A pathology atlas of the human cancer transcriptome

Citation for the original published paper (version of record):
Uhlen, M., Zhang, C., Lee, S. et al (2017). A pathology atlas of the human cancer transcriptome. Science, 357(6352): 660-+. http://dx.doi.org/10.1126/science.aan2507

(article starts on next page)

RESEARCH ARTICLE SUMMARY

CANCER

A pathology atlas of the human cancer transcriptome

Mathias Uhlen,* Cheng Zhang, Sunjae Lee, Evelina Sjöstedt, Linn Fagerberg, Gholamreza Bidkhori, Rui Benfeitas, Muhammad Arif, Zhengtao Liu, Fredrik Edfors, Kemal Sanli, Kalle von Feilitzen, Per Oksvold, Emma Lundberg, Sophia Hober, Peter Nilsson, Johanna Mattsson, Jochen M. Schwenk, Hans Brunnström, Bengt Glimelius, Tobias Sjöblom, Per-Henrik Edqvist, Dijana Djureinovic, Patrick Micke, Cecilia Lindskog, Adil Mardinoglu,† Fredrik Ponten†

INTRODUCTION: Cancer is a leading cause of death worldwide, and there is great need to define the molecular mechanisms driving the development and progression of individual tumors. The Hallmarks of Cancer has provided a framework for a deeper molecular understanding of cancer, and the focus so far has been on the genetic alterations in individual cancers, including genome rearrangements, gene amplifications, and specific cancer-driving mutations. Using systems-level approaches, it is now also possible to define downstream effects of individual genetic alterations in a genome-wide manner.

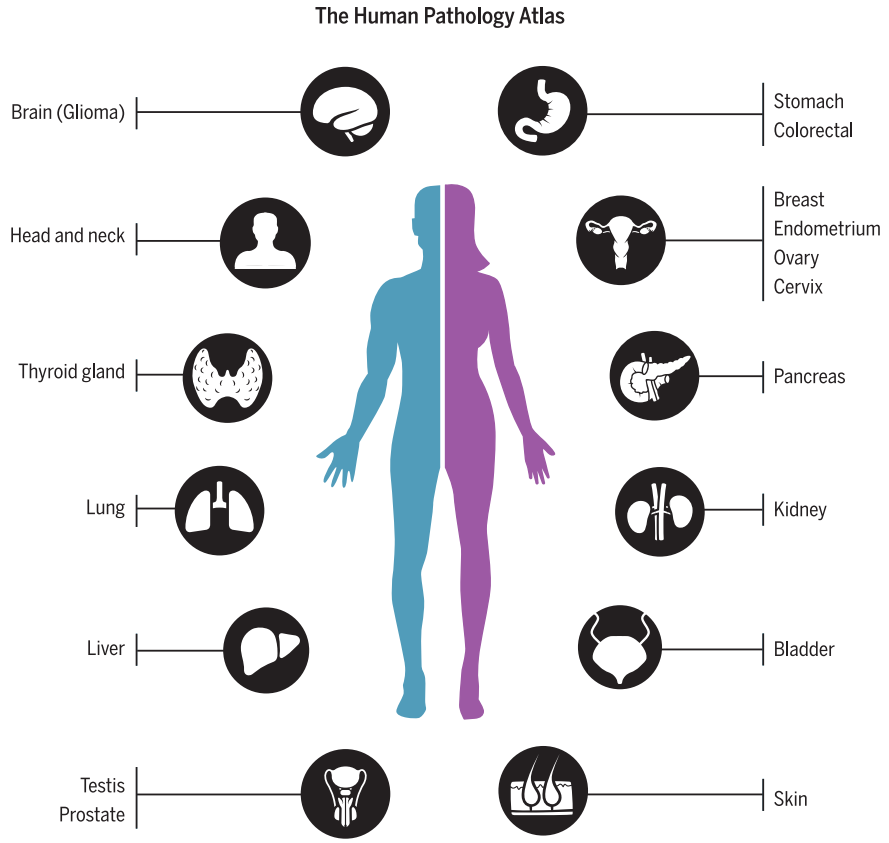
RATIONALE: In our study, we used a systems-level approach to analyze the transcriptome of 17 major cancer types with respect to clinical outcome, based on a genome-wide transcriptomics analysis of ~8000 individual patients with clinical metadata. The study was made possible through the availability of large open-access knowledge-based efforts such as the Cancer Genome Atlas and the Human Protein Atlas. Here, we used the data to perform a systems-level analysis of 17 major human cancer types, describing both interindividual and intertumor variation patterns.

RESULTS: The analysis identified candidate prognostic genes associated with clinical outcome for each tumor type; the results show that a large fraction of cancer protein-coding genes are differentially expressed and, in many cases, have an impact on overall patient survival. Systems biology analyses revealed that gene expression of individual tumors within a particular cancer varied considerably and could exceed the variation observed between distinct cancer types. No general prognostic gene necessary for clinical outcome was applicable to all cancers. Shorter patient survival was generally associated with up-regulation of genes involved in mitosis and cell growth and down-regulation of genes involved in cellular differentiation. The data allowed us to generate personalized genome-scale metabolic models for cancer patients to identify key genes involved in tumor growth. In addition, we explored tissue-specific genes associated with the dedifferentiation of tumor cells and the role of specific cancer testis antigens on a genome-wide scale. For lung and colorectal cancer, a selection of prognostic genes identified by the systems biology effort were analyzed in independent, prospective cancer cohorts using immunohistochemistry to validate the gene expression patterns at the protein level.

CONCLUSION: A Human Pathology Atlas has been created as part of the Human Protein Atlas program to explore the prognostic role of each protein-coding gene in 17 different cancers. Our atlas uses transcriptomics and antibody-based profiling to provide a standalone resource for cancer precision medicine. The results demonstrate the power of large systems biology efforts that make use of publicly available resources. Using genome-scale metabolic models, cancer patients are shown to have widespread metabolic heterogeneity, highlighting the need for precise and personalized medicine for cancer treatment. With more than 900,000 Kaplan-Meier plots, this resource allows exploration of the specific genes influencing clinical outcome for major cancers, paving the way for further in-depth studies incorporating systems-level analyses of cancer. All data presented are available in an interactive open-access database (www.proteinatlas.org/pathology) to allow for genome-wide exploration of the impact of individual proteins on clinical outcome in major human cancers. ■

ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aan2507



Schematic overview of the Human Pathology Atlas. A systems-level approach enables analysis of the protein-coding genes of 17 different cancer types from ~8000 patients. Results are available in an interactive open-access database.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: mathias.uhlen@scilifelab.se
†These authors contributed equally to this work.
Cite this article as M. Uhlen *et al.*, *Science* **357**, eaan2507 (2017). DOI: 10.1126/science.aan2507

RESEARCH ARTICLE

CANCER

A pathology atlas of the human cancer transcriptome

Mathias Uhlen,[1,2,3]* Cheng Zhang,[1] Sunjae Lee,[1] Evelina Sjöstedt,[1,4] Linn Fagerberg,[1] Gholamreza Bidkhori,[1] Rui Benfeitas,[1] Muhammad Arif,[1] Zhengtao Liu,[1] Fredrik Edfors,[1] Kemal Sanli,[1] Kalle von Feilitzen,[1] Per Oksvold,[1] Emma Lundberg,[1] Sophia Hober,[3] Peter Nilsson,[1] Johanna Mattsson,[4] Jochen M. Schwenk,[1] Hans Brunnström,[5] Bengt Glimelius,[4] Tobias Sjöblom,[4] Per-Henrik Edqvist,[4] Dijana Djureinovic,[4] Patrick Micke,[4] Cecilia Lindskog,[4] Adil Mardinoglu,[1,3,6]† Fredrik Ponten[4]†

Cancer is one of the leading causes of death, and there is great interest in understanding the underlying molecular mechanisms involved in the pathogenesis and progression of individual tumors. We used systems-level approaches to analyze the genome-wide transcriptome of the protein-coding genes of 17 major cancer types with respect to clinical outcome. A general pattern emerged: Shorter patient survival was associated with up-regulation of genes involved in cell growth and with down-regulation of genes involved in cellular differentiation. Using genome-scale metabolic models, we show that cancer patients have widespread metabolic heterogeneity, highlighting the need for precise and personalized medicine for cancer treatment. All data are presented in an interactive open-access database (www.proteinatlas.org/pathology) to allow genome-wide exploration of the impact of individual proteins on clinical outcomes.

Cancer is one of the leading causes of death worldwide, and both the incidence and prevalence of cancer continue to increase. Most current cancer drugs are effective only in a subgroup of patients owing to interindividual tumor heterogeneity, and large gaps remain in our current understanding of the best treatment approaches and the underlying molecular mechanisms driving cancer pathogenesis (*1*). There is therefore an urgent need for the development of personalized diagnostic and therapeutic strategies using methods such as systems-level analysis (*2*–*4*). Such approaches can be used to study the genome-wide effect of gene rearrangements, amplifications, and specific cancer-driving mutations on protein-coding regions.

Thanks to large open-access knowledge-based efforts, such as The Cancer Genome Atlas (TCGA) (*5*), the Human Protein Atlas (HPA) (*6*), the GTEx consortium (*7*), and recount2 (*8*), it is now possible to explore the genome-wide expression of individual genes in different tissues and cancers (*9*). The database resource from TCGA represents a comprehensive and coordinated effort to accelerate our understanding of cancer (*5*), and the HPA and GTEx represent international efforts to map the expression of protein-coding genes in normal human tissues. Many of the patients included in the TCGA database are also accompanied by clinical survival metadata, allowing clinical outcomes to be associated with genome-wide expression patterns of protein-coding genes and metabolic modeling of individual cancer patients. Such analysis is facilitated by the recent suggestion that there is a gene-specific correlation between RNA and protein levels in human tissues and cells, allowing quantitative analyses of mRNA levels to be used as proxies for the corresponding protein levels (*10*).

