bioRxiv preprint doi: https://doi.org/10.1101/030924; this version posted November 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A pan-cancer analysis of prognostic genes

Jordan Anaya[1]

[1]omnesres.com, email: omnesresnetwork@gmail.com, twitter: @omnesresnetwork

Corresponding Author:

Jordan Anaya

Charlottesville, VA, US

email address: omnesresnetwork@gmail.com

bioRxiv preprint doi: https://doi.org/10.1101/030924; this version posted November 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

Numerous studies have identified prognostic genes in individual cancers, but a thorough pan-cancer analysis has not been performed. In addition, previous studies have mostly used microarray data instead of RNA-SEQ, and have not published comprehensive lists of associations with survival. Using recently available RNA-SEQ and clinical data from the The Cancer Genome Atlas for 6,495 patients, I have investigated every annotated and expressed gene's association with survival across 16 cancer types. The most statistically significant harmful and protective genes were not shared across cancers, but were enriched in distinct gene sets which were shared across certain groups of cancers. These groups of cancers were independently recapitulated by both unsupervised clustering of Cox coefficients (a measure of association with survival) for individual genes, and for gene programs. This analysis has revealed unappreciated commonalities among cancers which may provide insights into cancer pathogenesis and rationales for co-opting treatments between cancers.

Main article text

Introduction

Led by The Cancer Genome Atlas, unprecedented efforts have been made to understand the molecular basis of cancer (http://cancergenome.nih.gov). Using standardized procedures, the TCGA Research Network has used whole genome sequencing, exome sequencing, RNA-SEQ, small RNA-SEQ, bisulfite-SEQ, and reverse phase arrays to identify the pathways commonly altered in different cancers [1-12]. As a result, we now know the most commonly mutated genes in dozens of cancers and can use this information to give patients targeted therapeutics.

Whereas well established statistical techniques exist for identifying mutations which are drivers instead of simply passengers (mut-drivers), identifying copy number aberrations, methylation changes, or non-coding mutations that alter expression of a gene and result in a growth advantage (epi-drivers) are more difficult to identify and represent a "dark matter" of cancer [13]. Although it currently is challenging to identify epi-drivers which lead to development of a cancer (tumorigenesis), by correlating these changes to survival it is possible to detect their role in disease progression (pathogenesis), which is one of main goals of cancer research.

Of the possible genomic measures that can be correlated with survival, gene expression has been shown to be the strongest predictor of survival [14], which is intuitive given that gene levels together with protein levels and posttranslational modifications are the final readout of the different possible alterations in a cell and are the final effectors of phenotype. To date many attempts have been made to identify genes whose expression is associated with survival to either identify markers that can predict patient survival or to identify mechanisms of pathogenesis [15-17]. One of the success stories of this approach is the identification of HER2 in breast cancer patients and the development of herceptin [18]. This story also highlights the complications treatment

bioRxiv preprint doi: https://doi.org/10.1101/030924; this version posted November 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

regimens can have on interpreting survival data. Whereas HER2 overexpression used to predict poor survival for breast cancer patients, because of the progress of personalized medicine these patients now do well and HER2 would not show up as a prognostic marker in a data set with HER2 positive patients on herceptin. While treatments may introduce a confounding variable in understanding a disease, the ultimate goal of cancer studies is to improve patient outcome, and adding treatment to the equation adds more information and provides an opportunity to study genes in the context of the current standard of care.

The vast majority of studies to identify prognostic genes have focused on a single disease and have utilized microarrays instead of RNA-SEQ. In addition, these studies often only publish a small set of genes that together most significantly stratify patients. Even the TCGA Research Network publications do not provide lists of genes associated with survival. cBioPortal does allow users to make Kaplan Meier plots for most of the cancers which contain survival information, but users still have to input one gene at a time, leaving one to wonder where researchers should go to find the genes which are most highly correlated with survival for their disease of interest.

Through the TCGA Network, RNA-SEQ has only very recently become available for thousands of human cancer samples. RNA-SEQ has multiple advantages over microarray data, including having a higher dynamic range, no probe affinity effects, ability to identify novel transcripts, and lower and consistently falling cost. I took advantage of the availability of this data to 1) investigate the ability of RNA-SEQ to associate expression with clinical outcome in a range of cancers, 2) perform the largest analysis of prognostic genes to date, and 3) provide every gene's correlation with survival for hypothesis testing and further investigations by the scientific community. In addition, attempts are now being made to identify commonalities between cancers with the hope that this type of analysis may be able to identify treatments that can be co-opted for a molecularly similar cancer. Given that only survival correlations integrate treatment with the genomic data, prognostic genes represent an exclusive window for understanding how different cancers in the context of their individualized treatments relate to one another. By performing the largest pan-cancer analysis of prognostic genes, I identified groupings of cancers which were robustly replicated through several methods. This study serves as a starting point for better understanding how survival data can be used to understand the commonalities and differences of cancers.

