	breast invasive carcinoma	955	106
COAD	colon adenocarcinoma	244	23
HNSC	head and neck squamous cell carcinoma	353	39
LIHC	liver hepatocellular carcinoma	123	46
LUAD	lung adenocarcinoma	450	58
LUSC	lung squamous cell carcinoma	398	44
KICH	kidney chromophobe	66	25
KIRC	kidney renal clear cell carcinoma	497	72
KIRP	kidney renal papillary cell carcinoma	127	28
PRAD	prostate adenocarcinoma	166	38
THCA	thyroid carcinoma	479	53

Table 4. Number of cancer and normal samples of 12 cancer types.

different organs or different tissues in the same organ. Interestingly, three types of kidney tumors don't show these patterns. We found that KIRC and KIRP are more similar to each other than KICH since they share 36% of DE genes (Table 1). Studies have shown that KICH is less aggressive than KIRC and KIRP[149,150].

Gene expression changes with phenotypic consequences are driven by mutations and epimutations. The driver mutations and epimutations may be scattered in different pathways. We hypothesize that some of these mutations or epimutations may disrupt a pathway responsible for cell cycle regulation that directly drives cells into uncontrolled proliferation, while others may lie within an organ-specific

pathway that turn a healthy cell into a cancer cell by altering the expression of cell cycle-associated pathways. We were not able to directly detect which pathways harbor the driver mutations through gene expression analysis, but we observed evidence that at least partially support this hypothesis: 1) aberrations in the cell cycle are a common feature shared by multiple cancer types since all of the cross-cancer gene signatures we identified involved cell cycle processes; 2) we found that some cancer-specific gene signatures contain genes implicated in corresponding organ-specific diseases; 3) each type of cancer has its unique features in terms of their DE profiles. It would be very interesting for future studies to explore the connections between mutational profiles and DE profiles and how gene expression patterns change surrounding driver mutations.

We identified some gene sets that were only significantly altered in one type of cancer. Some of these gene sets may be cancer-specific gene signatures, say CLUSTER1520 and CLUSTER2318, that shed light on the mechanisms underlying cancer-driving abnormalities in a specific organ, while many of them may still represent a cellular process broadly perturbed across cancer types, and the differentiation is just stronger in one cancer type than other cancer types. Therefore, when we look for mechanisms underlying a specific cancer type, we should treat these signatures with caution. It could be a way to reveal cancer-specific events by comparing various tumor types and looking into the differential gene sets between tumor types in the future.

A question arising from this study is how to make connections between the mutational profiles and DE profiles of human cancers. Some genes, for example tumor protein p53 (*TP53*/*p53*), phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha (*PIK3CA*), and retinoblastoma 1 (*RB1*), are frequently mutated in a number of cancers and are key genes contributing to tumorigenesis[151–153]. However, among 20530 genes in the BLCA, BRCA, COAD, HNSC, LIHC, LUAD, LUSC, KICH, KIRC, KIRP, PRAD, and THCA datasets, we found that *TP53* was ranked at 11834, 14601, 7752, 18515, 17359, 8769, 11116, 14995, 4200, 986, 4776, and 2094, *PIK3CA* at 16090, 7012, 14799, 4118, 13535, 13228, 6632, 12475, 14634, 17065, 19153, and 18438, *RB1* at 15691, 15157, 6836, 16116, 16543, 9168, 19208, 8333, 9565, 4233, 16756, and 11160, respectively (Supplementary Table S1). These genes are not at the top of the DE gene list. One reason could be that mutations in *TP53*/*PIK3CA*/*RB1* substantially change the expression of its downstream target genes rather than genes harboring them[154]. For example, the expression levels of two *p53* targets, *E2F7* and *HDGF*[69,130], are significantly altered across multiple cancer types. These results indicate that cancer-causing genes may only have subtle expression changes, it is thus crucial to measure the total effect of a pathway or integrate mutation analysis into gene expression analysis.

Batch effects in high-throughput data might lead to inaccurate results when dealing with samples from multiple cancer types or data from different sequencing platforms[155]. In our study, all of the RNA-Seq data we used were from the same sequencing platform and same sequencing center, and this design minimizes the impact of batch effects on our analyses. Second, we clustered genes by their changing tendency in expression over samples, and this can eliminate the impact of batch effects on clustering since these effects are global effects on every gene in a sample. Third, we carried out gene set association analysis for each cancer type independently, thus avoiding the cross-cancer bias from the batch effects. Of course, proper handling of batch effects could improve the cross-platform or cross-cancer consistency when performing analyses on data from different sequencing platforms or different cancer types.

Methods

Overview. The pipeline of our analysis is illustrated in Fig. 6. The details of each step are described below.

Data sets. Transcriptome data and clinical data were obtained from the TCGA Data Portal (https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp). In order to eliminate the heterogeneity introduced by different sequencing platforms, we only downloaded those data in the category of UNC (IlluminaHiSeq_RNASeqV2). We chose 12 cancer types with transcriptome data available for both cancer and normal tissue samples. The two classes of phenotypes we used were "primary tumor" and "solid tissue normal", namely only those samples in the clinical category of "primary tumor" or "solid tissue normal" were used for this study. The number of cancer and normal samples for each cancer type are listed in Table 4.

Differential expression analysis of individual genes. Differential expression analysis of individual genes was carried out using the edgeR Bioconductor package[156] (http://www.bioconductor.org/packages/release/bioc/html/edgeR.html). For each cancer type, we divide samples into two phenotypic groups, primary tumor and solid tissue normal, based on their clinical labels. Raw counts were extracted for these samples and edgeR was employed to find the differentially expressed genes between the two phenotypic groups.