Here, we used data from TCGA and the HPA efforts to perform a systems-level analysis of 17 major human cancer types corresponding to 7932 tumor samples, and describe both interindividual and intertumor variation patterns. The analysis identified candidate prognostic genes associated with clinical outcome for each tumor type and generated metabolic models for individual patients. A Human Pathology Atlas has been created as part of the Human Protein Atlas program to explore the prognostic role of each protein-coding gene in each cancer type by means of transcriptomics and antibody-based profiling (Fig. 1A). More than 100 million Kaplan-Meier survival plots were generated as part of the genome-wide analysis of potential prognostic genes in these cancers. More than 900,000 survival plots—each accompanied with statistical significance—can be visualized at the new pathology resource.

To investigate the key prognostic genes affecting patient survival, we generated cancer-specific coexpression networks for each of the studied cancer types and examined the functional relationship between the prognostic genes and the genes associated with Hallmarks of Cancer (*11*). Personalized genome-scale metabolic models (GSMMs) for the tumors in each cancer patient were generated to study the individual metabolic differences among tumors. This analysis also allowed us to study the role of tissue-specific genes in the "dedifferentiation" of cancer and the role of specific cancer testis antigens (CTAs) on a genome-wide scale. For two of the cancer types, lung and colorectal cancer, a selection of prognostic genes identified by the systems biology effort were analyzed in independent prospective cancer cohorts, using immunohistochemistry (IHC) to validate the gene expression patterns at the protein level.

All primary Human Pathology Atlas data are freely available without restrictions in the public open access database (www.proteinatlas.org/pathology) that is part of the Human Protein Atlas program. Significant prognostic genes in each cancer type are highlighted together with Kaplan-Meier plots based on overall survival and accompanied with data for individual gene expression heterogeneity of prognostic genes at the time of diagnosis.

Transcriptome analysis of human cancers

We retrieved RNA sequencing (RNA-seq) data together with clinical metadata corresponding to the 33 different human cancers that are available in TCGA (table S1). As a result, data were collected from 9666 individuals out of the 11,000 cancer patients included in the TCGA project from the Genomic Data Commons (GDC) Data Portal (https://gdc-portal.nci.nih.gov/). First, using hierarchical clustering, we investigated the relationship between the global gene expression patterns of all protein-coding genes in the 33 cancer types (n = 19,571) and the gene expression patterns in 37 normal human tissues obtained from 162 healthy subjects in the HPA project (*6*) (fig. S1). RNA-seq data from all cancer tissues and all normal tissues were processed in the same bioinformatics pipeline and normalized as fragments per kilobase of exon per million fragments mapped (FPKM). We found that a majority of all cancers (26 of 33) clustered in the same group, while the majority of the normal tissues (33 of 37) clustered in a different group, indicating that most cancer types share expression features that render them significantly different from normal tissues. Notably, we found that liver tissue and the primary form of liver cancer, hepatocellular carcinoma, as well as bone marrow and acute myeloid leukemia clustered together, suggesting that these phenotypes are more closely related independent of a benign or malignant status.

We previously classified all protein-coding genes into six different categories according to their expression across normal tissues and organs (*6*). The classification, based on a FPKM cut-off >1, ranged from genes expressed in all tissues to those

[1]Science for Life Laboratory, KTH–Royal Institute of Technology, Stockholm, Sweden. [2]Center for Biosustainability, Danish Technical University, Copenhagen, Denmark. [3]School of Biotechnology, AlbaNova University Center, KTH–Royal Institute of Technology, Stockholm, Sweden. [4]Department of Immunology Genetics and Pathology, Uppsala University, Uppsala, Sweden. [5]Division of Pathology, Lund University, Skåne University Hospital, Lund, Sweden. [6]Department of Biology and Biological Engineering, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden.
***Corresponding author. Email: mathias.uhlen@scilifelab.se**
†These authors contributed equally to this work.

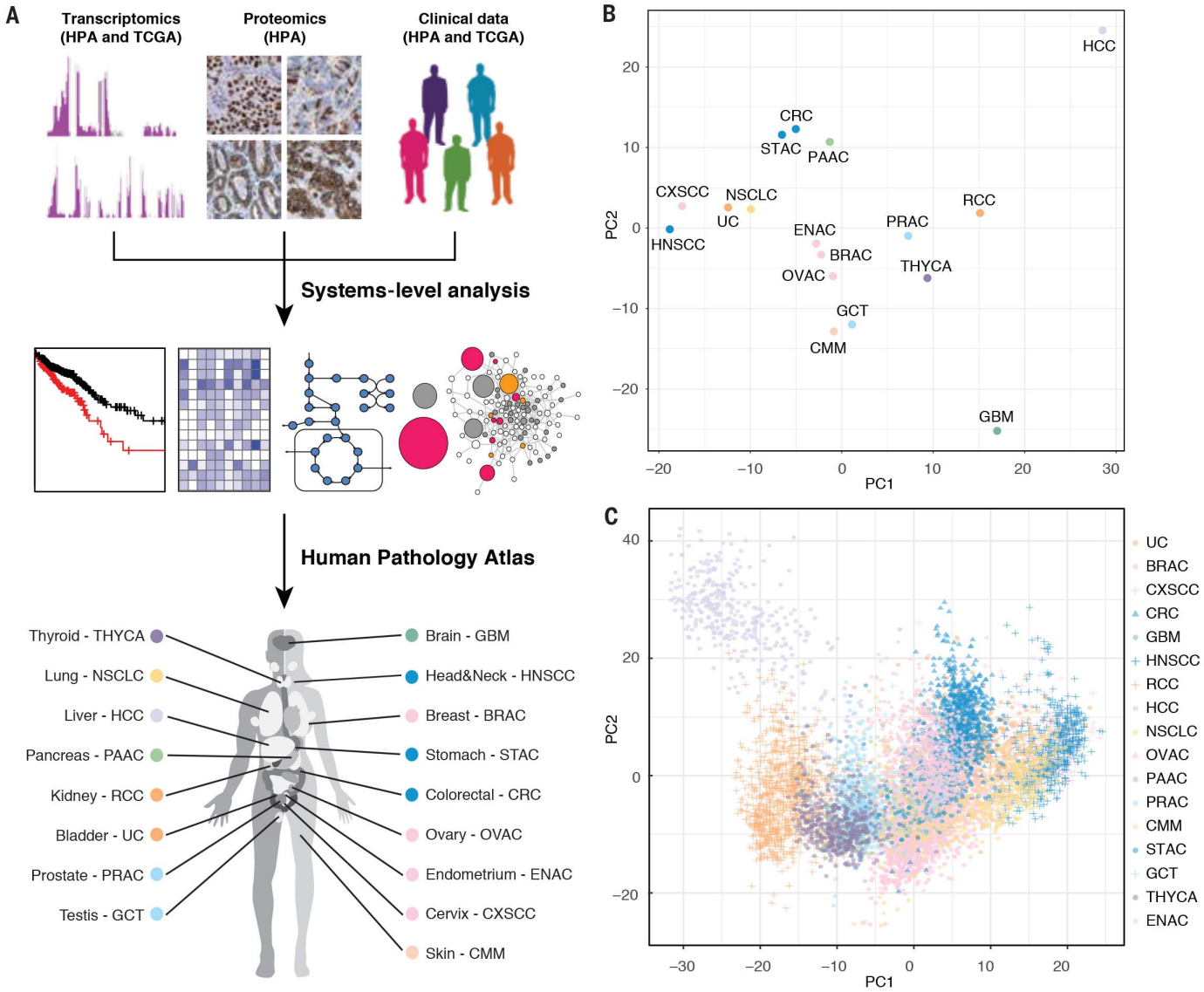


Fig. 1. Analysis of the global expression patterns of protein-coding genes in human cancers. (**A**) Schematic drawing of the Human Pathology Atlas effort described herein. (**B**) Principal components analysis (PCA) showing the similarities in expression of 19,571 protein-coding genes among 17 cancer types. See fig. S4 for additional PCA analysis with more stratified patient cohorts. (**C**) PCA plot showing the individual differences in the genome-wide global expression profiles among the 17 cancer types in 9666 individual patients.

with tissue-restricted expression and those not detected in any of the analyzed tissues. The transcriptomics data for the 33 different cancers allowed us to classify the protein-coding genes into six different categories based on the expression level. Our analysis revealed that a large fraction (41%) of the protein-coding genes were expressed in all analyzed cancers, while approximately 46% (n = 9057) displayed more tumor type-restricted expression. Among the protein-coding genes, 13% were not detected in any tumor types investigated (fig. S2 and table S2). The majority of the genes (n = 5772) detected in all samples were shared between cancers and normal tissues, whereas 2401 additional genes were expressed in all cancers analyzed, but with more restricted expression in the normal tissues. These "housekeeping" genes in tumors are enriched in biological functions related to DNA replication and the regulation of apoptosis and mitosis (table S3 and fig. S3).

Subsequently, we focused our analysis on 17 tumor types with large numbers of patients available in the TCGA data set accompanied by clinical metadata (Fig. 1A and table S4). The connectivity among these 17 cancers was determined using principal components analysis (PCA) based on the expression pattern of all protein-coding genes (Fig. 1B and fig. S4). We observed a relationship among cancer types that shared a similar tissue type of origin or similar morphological features and phenotypic expression patterns. For example, cancers with a dominating squamous cell carcinoma phenotype, such as cervical or head and neck cancer, clustered together close to the related urothelial cell carcinoma and non–small cell lung cancer (NSCLC), which also contains a large fraction of squamous cell carcinoma. Adenocarcinomas that originate from the gastrointestinal tract, including pancreatic cancer, also clustered separately from the cluster containing the three adenocarcinomas representing female cancer (i.e., breast, endometrial, and ovarian cancer). Interestingly, testicular germ cell tumors were located close to melanoma and were well separated from the more classical epithelial tumor types, whereas glioma (brain) and hepatocellular (liver) carcinoma clearly represented the most divergent tumor types in this global expression analysis.

Individual variation among cancers

To determine the individual gene expression patterns within and among certain cancer types, we used PCA to visualize the global expression patterns for all 9666 individual tumors that were

included in the patient cohorts, representing the 17 major cancer types (Fig. 1C). The results showed that the interindividual variation within each type of cancer was considerable, and that there was a large overlap in expression among individuals with different cancer types. One exception was liver cancer (Fig. 1C, upper left), in which the individual tumors showed relatively unique global expression patterns with little overlap with the other cancer types. Thus, gene expression varies considerably in individual tumors within a particular cancer subtype. For some patient tumors, the global expression pattern resembles other cancer types more than it does the given type of diagnosed cancer, which reinforces previous discoveries (12).

Clinical outcome based on gene expression analysis

First, we analyzed the survival data from the TCGA metadata (fig. S5 and table S4). Prostate cancer and testis cancer (germ cell tumors) have the most favorable 3-year survival rates (98% and 97%, respectively), while high-grade glioma and pancreatic cancer have the lowest 3-year survival rates (8% and 35%, respectively). The patient survival data and matched transcriptomic data enabled us to perform gene-centric and genome-wide survival analyses to identify prognostic genes across the 17 cancer types. For each cancer, all patients with survival data were included in the Kaplan-Meier survival analysis spanning 10 years as extracted from the metadata. The RNA levels at the time of diagnosis were plotted against the survival data as extracted from the follow-up clinical data (see examples in Fig. 2A). For each gene and cancer type, the patient cohort was stratified into two groups with the highest and lowest expression (FPKM) based on individual expression levels. To choose the best FPKM cutoffs for grouping the patients most significantly, we used all FPKM values from the 20th to 80th percentiles to group the patients, examined significant differences in the survival outcomes of the groups, and selected the value yielding the lowest log-rank P value. In total, more than 100 million Kaplan-Meier plots were generated that corresponded to all 19,571 protein-coding genes across the 17 cancer types. As a comparison, we also tested the method described by Hothorn and Lausen (13) and the results were highly similar (fig. S6). Two examples of genes in the liver cancer cohort are shown in Fig. 2B, including the survival data for the individual patients in the liver cancer cohort.

We identified two types of prognostic marker genes in terms of clinical outcome: (i) unfavorable prognostic genes, for which higher expression of a given gene was correlated with a poor patient survival outcome, and (ii) favorable prognostic genes, for which higher expression of a given gene was correlated with a longer patient survival outcome. A prognostic gene for a given cancer was defined as a gene for which the expression level above or below the experimentally determined cutoff in an individual patient yields a significant ($P < 0.001$) difference in overall survival. The ratios of favorable and unfavorable prognostic genes varied among the different types of cancer. In Fig. 2C, the numbers of prognostic genes for each of the 17 cancer types are shown, with more detailed information provided in table S5. It is noteworthy that 2375 genes showed opposite effects on prognosis depending upon cancer type and location, highlighting the need to perform functional studies of prognostic genes. See table S6 for a complete list of the prognostic association of all genes in all cancers.

In Fig. 2A, examples of favorable and unfavorable prognostic genes are shown for five of the cancer types, based on the optimal stratification P value calculated for each gene and cancer. In each case, a significant separation ($P < 0.001$) of the survival rate could be observed on the basis of differences in the expression levels of the respective gene. For some genes, the prognostic value has previously been reported in the literature; one example is RBM3 (RNA binding motif protein 3) (Fig. 2A), which has been implicated in survival of colorectal cancer (14). However, most of the identified prognostic genes lacked prior reports of a survival link to a given cancer, making them potential candidates for follow-up studies.

We extended the survival analysis by constructing panels of the five most significant favorable and unfavorable prognostic genes (table S7) for each tumor type and used them to predict the clinical outcome (Fig. 2A). Each of the five panels generated a prognostic panel of high significance ($P < 10^{-5}$). Similarly, all of the other 12 cancer types yielded prognostic panels in the same manner with very high significance (table S7). It is noteworthy that for cancers with more favorable survival rates (e.g., testicular or prostate cancer), a limited number of prognostic genes have been identified, perhaps because the 3-year survival probability for these cancers exceeds 95% and thus larger patient cohorts are needed to obtain prognostic genes with high significance. For two of the tumors (i.e., renal and liver cancer), the numbers of prognostic genes were much larger than for the other cancers (6070 and 2892, respectively) (Fig. 2C). This observation is interesting because both are cancers with distinct features and morphology, and liver cancer especially appears to be distantly related to other cancer types (Fig. 1B). For renal cancer, the number of favorable (n = 2782) and unfavorable genes (n = 3288) was balanced, whereas there were a large number of unfavorable prognostic genes (n = 2629) for liver cancer. An earlier study of renal cancer based on TCGA data showed distinctly different groups of patients that are not reflected by morphological subtypes (e.g., clear cell, papillary, and chromophobe phenotypes) (15). Thus, the large number of prognostic genes may simply reflect large global expression differences between these two subtypes, resulting in a large number of "passenger" genes and a much smaller set of driver genes affecting the clinical course of the patient.

Overlap of prognostic genes across cancer types

We examined the extent of overlap of prognostic genes among different cancer types. The correlation among the 17 cancer types for favorable and unfavorable prognostic genes was investigated in a pairwise manner (Fig. 3A). For most cancers, little correlation was observed, suggesting a relatively limited number of common prognostic genes. In contrast, a significant overlap of favorable prognostic genes was observed for other cancers (e.g., renal, breast, lung, and pancreatic cancers). Similarly, unfavorable prognostic genes for some cancers, including renal, liver, lung, and pancreatic cancer, clustered together. However, a detailed analysis revealed that no prognostic genes were shared among more than 7 of the cancer types (table S8).

Functional analysis of prognostic genes

A functional gene ontology (GO) analysis was performed for the most significant prognostic genes shared among the 17 major cancers, including both favorable and unfavorable genes (table S9). The results (Fig. 3B) suggest that many of the common unfavorable genes are related to cell proliferation, including mitosis, cell cycle regulation, and nucleic acid metabolism. In contrast, few GO functions were significantly overrepresented by the common favorable genes; the most enriched GO functions were positive regulation of cell activation, regulation of immune cell activation, and cell-cell adhesion.

Because genes associated with proliferation were identified by the functional analysis, we investigated the prognostic effect of all 314 cell cycle genes defined by the Molecular Signature database (16) in various cancer types. Interestingly, more than 60% (n = 194) of these genes were associated with an unfavorable clinical outcome, with increased expression in at least one of the analyzed cancer types (table S10). However, these prognostic cell cycle genes were generally only shared among a few cancers (Fig. 3C), which suggests that although cell cycle genes are commonly unfavorable genes, the use of a particular set of cell cycle genes and their effect on clinical outcome may differ among individual cancer types.

Tissue-enriched genes and dedifferentiation in cancer

We further analyzed genes with high relative expression that correlated with prolonged overall survival, for which a high expression level of a particular gene was associated with a good clinical outcome. Many of these favorable genes have previously (6) been classified as elevated in certain normal tissues (table S11), as exemplified in liver cancer (Fig. 3D), for which more than half (n = 150) of the 263 favorable prognostic genes were defined as tissue-elevated. To further investigate the molecular signatures related to differentiation, we analyzed alterations in liver-enriched genes (n = 154) defined by tissue-wide expression studies of normal hepatocytes. Samples from normal liver tissue were analyzed and compared with the transcriptomics patterns of the primary liver cancer biopsies and the liver cancer–derived HepG2 cell line. To further compare the expression levels of the tissue-enriched proteins, we plotted

the genome-wide transcriptomics data using the relative changes between cancer/normal tissue and cell line/normal tissue, respectively, for all genes expressed in the normal liver. The liver-enriched genes (red), liver group–enriched genes (orange), and all other expressed genes (black) are summarized in Fig. 4A. The global analysis demonstrates a down-regulation in both the liver cancer and the cancer cell line as compared with the expression levels in normal liver tissue (lower left quadrant). This quadrant contains 102 of the 154 liver-enriched genes (66%), which suggests that liver-enriched genes are down-regulated as a sign of dedifferentiation in both liver cancer and liver cancer cell lines.