Clustering and gene set association analysis. In order to perform gene set association analysis, we first clustered genes into co-regulated sets based on their expression profiles over all normal samples across 12 cancer types. First, RNA-Seq data were normalized by the DESeq normalization method[157], clustering was then performed using APCluster[158,159] (http://www.bioinf.jku.at/software/apcluster/). We used the Pearson correlation coefficient to measure the similarity between genes. Pearson correlation

measures the similarity in shape between two expression profiles, so this metric partitions genes into gene groups whose expression levels rise or fall synchronously under varying conditions or in response to a sequence of environment stimuli. We consider genes with coherent changing tendency in expression as co-regulated genes possibly functional in a same pathway. The number of clusters generated by APCluster is largely determined by the input preference, so we set the input preference (q) to 0.98 to obtain precise clusters, namely the expression profiles of genes in the same cluster are highly correlated.

We performed gene set association analysis for each cancer type to identify gene sets/clusters significantly associated with cancers. Gene set association analysis was carried out using GSAASeqSP, a software newly developed by our group[160] (http://gsaa.unc.edu). RNA-Seq raw counts were normalized by the DESeq normalization in GSAASeqSP which is same as that in the DESeq Bioconductor package[157]. We chose Signal2Noise for gene-level differential expression analysis and Weighted_KS for gene set association analysis. Gene sets are gene clusters generated by APCluster, and one cluster represents one gene set. Gene sets with less than 15 genes or more than 100 genes were filtered to avoid overly narrow or broad functional categories. In this study, we set the FDR cutoff to 0.15, namely gene sets with FDR < 0.15 were considered to be statistically significantly associated with cancers.

Pathway enrichment analysis and disease association analysis. To gain biological understanding of those gene sets statistically significantly associated with cancers, we carried out pathway enrichment analysis and disease association analysis using WebGestalt[161,162] (http://bioinfo.vanderbilt.edu/webgestalt/). We conducted three types of pathway enrichment analyses for genes of significant gene sets: GO analysis, KEGG analysis, and Pathway Commons analysis. GO analysis is to find which GO terms are over-represented in a gene set based on the functional annotation of genes. KEGG analysis and Pathway Commons analysis are to discover pathways enriched in genes in a gene set, the difference between these two analyses is that KEGG analysis is based on the KEGG PATHWAY Database[163] (http://www.genome.jp/kegg/pathway.html) while Pathway Commons analysis uses pathways collected by Pathway Commons[164]. Disease association analysis identifies diseases in which genes in a gene set are over-represented. We adopted the default values for parameters in WebGestalt when performing pathway enrichment analysis and disease association analysis.

References

1. Hudson, T. J. *et al.* International network of cancer genome projects. *Nature* **464,** 993–998 (2010).
2. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).
3. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep* **3,** 246–259 (2013).
4. Tamborero, D. *et al.* Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci Rep* **3,** 2650; doi: 10.1038/srep02650 (2013).
5. Abaan, O. D. *et al.* The exomes of the NCI-60 panel: a genomic resource for cancer biology and systems pharmacology. *Cancer Res* **73,** 4372–4382 (2013).
6. Cheng, W. Y., Ou Yang, T. H. & Anastassiou, D. Biomolecular events in cancer revealed by attractor metagenes. *PLoS Comput Biol* **9,** e1002920; doi: 10.1371/journal.pcbi.1002920 (2013).
7. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10,** 57–63 (2009).
8. Pharoah, P. D., Dunning, A. M., Ponder, B. A. & Easton, D. F. Association studies for finding cancer-susceptibility genetic variants. *Nat Rev Cancer* **4,** 850–860 (2004).
9. Loeb, K. R. & Loeb, L. A. Significance of multiple mutations in cancer. *Carcinogenesis* **21,** 379–385 (2000).
10. Banno, K. *et al.* Epimutation and cancer: a new carcinogenic mechanism of Lynch syndrome (Review). *Int J Oncol* **41,** 793–797 (2012).
11. Tomlinson, I. P., Novelli, M. R. & Bodmer, W. F. The mutation rate and cancer. *Proc Natl Acad Sci USA* **93,** 14800–14803 (1996).
12. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458,** 719–724 (2009).
13. Gibson, G. Cancer: Directions for the drivers. *Nature* **512,** 31–32 (2014).
14. Bashashati, A. *et al.* DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biol* **13,** R124; doi: 10.1186/gb-2012-13-12-r124 (2012).
15. Khaitovich, P., Enard, W., Lachmann, M. & Paabo, S. Evolution of primate gene expression. *Nat Rev Genet* **7,** 693–702 (2006).
16. Williams, G. H. & Stoeber, K. The cell cycle and cancer. *J Pathol* **226,** 352–364 (2012).
17. Collins, K., Jacks, T. & Pavletich, N. P. The cell cycle and cancer. *Proc Natl Acad Sci USA* **94,** 2776–2778 (1997).
18. Hartwell, L. H. & Kastan, M. B. Cell cycle control and cancer. *Science* **266,** 1821–1828 (1994).
19. Negrini, S., Gorgoulis, V. G. & Halazonetis, T. D. Genomic instability–an evolving hallmark of cancer. *Nat Rev Mol Cell Biol* **11,** 220–228 (2010).
20. Sen, S. Aneuploidy and cancer. *Curr Opin Oncol* **12,** 82–88 (2000).
21. Gordon, D. J., Resio, B. & Pellman, D. Causes and consequences of aneuploidy in cancer. *Nat Rev Genet* **13,** 189–203 (2012).
22. Hung, P. F. *et al.* The motor protein KIF14 inhibits tumor growth and cancer metastasis in lung adenocarcinoma. *PLoS One* **8,** e61664; doi: 10.1371/journal.pone.0061664 (2013).
23. Liu, X., Gong, H. & Huang, K. Oncogenic role of kinesin proteins and targeting kinesin therapy. *Cancer Sci* **104,** 651–656 (2013).
24. de Azambuja, E. *et al.* Ki-67 as prognostic marker in early breast cancer: a meta-analysis of published studies involving 12,155 patients. *Br J Cancer* **96,** 1504–1513 (2007).
25. Tawfik, K., Kimler, B. F., Davis, M. K., Fan, F. & Tawfik, O. Ki-67 expression in axillary lymph node metastases in breast cancer is prognostically significant. *Hum Pathol* **44,** 39–46 (2013).
26. Liang, Z. *et al.* Analysis of EGFR, HER2, and TOP2A gene status and chromosomal polysomy in gastric adenocarcinoma from Chinese patients. *BMC Cancer* **8,** 363; doi: 10.1186/1471-2407-8-363 (2008).
27. Arriola, E. *et al.* Genomic analysis of the HER2/TOP2A amplicon in breast cancer and breast cancer cell lines. *Lab Invest* **88,** 491–503 (2008).