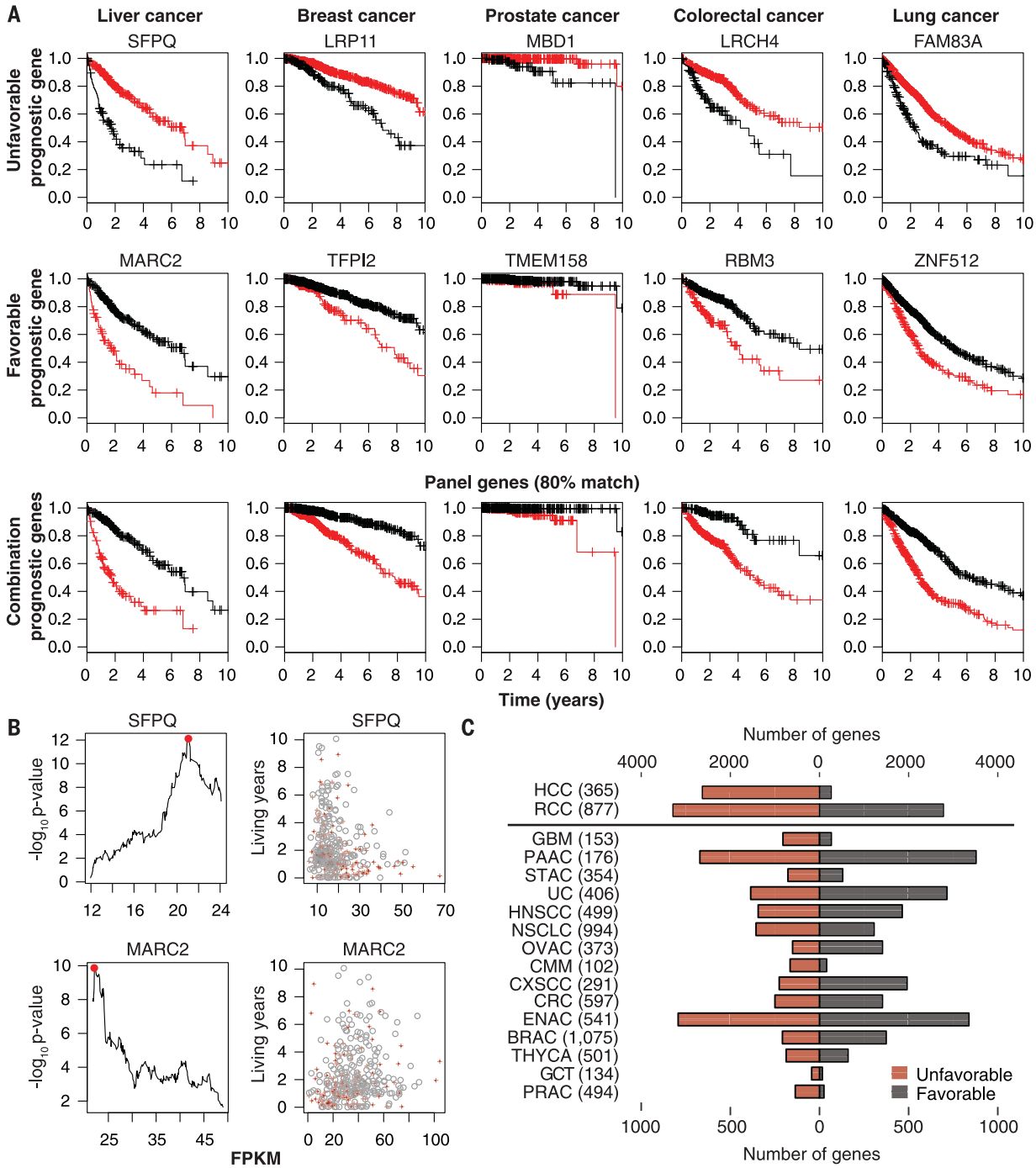


Fig. 2. Identification of prognostic genes based on expression coupled with clinical survival for 17 different cancer types. (**A**) Examples of Kaplan-Meier plots for five major cancer patients stratified by the expression of an unfavorable prognostic gene (first row), a favorable prognostic gene (second row), and a combination of 10 prognostic genes (third row). The selected unfavorable and favorable genes had the best log-rank *P* value based on the Kaplan-Meier analysis, with average RNA expression levels more than the median average expression of all protein-coding genes; the 10 marker genes were a combination of the top five favorable and unfavorable genes with expression higher than the median average expression. Black and red lines show high and low (or, in the third row, favorable and unfavorable) expression, respectively. (**B**) Examples of two prognostic genes in liver cancer. Left: Distribution of log-rank *P* values against the RNA expression with different RNA-level (FPKM) cutoffs. Right: Patient-centric scatterplot showing the relationships between living years and RNA expression of the prognostic genes. (**C**) Numbers of genes showing favorable and unfavorable prognostic effects in the 17 Human Pathology Atlas cancer types. Patient numbers for each cancer are shown in parentheses.

Metadata for the grade of malignancy (i.e., the degree of differentiation) are available in the TCGA database, and this allowed us to analyze the relative expression level of liver-enriched genes in liver cancer and to compare different grades of malignancy. The tumor grade was scored using the modified nuclear grading scheme outlined by Edmondson and Steiner (17), with the tumor grade categorized as low, intermediate, or high. The malignancy grade (G1 to G3) (18) was available for 341 cases. The analysis revealed a significant correlation between the malignancy grade and the expression pattern of liver-enriched genes that were significantly down-regulated in liver cancer. In Fig. 4B, examples of IHC-based protein expression levels of a liver-enriched gene (CYP2C9) are displayed for normal liver versus liver cancer with differing tumor grade. The gene expression levels of CYP2C9 across all patients are also shown as box plots for different tumor grades (Fig. 4C). In addition, we analyzed the distribution of correlation coefficients for all analyzed liver-enriched genes compared with that of a randomly selected set of genes (Fig. 4D). Randomly selected genes showed no correlation (median rho = 0.07), whereas the tissue-enriched genes showed a negative correlation, with reduced

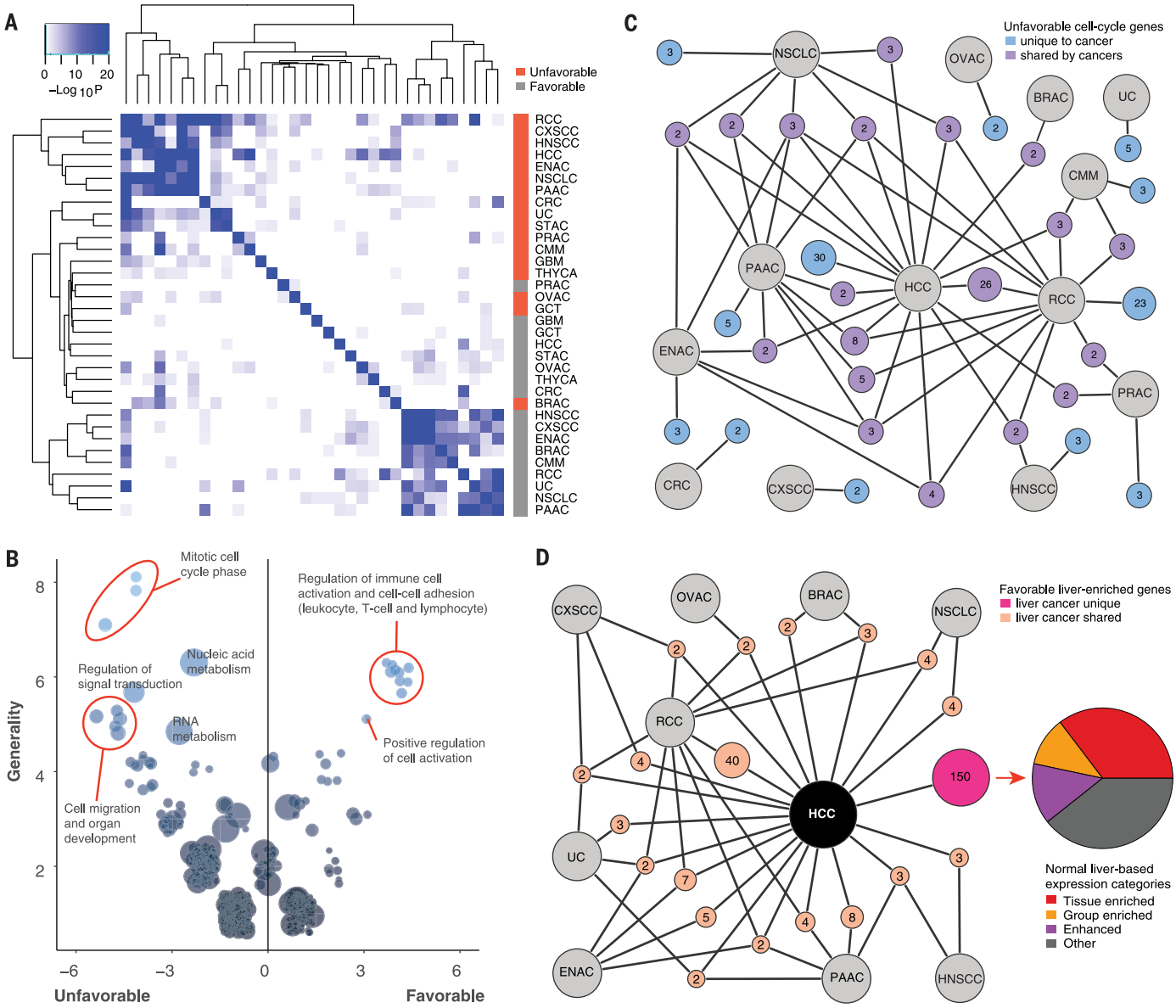


Fig. 3. Network analysis of prognostic genes. (**A**) Heat map showing the hypergeometric *P* value for the pairwise overlap of prognostic genes between the cancer types. (**B**) Bubble plot showing the common enriched Gene Ontology (GO) functions among the 17 Human Pathology Atlas cancer types. Bubble sizes represent numbers of genes in GO function; the *x* and *y* axes indicate the directionalities and generalities of the GO terms. Generality is defined by the number of cancers with their prognostic genes overrepresenting the GO function; directionality is defined by the number of cancers with their favorable genes overrepresenting the GO function minus the number of cancers with unfavorable genes overrepresenting the GO function. Note that only functions with more than five generalities are labeled. All GO terms for each cancer are provided in table S9. Results based on optional *P* value or hazard ratio cutoff–defined prognostic genes are provided in fig. S7 and table S9. (**C**) Network plot showing the number of cancer-specific and shared unfavorable cell cycle genes in all cancer types. Note that all groups with only one gene were removed from the plot. (**D**) Network plot showing the number of liver cancer–specific favorable genes and the favorable genes shared among liver and other cancers in the Human Pathology Atlas. Inset: Pie chart showing the fraction of elevated normal liver genes among the liver cancer–specific favorable genes.

expression of tissue-enriched genes in high-grade tumors (grade G3). The results demonstrated a molecular correlation between the expression levels of tissue-enriched genes and tumor grade, supporting the concept that dedifferentiated cancers are associated with decreased patient survival.

Cancer testis antigens in liver cancer

Cancer testis antigens are expressed in a wide range of cancer types, whereas their expression in normal tissues is restricted to immune-privileged sites such as the testis and placenta. To explore this observation further, we investigated the differential expression patterns of testis-enriched genes in normal liver, primary liver biopsies, and a liver cancer–derived cell line (HepG2). A global analysis, shown in Fig. 4E (upper right quadrant), showed that many of the testis-enriched genes had higher expression in the patient biopsy and cell line than in normal liver tissue. The results support many previous studies (*19*) that testis-enriched genes have higher expression in cancer than in the corresponding normal tissues.

Coexpression networks of human cancers

The Hallmarks of Cancer (*11*) has laid an important foundation for understanding cancer pathogenesis, and from the corresponding cellular processes, 2172 genes have recently been defined as hallmark-related genes (*16*, *20*). We thus decided to investigate their relationship with the prognostic genes reported here. Approximately two-thirds (65%) of the "hallmark genes" were predictive for clinical outcome in at least one of the cancers analyzed, but a network analysis revealed that none of the genes were shared among the majority of cancers, with most genes consequently affecting only a few of the cancer types (Fig. 5A and figs. S8 and S9). Subsequently, a cancer-specific coexpression network analysis for all 17 major cancers (table S12; available at http://inetmodels.com) was performed to identify genes that are expressed concurrently during tumorigenesis. Figure 5B shows a coexpression cluster in the lung cancer cohort, with enrichment for both prognostic and hallmark genes. Within this cluster, the hub genes (located in the center) are generally more prognostic than those with less coexpression. It is tempting to speculate

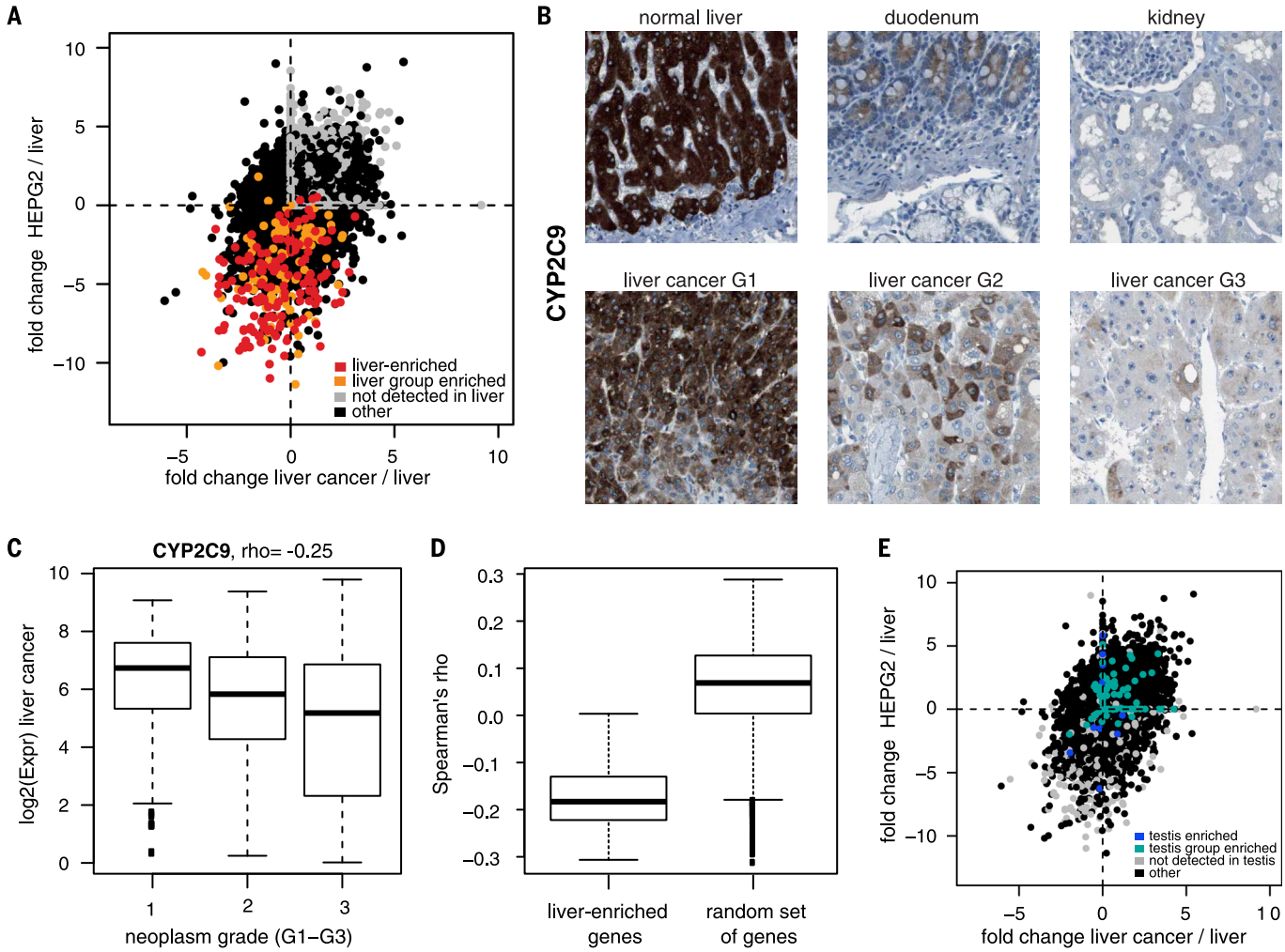


Fig. 4. Correlation between tumor differentiation and expression of liver-enriched genes. (**A**) Scatterplots showing the relative (fold) change between the transcript expression level in liver cancer and normal liver tissue (*x* axis) and the HepG2 cell line and normal tissue (*y* axis) for all protein-coding genes. Individual genes are colored according to their expression-based category in liver. All FPKM values less than 1 were set to 1 for the fold change calculation. (**B**) IHC staining of CYP2C9 proteins in four normal tissues and different hepatocellular carcinoma samples. For full IHC protein profiles, view the gene at www.proteinatlas.org/pathology. (**C**) Box plots showing the expression levels of liver tumor samples of different neoplasm grades for three representative liver-enriched genes for CYP2C9. (**D**) Box plot showing the distribution of correlation coefficients (Spearman's rho) between the neoplasm grade and expression for a random set of genes and all liver-enriched genes in liver tumors. (**E**) Scatterplots for all protein-coding genes showing the fold change in testis-specific antigen in liver cancer and normal liver tissue (*x* axis) and in the HepG2 cell line and normal liver tissue (*y* axis). Individual genes are colored according to their expression-based category in the testis.

that the hub genes in this cluster are lung cancer "drivers" and that the genes located around the outer boundary are lung cancer "passengers." Using somatic copy number alteration data in a TCGA pan-cancer analysis, we found that 36.4% of the genes in this cluster (table S13) were amplified or deleted in their chromosomal regions (*21*).

Among cancer-specific coexpression clusters, those that were significantly enriched with prognostic genes (hypergeometric test, $P \leq 0.05$) were named prognostic clusters, and an average of 13.9 clusters per cancer were enriched with prognostic genes (fig. S10 and table S14). A functional analysis, as exemplified by lung cancer (Fig. 5C and fig. S9), showed that many prognostic clusters were enriched with genes associated with the hallmark genes, such as those involved in DNA repair, cell proliferation, angiogenesis, and cell-cell signaling, implying that those processes or pathways may be associated with lung cancer progression. Across the 17 cancer types, the fractions of prognostic genes associated with the hallmark genes were determined (Fig. 5D and fig. S9); more than half (57% on average) of the prognostic genes were not identified as hallmark genes but were coexpressed with hallmark genes. It remains to be determined whether many of the prognostic genes identified herein have a passive or dominant role in the development of cancer.

Personalized metabolic networks for cancer patients

Tumors increase the nutrient import from the environment to fulfill biosynthetic demands

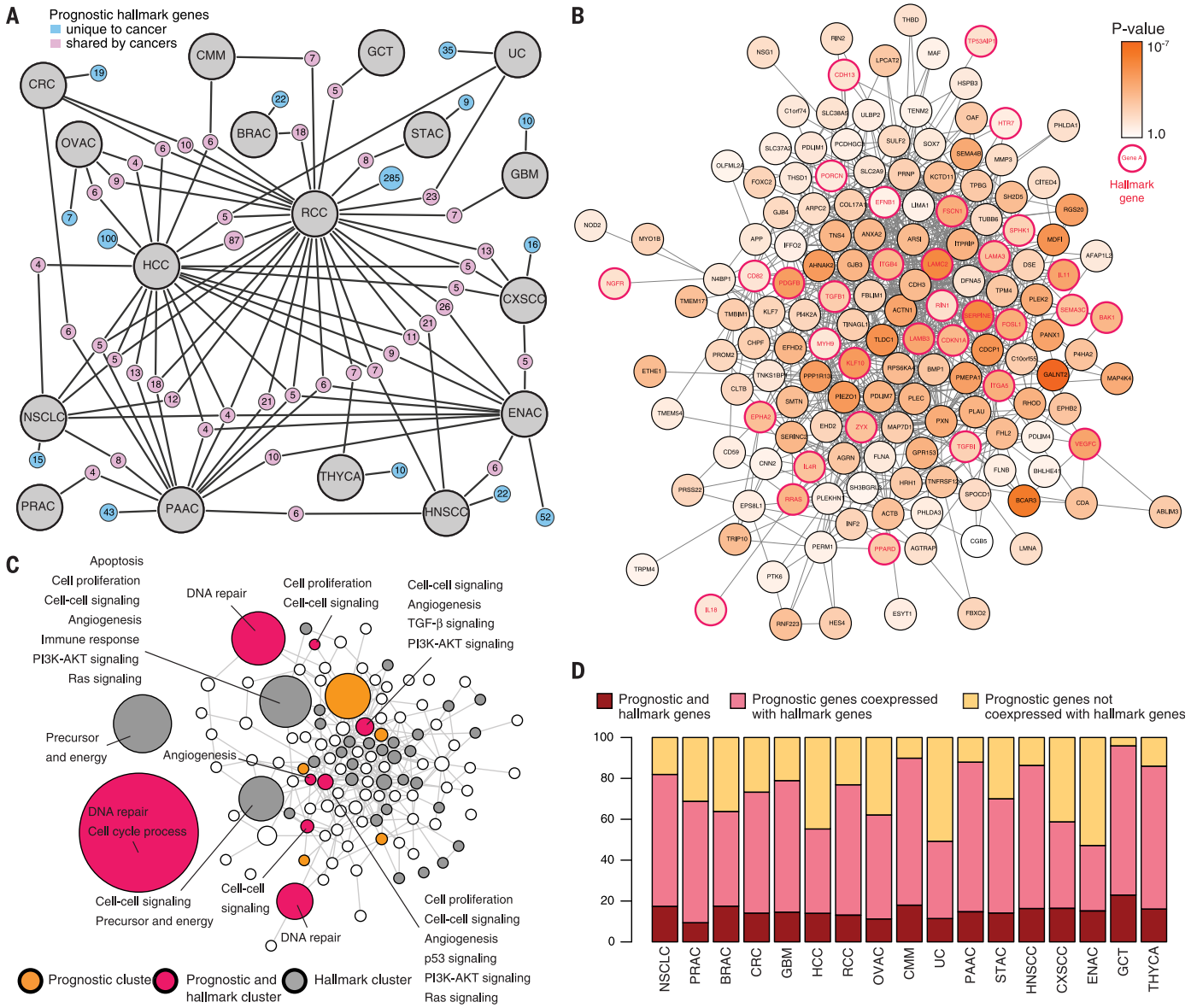


Fig. 5. Coexpression analysis reveals the relationship with the Hallmarks of Cancer and clues for drivers among prognostic genes. Gene coexpression of 17 cancers was investigated on the basis of established cancer coexpression networks. (**A**) Network plot showing the number of cancer-specific and shared prognostic cancer hallmark genes in all cancer types. Note that all groups with fewer than four genes were removed from the plot. (**B**) A gene coexpression cluster from the coexpression network of lung cancer enriched with both hallmark and prognostic genes. (**C**) Network plot showing coexpression clusters of lung cancer. All nodes indicate gene coexpression clusters; edges indicate significant coexpression links between clusters. The gray, yellow, and red color of the nodes indicates that the cluster was significantly enriched with hallmark genes, prognostic genes, and both cases, respectively. (**D**) Bar plot showing the fraction of prognostic genes that are mere hallmark genes (red), coexpressed in hallmark gene clusters (pink), or not coexpressed with hallmark genes (gold).