28. Kleer, C. G. *et al.* EZH2 is a marker of aggressive breast cancer and promotes neoplastic transformation of breast epithelial cells. *Proc Natl Acad Sci USA* **100,** 11606–11611 (2003).
29. Chase, A. & Cross, N. C. Aberrations of EZH2 in cancer. *Clin Cancer Res* **17,** 2613–2618 (2011).
30. Lee, D. F. *et al.* Regulation of embryonic and induced pluripotency by aurora kinase-p53 signaling. *Cell Stem Cell* **11,** 179–194 (2012).
31. Chou, C. H. *et al.* Chromosome instability modulated by BMI1-AURKA signaling drives progression in head and neck cancer. *Cancer Res* **73,** 953–966 (2013).
32. Zhu, J., Abbruzzese, J. L., Izzo, J., Hittelman, W. N. & Li, D. AURKA amplification, chromosome instability, and centrosome abnormality in human pancreatic carcinoma cells. *Cancer Genet Cytogenet* **159,** 10–17 (2005).
33. Nassar, A., Lawson, D., Cotsonis, G. & Cohen, C. Survivin and caspase-3 expression in breast cancer: correlation with prognostic parameters, proliferation, angiogenesis, and outcome. *Appl Immunohistochem Mol Morphol* **16,** 113–120 (2008).
34. Boidot, R. *et al.* The expression of BIRC5 is correlated with loss of specific chromosomal regions in breast carcinomas. *Genes Chromosomes Cancer* **47,** 299–308 (2008).
35. Wang, C., Zheng, X., Shen, C. & Shi, Y. MicroRNA-203 suppresses cell proliferation and migration by targeting BIRC5 and LASP1 in human triple-negative breast cancer cells. *J Exp Clin Cancer Res* **31,** 58; doi: 10.1186/1756-9966-31-58 (2012).
36. Alegre, M. M., Robison, R. A. & O'Neill, K. L. Thymidine Kinase 1: A Universal Marker for Cancer. *Cancer and Clinical Oncology* **2,** 159–167 (2013).
37. Alegre, M. M., Robison, R. A. & O'Neill, K. L. Thymidine kinase 1 upregulation is an early event in breast tumor formation. *J Oncol* **2012,** 575647; doi: 10.1155/2012/575647 (2012).
38. Degenhardt, Y. & Lampkin, T. Targeting Polo-like kinase in cancer therapy. *Clin Cancer Res* **16,** 384–389 (2010).
39. Weiss, L. & Efferth, T. Polo-like kinase 1 as target for cancer therapy. *Exp Hematol Oncol* **1,** 38; doi: 10.1186/2162-3619-1-38 (2012).
40. Hu, K., Law, J. H., Fotovati, A. & Dunn, S. E. Small interfering RNA library screen identified polo-like kinase-1 (PLK1) as a potential therapeutic target for breast cancer that uniquely eliminates tumor-initiating cells. *Breast Cancer Res* **14,** R22; doi: 10.1186/bcr3107 (2012).
41. Lord, C. J. & Ashworth, A. RAD51, BRCA2 and DNA repair: a partial resolution. *Nat Struct Mol Biol* **14,** 461–462 (2007).
42. Nagathihalli, N. S. & Nagaraju, G. RAD51 as a potential biomarker and therapeutic target for pancreatic cancer. *Biochim Biophys Acta* **1816,** 209–218 (2011).
43. Tilghman, J. *et al.* HMMR maintains the stemness and tumorigenicity of glioblastoma stem-like cells. *Cancer Res* **74,** 3168–3179 (2014).
44. Pujana, M. A. *et al.* Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nat Genet* **39,** 1338–1349 (2007).
45. Suzuki, T. *et al.* Nuclear cyclin B1 in human breast carcinoma as a potent prognostic factor. *Cancer Sci* **98,** 644–651 (2007).
46. Hassan, K. A. *et al.* Clinical significance of cyclin B1 protein expression in squamous cell carcinoma of the tongue. *Clin Cancer Res* **7,** 2458–2462 (2001).
47. Hu, F. *et al.* PBK/TOPK interacts with the DBD domain of tumor suppressor p53 and modulates expression of transcriptional targets including p21. *Oncogene* **29,** 5464–5474 (2010).
48. Shih, M. C. *et al.* TOPK/PBK promotes cell migration via modulation of the PI3K/PTEN/AKT pathway and is associated with poor prognosis in lung cancer. *Oncogene* **31,** 2389–2400 (2012).
49. Li, T., Xue, H., Guo, Y. & Guo, K. CDKN3 is an independent prognostic factor and promotes ovarian carcinoma cell proliferation in ovarian cancer. *Oncol Rep* **31,** 1825–1831 (2014).
50. D'Andrea, A. D. & Grompe, M. The Fanconi anaemia/BRCA pathway. *Nat Rev Cancer* **3,** 23–34 (2003).
51. Wang, W. Emergence of a DNA-damage response network consisting of Fanconi anaemia and BRCA proteins. *Nat Rev Genet* **8,** 735–748 (2007).
52. Rafnar, T. *et al.* Mutations in BRIP1 confer high risk of ovarian cancer. *Nat Genet* **43,** 1104–1107 (2011).
53. Seal, S. *et al.* Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. *Nat Genet* **38,** 1239–1241 (2006).
54. De Nicolo, A. *et al.* A novel breast cancer-associated BRIP1 (FANCJ/BACH1) germ-line mutation impairs protein stability and function. *Clin Cancer Res* **14,** 4672–4680 (2008).
55. Kim, J. S., Kim, E. J., Oh, J. S., Park, I. C. & Hwang, S. G. CIP2A modulates cell-cycle progression in human cancer cells by regulating the stability and activity of Plk1. *Cancer Res* **73,** 6667–6678 (2013).
56. Liu, N. *et al.* Overexpression of CIP2A is an independent prognostic indicator in nasopharyngeal carcinoma and its depletion suppresses cell proliferation and tumor growth. *Mol Cancer* **13,** 111; doi: 10.1186/1476-4598-13-111 (2014).
57. Vaarala, M. H., Vaisanen, M. R. & Ristimaki, A. CIP2A expression is increased in prostate cancer. *J Exp Clin Cancer Res* **29,** 136; doi: 10.1186/1756-9966-29-136 (2010).
58. Zhao, X. *et al.* Interruption of cenph causes mitotic failure and embryonic death, and its haploinsufficiency suppresses cancer in zebrafish. *J Biol Chem* **285,** 27924–27934 (2010).
59. Liao, W. T. *et al.* Centromere protein H is a novel prognostic marker for nasopharyngeal carcinoma progression and overall patient survival. *Clin Cancer Res* **13,** 508–514 (2007).
60. Liao, W. T. *et al.* Centromere protein H is a novel prognostic marker for human nonsmall cell lung cancer progression and overall patient survival. *Cancer* **115,** 1507–1517 (2009).
61. Liao, W. T. *et al.* Overexpression of centromere protein H is significantly associated with breast cancer progression and overall patient survival. *Chin J Cancer* **30,** 627–637 (2011).
62. Potemski, P. *et al.* Cyclin E expression in breast cancer correlates with negative steroid receptor status, HER2 expression, tumor grade and proliferation. *J Exp Clin Cancer Res* **25,** 59–64 (2006).
63. Sieuwerts, A. M. *et al.* Which cyclin E prevails as prognostic marker for breast cancer? Results from a retrospective study involving 635 lymph node-negative breast cancer patients. *Clin Cancer Res* **12,** 3319–3328 (2006).
64. Nakayama, N. *et al.* Gene amplification CCNE1 is related to poor survival and potential therapeutic target in ovarian cancer. *Cancer* **116,** 2621–2634 (2010).
65. Gonzalez, S. *et al.* Oncogenic activity of Cdc6 through repression of the INK4/ARF locus. *Nature* **440,** 702–706 (2006).
66. Liu, Y., Gong, Z., Sun, L. & Li, X. FOXM1 and androgen receptor co-regulate CDC6 gene transcription and DNA replication in prostate cancer cells. *Biochim Biophys Acta* **1839,** 297–305 (2014).
67. Robles, L. D. *et al.* Down-regulation of Cdc6, a cell cycle regulatory gene, in prostate cancer. *J Biol Chem* **277,** 25431–25438 (2002).
68. Endo-Munoz, L. *et al.* E2F7 can regulate proliferation, differentiation, and apoptotic responses in human keratinocytes: implications for cutaneous squamous cell carcinoma formation. *Cancer Res* **69,** 1800–1808 (2009).
69. Carvajal, L. A., Hamard, P. J., Tonnessen, C. & Manfredi, J. J. E2F7, a novel target, is up-regulated by p53 and mediates DNA damage-dependent transcriptional repression. *Genes Dev* **26,** 1533–1545 (2012).