associated with proliferation, making use of these nutrients to both maintain viability and build new biomass (22–24). To investigate the metabolic reprogramming of each tumor, we generated personalized GSMMs for tumors from more than 7000 of the 17 major cancer patients based on transcriptomics data and generic human GSMM HMR2 (25) as previously described (26) (Fig. 6A). The resulting personalized GSMMs ranged in size from 2070 to 4058 metabolites, 2093 to 5261 reactions, and 978 to 2102 associated genes (fig. S11 and table S15). A total of 4889 metabolites, 6977 reactions, and 2760 genes were shared across the models; 1419 metabolites, 1020 of the reactions, and 334 of the genes were present in all personalized GSMMs. The personalized GSMMs are available in the BioModels Database (www.ebi.ac.uk/biomodels) with accession numbers MODEL1707110000 to MODEL1707116752.

Personalized GSMMs may allow for the investigation of common and unique biological functions for each patient (27). Using personalized GSMM and constraint-based modeling, we investigated heterogeneities of individual cancers by identifying genes within a tumor that are important for its growth (3). This method is suitable for studying cancer metabolism because it assumes an increase in tumor growth rate under optimal conditions and hence searches for metabolic flux distributions to produce essential biomass precursors at high rates (2, 28, 29). We found significant differences in the essential genes catalyzing tricarboxylic acid (TCA) cycle metabolism in liver cancer (Fig. 6B). As shown, the enzyme FH (fumarate hydratase) is identified as a conserved gene for tumor growth in all liver cancer patients analyzed, whereas SDHA (succinate dehydrogenase complex, subunit A) is important for tumor growth in ~60% of liver cancer patients, and ACLY (ATP citrate lyase) is key for tumor growth in fewer than 5% of liver cancer patients. In total, we identified 2553 essential genes that can inhibit or kill tumor growth in any of the analyzed samples and found that 55 (2%) of the key genes are common in all cancer patients analyzed, regardless of the cancer type (table S14). Notably, we found that only 10% to 25% of the essential genes were conserved in more than 80% of patients of each cancer type (Fig. 6C).

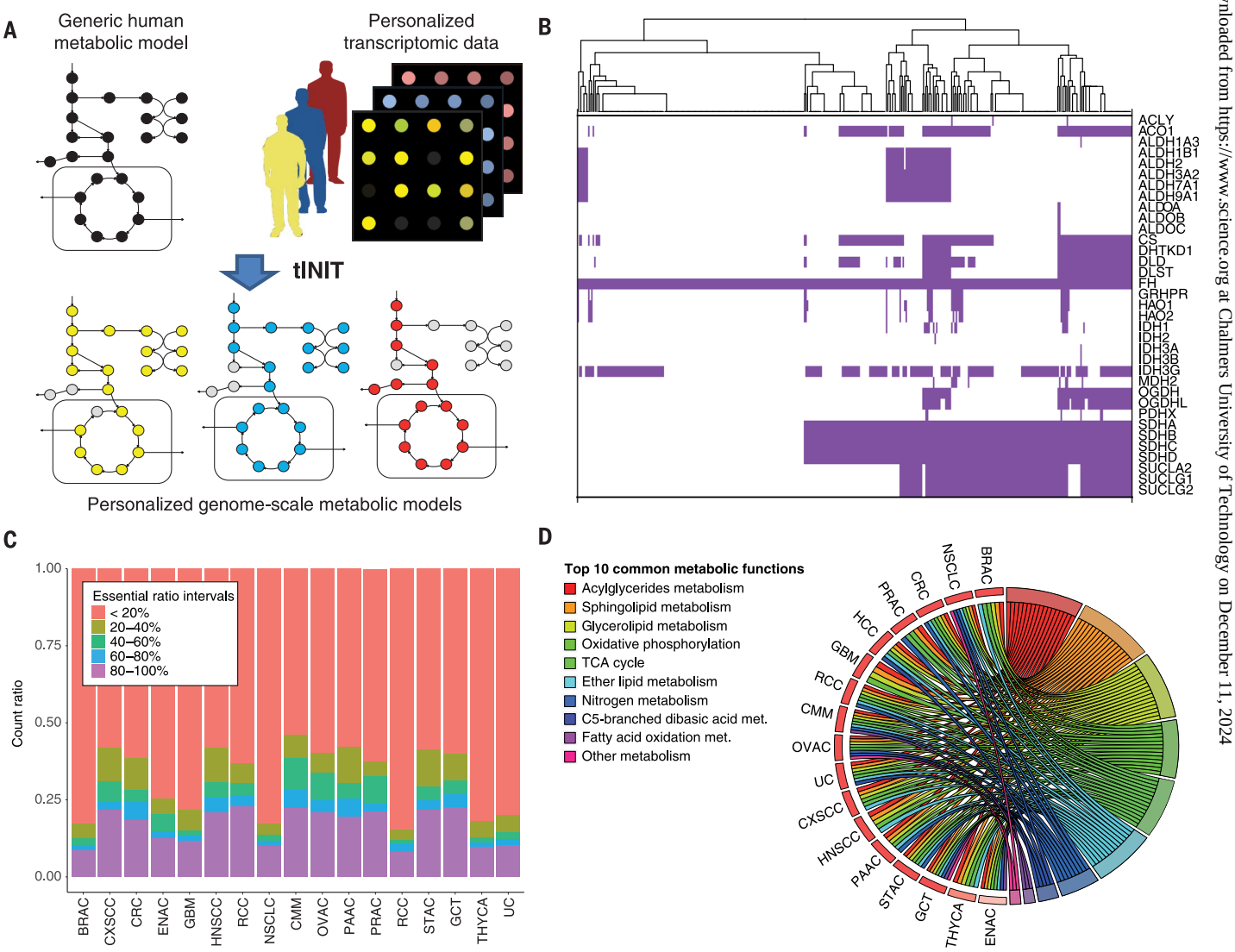


Fig. 6. Genome-scale metabolic models (GSMMs) of cancers. (**A**) Concept of personalized GSMMs, which are comprehensive compilations of all the metabolic reactions within a particular cell, tissue, organ, or organism. By mapping the transcriptomic data from cancer patients, personalized GSMMs could be reconstructed for investigation of the specific metabolic viabilities for each individual. (**B**) Heat map showing the essential enzymes in the TCA cycle for all glioma patients to exemplify the heterogeneity within the same cancer patient group. Only enzymes that were key in at least one patient are shown. (**C**) Bar plot showing the fraction of genes that were common in key genes in different proportions of patients for 17 Human Pathology Atlas cancers. (**D**) Circos plot showing the top 10 common metabolic pathways that were over-represented by key genes in 17 Human Pathology Atlas cancers. Abbreviated names are provided in Fig. 1A and table S17.

When we investigated the associated biological functions, a vast majority of these genes were associated with central metabolic functions that are essential for normal tissues (Fig. 6D and table S16), and the corresponding proteins are thus not suitable as targets for drug development. Therefore, we performed toxicity tests using the models generated for healthy tissues and observed that, in many cases, the potential inhibition of 76 to 81% of these targets could be predicted to have severe side effects, because the target is essential in at least some normal tissues. Moreover, we also predicted that 32 gene targets that are mainly involved in nucleotide metabolism were predicted to be nontoxic in healthy tissues (fig. S12) but key in more than 80% of the tumor of the patient, regardless of the cancer type. These genes may therefore hold promise as potential targets for cancer treatment. In general, gene targets with less toxicity in normal tissue were key for tumor growth in fewer than 20% of cancer patients. Our analysis thus demonstrates the large heterogeneities in different cancer patients from a metabolic perspective and shows a path to individualized treatment of patients based on metabolic modeling, thereby highlighting the importance of systems-level analysis for precision cancer treatment.

Examination of genes in lung cancer

Further validation of prognostic genes identified through analyses of TCGA data was performed using an independent cohort of lung cancer

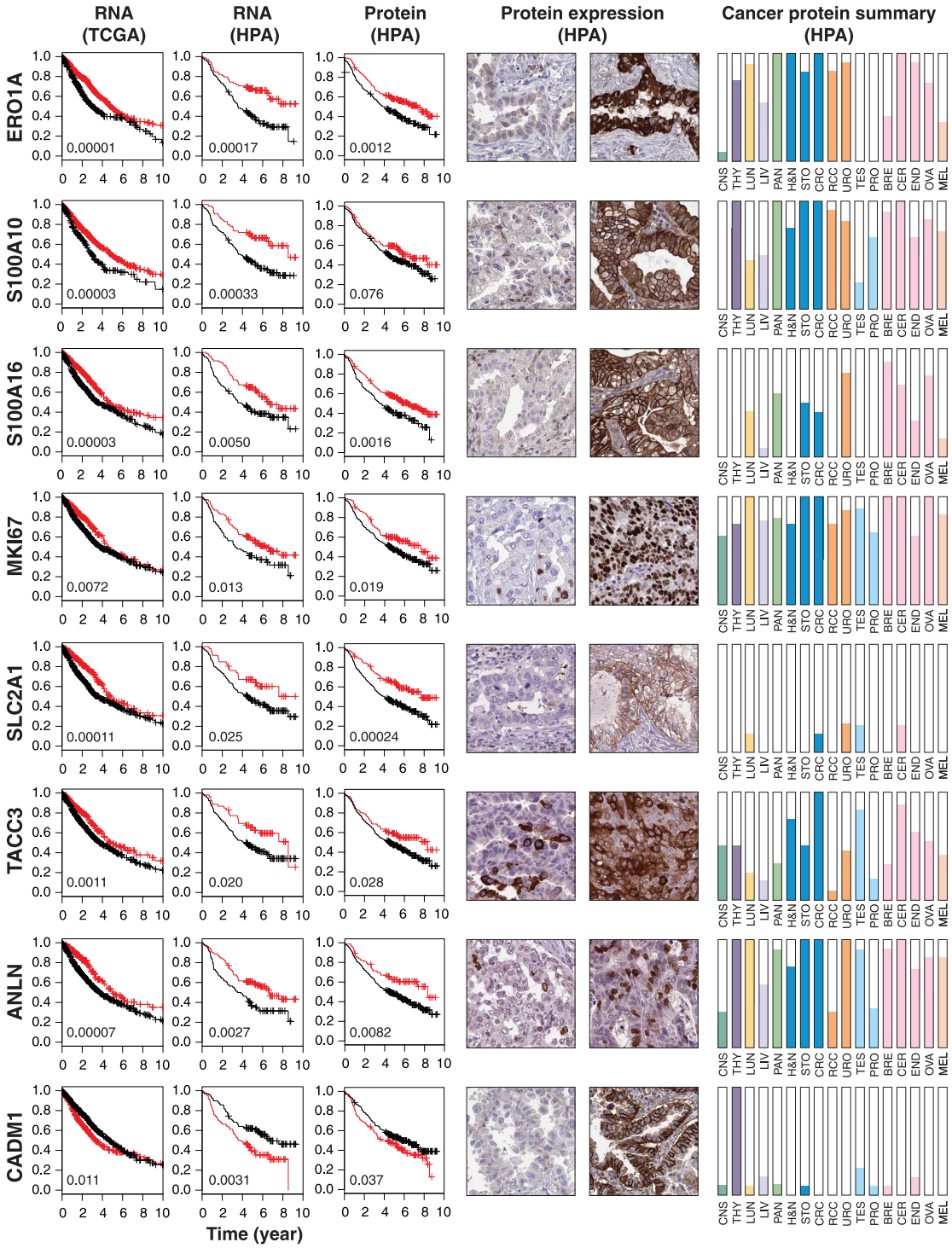


Fig. 7. Validation of selected genes with a prognostic effect in lung cancer. Kaplan-Meier plots for RNA level separation from the TCGA cohort, RNA level separation from the HPA cohort, and protein-level separation are shown in the first, second, and third columns, respectively. The log-rank *P* values are shown in the lower left corner of each Kaplan-Meier plot. IHC stained tissues representing high and low protein expression are shown in the fourth and fifth columns, respectively. The protein expression levels across 17 cancer types analyzed by IHC in the Human Pathology Atlas are shown at the right.

(NSCLC) patients (n = 357). We used available RNA-seq data from 199 individual tumors (30) and paraffin-embedded tumor tissue material in a tissue microarray (TMA) format from 357 patients (31). On the basis of transcriptomic data, the 100 most significant lung cancer prognostic genes identified in the TCGA analysis showed a high degree of overlap with prognostic genes in the independent NSCLC cohort (74% with P < 0.1, 45% with P < 0.01). In addition, the panel for lung cancer shown in Fig. 2A was also validated in this independent cohort (fig. S13).

To further investigate whether prognostic genes identified through genome-wide transcriptomics analyses could be verified at the protein level, we performed antibody-based IHC analyses of TMAs with tumor tissue (n = 357) for eight selected targets (Fig. 7). The IHC-based analysis confirmed that the corresponding protein expression pattern was also significantly associated with prognosis, and this was also supported by the RNA-seq data in the independent NSCLC cohort. Examples (Fig. 7) include the endoplasmic reticulum oxidoreductase α protein ERO1A (32) and two members of the S100 family (S100A10 and S100A16). The latter two proteins have been suggested as prognostic markers at the protein level in NSCLC adenocarcinoma (33, 34). We could confirm the prognostic association of both S100A10 and S100A16 in the NSCLC cohort containing both adenocarcinomas and squamous cell carcinomas. The proliferation marker MKI67 has been studied in a number of cancer types; however, its clinical application has been debated (35), and MKI67 has not been included in routine NSCLC diagnostics (36). In the present investigation, MKI67 was associated with an unfavorable prognosis in the TCGA data set, which was also confirmed at both the RNA and protein level in the independent NSCLC cohort. SLC2A1 (solute carrier family 2 member 1), also known as GLUT1, is a downstream gene of the hypoxic marker HIF1A and plays a role in glucose transport. TACC3 (transforming acidic coiled coil–containing protein 3) is involved in controlling normal cell growth and differentiation. Overexpression of SLC2A1 and TACC3 was previously associated with a poor prognosis in lung cancer (37, 38), and here we found that expression level associates with clinical outcome in lung cancer. Anillin (ANLN), an actin-binding protein required for cytokinesis, plays an important role in cell division and has been suggested as a prognostic marker in breast cancer (39) and lung cancer (40). Here, our TCGA analysis show prognostic value in lung, renal, pancreatic, and liver cancers, and the analysis of the independent lung cohort implies that this may be a favorable prognostic gene for clinical outcome.

Examination of genes in colon cancer

We investigated a large, independent, prospectively collected population-based cohort of colorectal cancer patients available in TMA format to assess possible prognostic protein signatures. In this cohort, mRNA expression data (RNA-seq) were also available for a smaller subset of the patients (n = 60). Six targets with prognostic significance in colorectal cancer based on TCGA data were selected for IHC staining on the TMAs. All six genes were verified as related to prognosis at both the RNA level (n = 60) and protein level (n = 745) (fig. S14).

The Human Pathology Atlas

As part of this publication, we launch a new open-access resource named the Human Pathology Atlas as part of the Human Protein Atlas (www.proteinatlas.org/pathology), presenting the Kaplan-Meier survival plots for all protein-coding genes in 17 different tumor types. A survival plot of the patient cohort, with the respective cancer and gene divided into two equal groups (median), is presented on the basis of RNA levels. More than 900,000 survival plots (as exemplified by Fig. 2C) are presented in the new pathology resource to allow investigators to explore the clinical significance of patient survival related to specific genes in specific cancers, together with the associated transcriptomic, proteomic, and clinical information. A total of 13,088 Kaplan-Meier plots with high significance (P < 0.001) are highlighted, and the data are presented in a gene-centric manner for all human protein-coding genes across the analyzed cancer types. Each prognostic gene for a given cancer type is shown, including the Kaplan-Meier plots (Fig. 2A), together with the underlying data for the selection of suitable FPKM cutoffs for patient stratification (Fig. 2B) and the individual survival data for all patients (Fig. 2B). In addition, IHC analysis using a TMA-based analysis of the corresponding proteins in patients with the respective cancer types is presented for a majority of the protein-coding genes. More than 5 million IHC-based cancer tissue images are included in the atlas, showing protein expression patterns for individual tumors of each cancer type. All IHC images have been manually annotated by certified pathologists. Thus, the resource allows researchers to explore the possible prognostic value of all human protein-coding genes related to expression levels in different forms of human cancer.

Discussion

Our results demonstrate the power of large systematic "big data" efforts that make use of publicly available resources, such as the TCGA database used herein. The compiled data show that a large fraction of human protein-coding genes are differentially expressed in cancer and that this differential expression in many cases has an impact on patient survival. Prognostic genes appear to be restricted to only a few cancer types, and no genes were informative across a large set of cancer patients. A general pattern emerged, with unfavorable genes showing an up-regulation associated with mitosis and cell growth, whereas the down-regulation of genes associated with cellular differentiation was associated with shorter patient survival. However, it is important to point out that for a given prognostic gene, we observe a huge variation in terms of clinical outcome for the corresponding patient, implying the need for further sophisticated studies to better comprehend the concept of prognostic genes.

The prognostic genes we identified should be validated in independent patient cohorts, as exemplified by the validation using antibody-based TMAs of a selection of the genes identified in lung cancer. The clinical metadata in the TCGA resource did not include therapeutic regimens for the patients, nor whether death was related to the diagnosed cancer. In addition, the different sample and effect sizes for different cancers would affect the number of prognostic genes obtained by survival analysis and log-rank test. Moreover, the purity of the tumor samples should also affect the survival analysis, as previously reported (41). Hence, there is a need for follow-up validation studies in more controlled independent cancer cohorts before a potential prognostic gene can be confirmed. An important quest for the near future is to identify which prognostic genes are functionally important ("drivers") with functional consequences that are required for carcinogenesis and tumor progression, and which of the apparent prognostic genes are merely coexpressed with these "driver" genes.

We generated cancer-specific coexpression networks to study the functional relationship between the prognostic genes and genes associated with Hallmarks of Cancer. This network-dependent analysis enabled the identification of genes with a key role in the survival of patients. The personalized genome-scale GSMMs allowed us to identify genes that were critical for tumor growth by demonstrating a huge heterogeneity among patients from a metabolic perspective, highlighting the need for precise and personalized medicine for cancer treatment. In this context, the new Human Pathology Atlas is a useful standalone resource for cancer precision medicine. With its more than 900,000 Kaplan-Meier plots, this resource enables insights concerning the specific involvement of genes in clinical outcome for all the major cancers, paving the way for further in-depth studies incorporating systems-level analyses of cancer. All data presented herein are available in an interactive open-access database (www.proteinatlas.org/pathology) to allow for genome-wide exploration of the impact of individual proteins on clinical outcome in major human cancer types.