70. Chen, H. Z., Tsai, S. Y. & Leone, G. Emerging roles of E2Fs in cancer: an exit from cell cycle control. *Nat Rev Cancer* **9,** 785–797 (2009).
71. Mudbhary, R. *et al.* UHRF1 overexpression drives DNA hypomethylation and hepatocellular carcinoma. *Cancer Cell* **25,** 196–209 (2014).
72. Raychaudhuri, P. & Park, H. J. FoxM1: a master regulator of tumor metastasis. *Cancer Res* **71,** 4329–4333 (2011).
73. Lokody, I. Signalling: FOXM1 and CENPF: co-pilots driving prostate cancer. *Nat Rev Cancer* **14,** 450–451 (2014).
74. Halasi, M. & Gartel, A. L. Targeting FOXM1 in cancer. *Biochem Pharmacol* **85,** 644–652 (2013).
75. Koumarianou, A. *et al.* Prognostic Markers in Early-stage Colorectal Cancer: Significance of TYMS mRNA Expression. *Anticancer Res* **34,** 4949–4962 (2014).
76. Conradi, L. C. *et al.* Thymidylate synthase as a prognostic biomarker for locally advanced rectal cancer after multimodal treatment. *Ann Surg Oncol* **18,** 2442–2452 (2011).
77. Hsu, N. Y. *et al.* Expression status of ribonucleotide reductase small subunits hRRM2/p53R2 as prognostic biomarkers in stage I and II non-small cell lung cancer. *Anticancer Res* **31,** 3475–3481 (2011).
78. Putluri, N. *et al.* Pathway-centric integrative analysis identifies RRM2 as a prognostic marker in breast cancer associated with poor survival and tamoxifen resistance. *Neoplasia* **16,** 390–402 (2014).
79. Zhang, K. *et al.* Overexpression of RRM2 decreases thrombspondin-1 and increases VEGF production in human cancer cells *in vitro* and *in vivo*: implication of RRM2 in angiogenesis. *Mol Cancer* **8,** 11; doi: 10.1186/1476-4598-8-11 (2009).
80. Sun, W., Yao, L., Jiang, B., Guo, L. & Wang, Q. Spindle and kinetochore-associated protein 1 is overexpressed in gastric cancer and modulates cell growth. *Mol Cell Biochem* **391,** 167–174 (2014).
81. Gstaiger, M. *et al.* Skp2 is oncogenic and overexpressed in human cancers. *Proc Natl Acad Sci USA* **98,** 5043–5048 (2001).
82. Lin, H. K. *et al.* Skp2 targeting suppresses tumorigenesis by Arf-p53-independent cellular senescence. *Nature* **464,** 374–379 (2010).
83. Wang, Z. *et al.* Skp2: a novel potential therapeutic target for prostate cancer. *Biochim Biophys Acta* **1825,** 11–17 (2012).
84. Jordheim, L. P., Seve, P., Tredan, O. & Dumontet, C. The ribonucleotide reductase large subunit (RRM1) as a predictive factor in patients with cancer. *Lancet Oncol* **12,** 693–702 (2011).
85. Ceppi, P. *et al.* ERCC1 and RRM1 gene expressions but not EGFR are predictive of shorter survival in advanced non-small-cell lung cancer treated with cisplatin and gemcitabine. *Ann Oncol* **17,** 1818–1825 (2006).
86. Agarwal, S. *et al.* Mahanine restores RASSF1A expression by down-regulating DNMT1 and DNMT3B in prostate cancer cells. *Mol Cancer* **12,** 99; doi: 10.1186/1476-4598-12-99 (2013).
87. Robert, M. F. *et al.* DNMT1 is required to maintain CpG methylation and aberrant gene silencing in human cancer cells. *Nat Genet* **33,** 61–65 (2003).
88. Li, A., Omura, N., Hong, S. M. & Goggins, M. Pancreatic cancer DNMT1 expression and sensitivity to DNMT1 inhibitors. *Cancer Biol Ther* **9,** 321–329 (2010).
89. Kullmann, K., Deryal, M., Ong, M. F., Schmidt, W. & Mahlknecht, U. DNMT1 genetic polymorphisms affect breast cancer risk in the central European Caucasian population. *Clin Epigenetics* **5,** 7; doi: 10.1186/1868-7083-5-7 (2013).
90. den Hollander, J. *et al.* Aurora kinases A and B are up-regulated by Myc and are essential for maintenance of the malignant state. *Blood* **116,** 1498–1505 (2010).
91. Morozova, O. *et al.* System-level analysis of neuroblastoma tumor-initiating cells implicates AURKB as a novel drug target for neuroblastoma. *Clin Cancer Res* **16,** 4572–4582 (2010).
92. Addepalli, M. K. *et al.* RNAi-mediated knockdown of AURKB and EGFR shows enhanced therapeutic efficacy in prostate tumor regression. *Gene Ther* **17,** 352–359 (2010).
93. Wagner, K. W. *et al.* Overexpression, genomic amplification and therapeutic potential of inhibiting the UbcH10 ubiquitin conjugase in human carcinomas of diverse anatomic origin. *Oncogene* **23,** 6621–6629 (2004).
94. Pallante, P. *et al.* UbcH10 overexpression in human lung carcinomas and its correlation with EGFR and p53 mutational status. *Eur J Cancer* **49,** 1117–1126 (2013).
95. Fujita, T. *et al.* Clinicopathological relevance of UbcH10 in breast cancer. *Cancer Sci* **100,** 238–248 (2009).
96. Rhee, I. *et al.* DNMT1 and DNMT3b cooperate to silence genes in human cancer cells. *Nature* **416,** 552–556 (2002).
97. Subramaniam, D., Thombre, R., Dhar, A. & Anant, S. DNA methyltransferases: a novel target for prevention and therapy. *Front Oncol* **4,** 80; doi: 10.3389/fonc.2014.00080 (2014).
98. Robertson, K. D. DNA methylation, methyltransferases, and cancer. *Oncogene* **20,** 3139–3155 (2001).
99. Garcia-Higuera, I. *et al.* Interaction of the Fanconi anemia proteins and BRCA1 in a common pathway. *Mol Cell* **7,** 249–262 (2001).
100. Silveyra, P., DiAngelo, S. L. & Floros, J. An 11-nt sequence polymorphism at the 3'UTR of human SFTPA1 and SFTPA2 gene variants differentially affect gene expression levels and miRNA regulation in cell culture. *Am J Physiol Lung Cell Mol Physiol* **307,** L106–119 (2014).
101. Grageda, M., Silveyra, P., Thomas, N. J., DiAngelo, S. L. & Floros, J. DNA methylation profile and expression of surfactant protein A2 gene in lung cancer. *Exp Lung Res* **41,** 93–102 (2015).
102. Wang, Y. *et al.* Genetic defects in surfactant protein A2 are associated with pulmonary fibrosis and lung cancer. *Am J Hum Genet* **84,** 52–59 (2009).
103. Lin, Z. *et al.* DNA methylation markers of surfactant proteins in lung cancer. *Int J Oncol* **31,** 181–191 (2007).
104. Maitra, M., Cano, C. A. & Garcia, C. K. Mutant surfactant A2 proteins associated with familial pulmonary fibrosis and lung cancer induce TGF-beta1 secretion. *Proc Natl Acad Sci USA* **109,** 21064–21069 (2012).
105. Maitra, M., Wang, Y., Gerard, R. D., Mendelson, C. R. & Garcia, C. K. Surfactant protein A2 mutations associated with pulmonary fibrosis lead to protein instability and endoplasmic reticulum stress. *J Biol Chem* **285,** 22103–22113 (2010).
106. Choi, E. H., Ehrmantraut, M., Foster, C. B., Moss, J. & Chanock, S. J. Association of common haplotypes of surfactant protein A1 and A2 (SFTPA1 and SFTPA2) genes with severity of lung disease in cystic fibrosis. *Pediatr Pulmonol* **41,** 255–262 (2006).
107. Heinrich, S., Hartl, D. & Griese, M. Surfactant protein A–from genes to human lung diseases. *Curr Med Chem* **13,** 3239–3252 (2006).
108. Zhang, Y. *et al.* Identification and examination of a novel 9-bp insert/deletion polymorphism on porcine SFTPA1 exon 2 associated with acute lung injury using an oleic acid-acute lung injury model. *Anim Sci J* **86,** 573–578 (2015).
109. Silveyra, P. & Floros, J. Genetic variant associations of human SP-A and SP-D with acute and chronic lung injury. *Front Biosci (Landmark Ed)* **17,** 407–429 (2012).
110. Deng, J. *et al.* Knockout of the tumor suppressor gene Gprc5a in mice leads to NF-kappaB activation in airway epithelium and promotes lung inflammation and tumorigenesis. *Cancer Prev Res (Phila)* **3,** 424–437 (2010).
111. Barta, P. *et al.* Enhancement of lung tumorigenesis in a Gprc5a Knockout mouse by chronic extrinsic airway inflammation. *Mol Cancer* **11,** 4; doi: 10.1186/1476-4598-11-4 (2012).
112. Chen, Y. *et al.* Gprc5a deletion enhances the transformed phenotype in normal and malignant lung epithelial cells by eliciting persistent Stat3 signaling induced by autocrine leukemia inhibitory factor. *Cancer Res* **70,** 8917–8926 (2010).