REFERENCES AND NOTES

1. D. J. Brennan, D. P. O'Connor, E. Rexhepaj, F. Ponten, W. M. Gallagher, Antibody-based proteomics: Fast-tracking molecular diagnostics in oncology. *Nat. Rev. Cancer* **10**, 605–617 (2010). doi: 10.1038/nrc2902; pmid: 20720569
2. E. Björnson *et al.*, Stratification of hepatocellular carcinoma patients based on acetate utilization. *Cell Rep.* **13**, 2014–2026 (2015). doi: 10.1016/j.celrep.2015.10.045; pmid: 26655911
3. A. Mardinoglu, J. Nielsen, New paradigms for metabolic modeling of human cells. *Curr. Opin. Biotechnol.* **34**, 91–97 (2015). doi: 10.1016/j.copbio.2014.12.013; pmid: 25559199
4. S. Lee, A. Mardinoglu, C. Zhang, D. Lee, J. Nielsen, Dysregulated signaling hubs of liver lipid metabolism reveal hepatocellular carcinoma pathogenesis. *Nucleic Acids Res.* **44**, 5529–5539 (2016). doi: 10.1093/nar/gkw462; pmid: 27216817
5. J. N. Weinstein *et al.*, The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013). doi: 10.1038/ng.2764; pmid: 24071849

6. M. Uhlén et al., Tissue-based map of the human proteome. Science 347, 1260419 (2015). doi: 10.1126/science.1260419; pmid: 25613900
7. J. Lonsdale et al., The Genotype-Tissue Expression (GTEx) project. Nat. Genet. 45, 580–585 (2013). doi: 10.1038/ng.2653; pmid: 23715323
8. L. Collado-Torres et al., Reproducible RNA-seq analysis using recount2. Nat. Biotechnol. 35, 319–321 (2017). doi: 10.1038/nbt.3838; pmid: 28398307
9. L. Peng 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 (2015). doi: 10.1038/srep13413; pmid: 26292924
10. F. Edfors et al., Gene-specific correlation of RNA and protein levels in human cells and tissues. Mol. Syst. Biol. 12, 883 (2016). doi: 10.15252/msb.20167144; pmid: 27951527
11. D. Hanahan, R. A. Weinberg, Hallmarks of cancer: The next generation. Cell 144, 646–674 (2011). doi: 10.1016/j.cell.2011.02.013; pmid: 21376230
12. C. Kandoth et al., Mutational landscape and significance across 12 major cancer types. Nature 502, 333–339 (2013). doi: 10.1038/nature12634; pmid: 24132290
13. T. Hothorn, B. Lausen, On the exact distribution of maximally selected rank statistics. Comput. Stat. Data Anal. 43, 121–137 (2003). doi: 10.1016/S0167-9473(02)00225-6
14. B. Hjelm et al., High nuclear RBM3 expression is associated with an improved prognosis in colorectal cancer. Proteomics Clin. Appl. 5, 624–635 (2011). doi: 10.1002/prca.201100020; pmid: 21956899
15. C. J. Creighton et al., Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature 499, 43–49 (2013). doi: 10.1038/nature12222; pmid: 23792563
16. A. Subramanian et al., Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U.S.A. 102, 15545–15550 (2005). doi: 10.1073/pnas.0506580102; pmid: 16199517
17. H. A. Edmondson, P. E. Steiner, Primary carcinoma of the liver: A study of 100 cases among 48,900 necropsies. Cancer 7, 462–503 (1954). doi: 10.1002/1097-0142(195405)7:3<462::AID-CNCR2820070308>3.0.CO;2-E; pmid: 13160935
18. T. M. Pawlik et al., Preoperative assessment of hepatocellular carcinoma tumor grade using needle biopsy: Implications for transplant eligibility. Ann. Surg. 245, 435–442 (2007). doi: 10.1097/01.sla.0000250420.73854.ad; pmid: 17435551
19. A. J. Simpson, O. L. Caballero, A. Jungbluth, Y. T. Chen, L. J. Old, Cancer/testis antigens, gametogenesis and cancer. Nat. Rev. Cancer 5, 615–625 (2005). doi: 10.1038/nrc1669; pmid: 16034368
20. M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, K. Morishima, KEGG: New perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 45, D353–D361 (2017). doi: 10.1093/nar/gkw1092; pmid: 27899662
21. T. I. Zack et al., Pan-cancer patterns of somatic copy number alteration. Nat. Genet. 45, 1134–1140 (2013). doi: 10.1038/ng.2760; pmid: 24071852
22. N. N. Pavlova, C. B. Thompson, The emerging hallmarks of cancer metabolism. Cell Metab. 23, 27–47 (2016). doi: 10.1016/j.cmet.2015.12.006; pmid: 26771115
23. M. G. Vander Heiden, R. J. DeBerardinis, Understanding the intersections between metabolism and cancer biology. Cell 168, 657–669 (2017). doi: 10.1016/j.cell.2016.12.039; pmid: 28187287
24. P. Ghaffari, A. Mardinoglu, J. Nielsen, Cancer metabolism: A modeling perspective. Front. Physiol. 6, 382 (2015). doi: 10.3389/fphys.2015.00382; pmid: 26733270
25. A. Mardinoglu et al., Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. Nat. Commun. 5, 3083 (2014). doi: 10.1038/ncomms4083; pmid: 24419221
26. R. Agren et al., Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. Mol. Syst. Biol. 10, 721 (2014). doi: 10.1002/msb.145122; pmid: 24646661
27. A. Mardinoglu et al., Personal model-assisted identification of NAD(+) and glutathione metabolism as intervention target in NAFLD. Mol. Syst. Biol. 13, 916 (2017). doi: 10.15252/msb.20167422; pmid: 28254760
28. L. Jerby-Arnon et al., Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality. Cell 158, 1199–1209 (2014). doi: 10.1016/j.cell.2014.07.027; pmid: 25171417
29. C. Zhang, Q. Hua, Applications of genome-scale metabolic models in biotechnology and systems medicine. Front. Physiol. 6, 413 (2016). doi: 10.3389/fphys.2015.00413; pmid: 26779040
30. D. Djureinovic et al., Profiling cancer testis antigens in non-small-cell lung cancer. Jci Insight 1, e86837 (2016). doi: 10.1172/jci.insight.86837; pmid: 27699219
31. P. Micke et al., The impact of the Fourth Edition of the WHO Classification of Lung Tumours on histological classification of resected pulmonary NSCCs. J. Thorac. Oncol. 11, 862–872 (2016). doi: 10.1016/j.jtho.2016.01.020; pmid: 26872818
32. T. Tanaka et al., Cancer-associated oxidoreductase ERO1-α drives the production of VEGF via oxidative protein folding and regulating the mRNA level. Br. J. Cancer 114, 1227–1234 (2016). doi: 10.1038/bjc.2016.105; pmid: 27100727
33. K. Katono et al., Clinicopathological significance of S100A10 expression in lung adenocarcinomas. Asian Pac. J. Cancer Prev. 17, 289–294 (2016). doi: 10.7314/APJCP.2016.17.1.289; pmid: 26838226
34. K. Saito et al., S100A16 is a prognostic marker for lung adenocarcinomas. Asian Pac. J. Cancer Prev. 16, 7039–7044 (2015). doi: 10.7314/APJCP.2015.16.16.7039; pmid: 26514487
35. F. Penault-Llorca, N. Radosevic-Robin, Ki67 assessment in breast cancer: An update. Pathology 49, 166–171 (2017). doi: 10.1016/j.pathol.2016.11.006; pmid: 28065411
36. J. N. Jakobsen, J. B. Sørensen, Clinical impact of Ki-67 labeling index in non-small cell lung cancer. Lung Cancer 79, 1–7 (2013). doi: 10.1016/j.lungcan.2012.10.008; pmid: 23137549
37. M. Younes, R. W. Brown, M. Stephenson, M. Gondo, P. T. Cagle, Overexpression of Glut1 and Glut3 in stage I nonsmall cell lung carcinoma is associated with poor survival. Cancer 80, 1046–1051 (1997). doi: 10.1002/(SICI)1097-0142(19970915)80:6<1046::AID-CNCR6>3.0.CO;2-7; pmid: 9305704
38. C. K. Jung et al., Expression of transforming acidic coiled-coil containing protein 3 is a novel independent prognostic marker in non-small cell lung cancer. Pathol. Int. 56, 503–509 (2006). doi: 10.1111/j.1440-1827.2006.01998.x; pmid: 16930330
39. K. Magnusson et al., ANLN is a prognostic biomarker independent of Ki-67 and essential for cell cycle progression in primary breast cancer. BMC Cancer 16, 904 (2016). doi: 10.1186/s12885-016-2923-8; pmid: 27863473
40. C. Suzuki et al., ANLN plays a critical role in human lung carcinogenesis through the activation of RHOA and by involvement in the phosphoinositide 3-kinase/AKT pathway. Cancer Res. 65, 11314–11325 (2005). doi: 10.1158/0008-5472.CAN-05-1507; pmid: 16357138
41. D. Aran, M. Sirota, A. J. Butte, Systematic pan-cancer analysis of tumour purity. Nat. Commun. 6, 8971 (2015). doi: 10.1038/ncomms9971; pmid: 26634437

ACKNOWLEDGMENTS

The data reported in this manuscript are tabulated in the main paper and the supplementary materials. We thank the entire staff of the Human Protein Atlas program and the Science for Life Laboratory, the National Genomics Infrastructure, and Swedish National Infrastructure for Computing at UPPMAX and C3SE for providing assistance in massive parallel sequencing and computational infrastructure. Supported by the Swedish Cancer Foundation (F.P.), Science for Life Laboratory infrastructure funding (C.L., E.L. and P.N.), the Erling Persson Foundation (M.U.), Elixir (EU infrastructure funding) (K.F. and M.U.), the Swedish Cancer Foundation and the Uppsala Lions Cancer Foundation (P.M.), the Knut and Alice Wallenberg Foundation, and the strategic research area "U-CAN" through Swedish Research Council grant CancerUU. We thank the Uppsala Biobank and Clinical Department of Pathology at the Uppsala University Hospital for providing specimens. We also thank The Cancer Genome Atlas for providing access to data. M.U. and F.P. are inventors on patents held by Atlas Antibodies that cover the analysis of RBM3 (U.S. patent 8728739) and PODXL (U.S. patent 8999656) as a prognostic marker in several cancers. The authors declare that they have no other conflict of interest. An interactive pathology atlas with gene-specific data for all human genes (including more than 900,000 Kaplan-Meier plots) is available at www.proteinatlas.org/pathology.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/357/6352/eaan2507/suppl/DC1
Materials and Methods
Figs. S1 to S14
Tables S1 to S21
References (*42–51*)

22 March 2017; resubmitted 2 June 2017
Accepted 14 July 2017
10.1126/science.aan2507