113. Fujimoto, J. *et al.* G-protein coupled receptor family C, group 5, member A (GPRC5A) expression is decreased in the adjacent field and normal bronchial epithelia of patients with chronic obstructive pulmonary disease and non-small-cell lung cancer. *J Thorac Oncol* **7,** 1747–1754 (2012).
114. Kadara, H. *et al.* A Gprc5a tumor suppressor loss of expression signature is conserved, prevalent, and associated with survival in human lung adenocarcinomas. *Neoplasia* **12,** 499–505 (2010).
115. Ohira, T. *et al.* WNT7a induces E-cadherin in lung cancer cells. *Proc Natl Acad Sci U S A* **100,** 10429–10434 (2003).
116. Tennis, M. A., Vanscoyk, M. M., Wilson, L. A., Kelley, N. & Winn, R. A. Methylation of Wnt7a is modulated by DNMT1 and cigarette smoke condensate in non-small cell lung cancer. *PLoS One* **7,** e32921; doi: 10.1371/journal.pone.0032921 (2012).
117. LaFemina, M. J. *et al.* Claudin-18 deficiency results in alveolar barrier dysfunction and impaired alveologenesis in mice. *Am J Respir Cell Mol Biol* **51,** 550–558 (2014).
118. Micke, P. *et al.* Aberrantly activated claudin 6 and 18.2 as potential therapy targets in non-small-cell lung cancer. *Int J Cancer* **135,** 2206–2214 (2014).
119. Torjussen, T. M. *et al.* Childhood lung function and the association with beta2-adrenergic receptor haplotypes. *Acta Paediatr* **102,** 727–731 (2013).
120. Marson, F. A., Bertuzzo, C. S., Ribeiro, A. F. & Ribeiro, J. D. Polymorphisms in ADRB2 gene can modulate the response to bronchodilators and the severity of cystic fibrosis. *BMC Pulm Med* **12,** 50; doi: 10.1186/1471-2466-12-50 (2012).
121. Byers, D. E. *et al.* Long-term IL-33-producing epithelial progenitor cells in chronic obstructive lung disease. *J Clin Invest* **123,** 3967–3982 (2013).
122. Li, D. *et al.* IL-33 promotes ST2-dependent lung fibrosis by the induction of alternatively activated macrophages and innate lymphoid cells in mice. *J Allergy Clin Immunol* **134,** 1422–1432 (2014).
123. Tanaka, S. *et al.* Interferon (alpha, beta and omega) receptor 2 is a prognostic biomarker for lung cancer. *Pathobiology* **79,** 24–33 (2012).
124. Shiao, Y. M. *et al.* Dysregulation of GIMAP genes in non-small cell lung cancer. *Lung Cancer* **62,** 287–294 (2008).
125. di Martino, E., Tomlinson, D. C. & Knowles, M. A. A Decade of FGF Receptor Research in Bladder Cancer: Past, Present, and Future Challenges. *Adv Urol* **2012,** 429213; doi: 10.1155/2012/429213 (2012).
126. Lamont, F. R. *et al.* Small molecule FGF receptor inhibitors block FGFR-dependent urothelial carcinoma growth *in vitro* and *in vivo*. *Br J Cancer* **104,** 75–82 (2011).
127. Heinrich, M., Oberbach, A., Schlichting, N., Stolzenburg, J. U. & Neuhaus, J. Cytokine effects on gap junction communication and connexin expression in human bladder smooth muscle cells and suburothelial myofibroblasts. *PLoS One* **6,** e20792; doi: 10.1371/journal.pone.0020792 (2011).
128. Zaravinos, A., Lambrou, G. I., Boulalas, I., Delakas, D. & Spandidos, D. A. Identification of common differentially expressed genes in urinary bladder cancer. *PLoS One* **6,** e18135; doi: 10.1371/journal.pone.0018135 (2011).
129. Hurley, P. J. *et al.* Secreted protein, acidic and rich in cysteine-like 1 (SPARCL1) is down regulated in aggressive prostate cancers and is prognostic for poor clinical outcome. *Proc Natl Acad Sci USA* **109,** 14977–14982 (2012).
130. Sasaki, Y. *et al.* p53 negatively regulates the hepatoma growth factor HDGF. *Cancer Res* **71,** 7038–7047 (2011).
131. Chen, S. C. *et al.* Hepatoma-derived growth factor regulates breast cancer cell invasion by modulating epithelial–mesenchymal transition. *J Pathol* **228,** 158–169 (2012).
132. Liu, G. & Chen, X. The ferredoxin reductase gene is regulated by the p53 family and sensitizes cells to oxidative stress-induced apoptosis. *Oncogene* **21,** 7195–7204 (2002).
133. Lacroix, M., Toillon, R. A. & Leclercq, G. p53 and breast cancer, an update. *Endocr Relat Cancer* **13,** 293–325 (2006).
134. Yang, N., Mosher, R., Seo, S., Beebe, D. & Friedl, A. Syndecan-1 in breast cancer stroma fibroblasts regulates extracellular matrix fiber organization and carcinoma cell motility. *Am J Pathol* **178,** 325–335 (2011).
135. Maeda, T., Desouky, J. & Friedl, A. Syndecan-1 expression by stromal fibroblasts promotes breast carcinoma growth *in vivo* and stimulates tumor angiogenesis. *Oncogene* **25,** 1408–1412 (2006).
136. Gascue, C., Katsanis, N. & Badano, J. L. Cystic diseases of the kidney: ciliary dysfunction and cystogenic mechanisms. *Pediatr Nephrol* **26,** 1181–1195 (2011).
137. Bollee, G. *et al.* Nephronophthisis related to homozygous NPHP1 gene deletion as a cause of chronic renal failure in adults. *Nephrol Dial Transplant* **21,** 2660–2663 (2006).
138. Saunier, S. *et al.* Characterization of the NPHP1 locus: mutational mechanism involved in deletions in familial juvenile nephronophthisis. *Am J Hum Genet* **66,** 778–789 (2000).
139. Konrad, M. *et al.* Large homozygous deletions of the 2q13 region are a major cause of juvenile nephronophthisis. *Hum Mol Genet* **5,** 367–371 (1996).
140. Parisi, M. A. *et al.* The NPHP1 gene deletion associated with juvenile nephronophthisis is present in a subset of individuals with Joubert syndrome. *Am J Hum Genet* **75,** 82–91 (2004).
141. Furey, T. S. *et al.* Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **16,** 906–914 (2000).
142. Zhang, C., Lu, X. & Zhang, X. Significance of gene ranking for classification of microarray samples. *IEEE/ACM Trans Comput Biol Bioinform* **3,** 312–320 (2006).
143. Saeys, Y., Inza, I. & Larranaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **23,** 2507–2517 (2007).
144. Sharma, A., Imoto, S. & Miyano, S. A top-r feature selection algorithm for microarray gene expression data. *IEEE/ACM Trans Comput Biol Bioinform* **9,** 754–764 (2012).
145. Joachims, T. Making large-Scale SVM Learning Practical. in *Advances in Kernel Methods - Support Vector Learning* (eds. Schölkopf, B., Burges, C. & Smola, A.) (MIT-Press, 1999).
146. Seo, J. S. *et al.* The transcriptional landscape and mutational profile of lung adenocarcinoma. *Genome Res* **22,** 2109–2119 (2012).
147. Kim, S. K. *et al.* A nineteen gene-based risk score classifier predicts prognosis of colorectal cancer patients. *Mol Oncol* **8,** 1653–1666 (2014).
148. Yu, K. *et al.* A precisely regulated gene expression cassette potently modulates metastasis and survival in multiple solid cancers. *PLoS Genet* **4,** e1000129; doi: 10.1371/journal.pgen.1000129 (2008).
149. Steffens, S. *et al.* Clinical behavior of chromophobe renal cell carcinoma is less aggressive than that of clear cell renal cell carcinoma, independent of Fuhrman grade or tumor size. *Virchows Arch* **465,** 439–444 (2014).
150. Onishi, T., Ohishi, Y., Goto, H., Suzuki, M. & Miyazawa, Y. Papillary renal cell carcinoma: clinicopathological characteristics and evaluation of prognosis in 42 patients. *BJU Int* **83,** 937–943 (1999).
151. Muller, P. A. & Vousden, K. H. p53 mutations in cancer. *Nat Cell Biol* **15,** 2–8 (2013).
152. Karakas, B., Bachman, K. E. & Park, B. H. Mutation of the PIK3CA oncogene in human cancers. *Br J Cancer* **94,** 455–459 (2006).
153. Kleinerman, R. A. *et al.* Hereditary retinoblastoma and risk of lung cancer. *J Natl Cancer Inst* **92,** 2037–2039 (2000).
154. Menendez, D., Inga, A. & Resnick, M. A. The expanding universe of p53 targets. *Nat Rev Cancer* **9,** 724–737 (2009).

155. Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* **11,** 733–739 (2010).
156. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).
157. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11,** R106; doi: 10.1186/gb-2010-11-10-r106 (2010).
158. Bodenhofer, U., Kothmeier, A. & Hochreiter, S. APCluster: an R package for affinity propagation clustering. *Bioinformatics* **27,** 2463–2464 (2011).
159. Frey, B. J. & Dueck, D. Clustering by passing messages between data points. *Science* **315,** 972–976 (2007).
160. Xiong, Q., Mukherjee, S. & Furey, T. S. GSAASeqSP: a toolset for gene set association analysis of RNA-Seq data. *Sci Rep* **4,** 6347; doi: 10.1038/srep06347 (2014).
161. Zhang, B., Kirov, S. & Snoddy, J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res* **33,** W741–748 (2005).
162. Wang, J., Duncan, D., Shi, Z. & Zhang, B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Res* **41,** W77–83 (2013).
163. Kanehisa, M. *et al.* Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* **42,** D199–205 (2014).
164. Cerami, E. G. *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res* **39,** D685–690 (2011).
165. Killock, D. Lung cancer: alternative rearrangements–targeting ROS1 in NSCLC. *Nat Rev Clin Oncol* **11,** 624; doi: 10.1038/nrclinonc.2014.180 (2014).
166. Bergethon, K. *et al.* ROS1 rearrangements define a unique molecular class of lung cancers. *J Clin Oncol* **30,** 863–870 (2012).
167. Campo, I. *et al.* A large kindred of pulmonary fibrosis associated with a novel ABCA3 gene variant. *Respir Res* **15,** 43; doi: 10.1186/1465-9921-15-43 (2014).
168. Wambach, J. A. *et al.* Genotype-phenotype correlations for infants and children with ABCA3 deficiency. *Am J Respir Crit Care Med* **189,** 1538–1543 (2014).
169. Agrawal, A. *et al.* An intronic ABCA3 mutation that is responsible for respiratory disease. *Pediatr Res* **71,** 633–637 (2012).
170. Gower, W. A. *et al.* Fatal familial lung disease caused by ABCA3 deficiency without identified ABCA3 mutations. *J Pediatr* **157,** 62–68 (2010).
171. Xie, Y. *et al.* Aquaporin 1 and aquaporin 4 are involved in invasion of lung cancer cells. *Clin Lab* **58,** 75–80 (2012).
172. Warth, A. *et al.* Loss of aquaporin-4 expression and putative function in non-small cell lung cancer. *BMC Cancer* **11,** 161; doi: 10.1186/1471-2407-11-161 (2011).
173. Yang, Q. *et al.* STAT3 activation and aberrant ligand-dependent sonic hedgehog signaling in human pulmonary adenocarcinoma. *Exp Mol Pathol* **93,** 227–236 (2012).
174. Zhou, X. *et al.* Identification of a chronic obstructive pulmonary disease genetic determinant that regulates HHIP. *Hum Mol Genet* **21,** 1325–1335 (2012).
175. Li, X. *et al.* Importance of hedgehog interacting protein and other lung function genes in asthma. *J Allergy Clin Immunol* **127,** 1457–1465 (2011).
176. Jonsson, A. L., Simonsen, U., Hilberg, O. & Bendstrup, E. Pulmonary alveolar microlithiasis: two case reports and review of the literature. *Eur Respir Rev* **21,** 249–256 (2012).
177. Ferreira Francisco, F. A., Pereira e Silva, J. L., Hochhegger, B., Zanetti, G. & Marchiori, E. Pulmonary alveolar microlithiasis. State-of-the-art review. *Respir Med* **107,** 1–9 (2013).
178. Edmiston, J. S. *et al.* Gene expression profiling of peripheral blood leukocytes identifies potential novel biomarkers of chronic obstructive pulmonary disease in current and former smokers. *Biomarkers* **15,** 715–730 (2010).
179. Hawkins, G. A. *et al.* The IL6R variation Asp(358)Ala is a potential modifier of lung function in subjects with asthma. *J Allergy Clin Immunol* **130,** 510–515 e511 (2012).

Acknowledgements

The results published here are based on data generated by the TCGA Research Network: http://cancergenome.nih.gov/. This work was supported by the Open Fund of State Key Laboratory of Silkworm Genome Biology (sklsgb2013005) (Q.X.).

Author Contributions

Q.X. designed the study. L.P., D.K.L. and Q.X. performed experiments. L.P., X.W.B., D.K.L., C.X., G.M.W., Q.Y.X. and Q.X. contributed to the interpretation of results and manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Peng, L. *et al.* Large-scale RNA-Seq Transcriptome Analysis of 4043 Cancers and 548 Normal Tissue Controls across 12 TCGA Cancer Types. *Sci. Rep.* **5**, 13413; doi: 10.1038/srep13413 (2015).