Materials and methods

Code and files

All of the Python and R code used to generate the figures and tables in this study, including intermediate and final files, tables, and figures, is available at https://github.com/OmnesRes/pan_cancer. All scripts were run on a HP dv7t laptop with an i7-3820QM processor and 16GB of RAM running Windows 7, Python 2.7.5, and R 3.0.1.

bioRxiv preprint doi: https://doi.org/10.1101/030924; this version posted November 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Construction of multivariate Cox models

RNA-SEQ and clinical data were downloaded from the TCGA data portal, https://tcga-data.nci.nih.gov/tcga/, in February 2015. For each cancer, survival information was parsed from the "clinical_follow_up" files and "clinical_patient" file, and for each patient the most recent follow up information found in the multiple files was kept. Sex, age, and histological grade data were extracted from the "clinical_patient" file. For each cancer, only patients that had a follow up time greater than 0 days and had complete clinical information were included in the model. TCGA has used two different methods of reporting expression values, RSEM and RPKM. RPKM is simply the reads per kilobase per million mapped reads, while RSEM is a normalized value outputted by the RSEM software [19]. For each cancer, only genes which had a median RSEM value greater than 1 (for RNASeqV2), or median RPKM value greater than .1 (for RNASeq), and had 0 expression in less than one fourth of patients were included in the analysis. RNASeq uses a different gene annotation file from RNASeqV2, and because RNASeqV2 represents the most recent analysis, for RNASeq analyses only those genes present in the RNASeqV2 gene annotation file were included. Multivariate Cox models were run with the coxph function from the R survival library, and the equation for each model is shown in Table S1. Grade information was included in the model by separate terms, which were either 1 or 0, and model input gene expression values were inverse normal transformed. If a patient had replicates for their primary tumor, those expression values were averaged prior to inverse normal transformation. The scripts for performing Cox regression for each cancer are named "cox_regression.py".

Gene set analysis

For each cancer, the 250 most significant protective genes and 250 harmful genes were inputted separately into MSigDB with the "positional genes sets", "chemical and genetic perturbations", "canonical pathways", "KEGG gene sets", "microRNA targets", "transcription factor targets", "cancer modules", "GO biological process", and "oncogenic signatures" sets selected: http://www.broadinstitute.org/gsea/msigdb. The FDR q-value threshold was set at .05 and the top 100 enriched gene sets were saved, except for the 250 protective genes in BLCA, which only contained 27 overlaps below .05.

Normalization of Cox coefficients

In order to compare the Cox coefficients between cancers I robustly scaled the negative and positive coefficients, x, to their 5th and 95th percentile values, respectively, using the following sigmoidal normalization function:

$$z_{\pm} = \frac{2}{1 + e^{-\frac{2x}{|u_{\pm}|}}} - 1$$

bioRxiv preprint doi: https://doi.org/10.1101/030924; this version posted November 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

where u_- and u_+ are the 5th and 95th percentile values of the negative and positive Cox coefficients, respectively. The implementation of this code is present in the files named "normalizing_coeffs.py", and all the normalized coefficients are listed in Table S1.

Construction of gene programs

Gene programs from Table S4 of [20] were used. In general a nonredundant set of genes from gene sets which had a Pearson correlation of at least .9 [20] was generated for each program. The exact gene sets used are listed in Table S3. Lists of genes for the gene sets were obtained from http://www.broadinstitute.org/gsea/msigdb and [21].

Results

Cancers vary in number of prognostic genes

In order to perform the most comprehensive cancer analysis possible, I selected TCGA cancers that had sufficient numbers of patients with RNA-SEQ data and mature clinical follow up information, and did not contain any publication restrictions. This resulted in me studying a total of 16 cancers, 10 of which were present in the original pan-cancer initiative [22]: acute myeloid leukemia (LAML), bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), and ovarian serous cystadenocarcinoma (OV), and 6 cancers which have been the focus of limited individual or pan-cancer studies: cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), brain lower grade glioma (LGG), kidney renal papillary cell carcinoma (KIRP), liver hepatocellular carcinoma (LIHC), skin cutaneous melanoma (SKCM), and stomach adenocarcinoma (STAD).

I was interested in the effect a gene has on prognosis independent of factors such as tumor grade and age of a patient. To achieve this I used a multivariate Cox proportions hazards model [23], which is a standard regression method for studying survival data [24-27]. For every cancer, a model was generated separately for each gene, with the number of covariates depending on the cancer. Unlike microarray data, RNA-SEQ data has extreme values which may affect regression. To account for this I inverse normal transformed the expression values of each gene, which has been shown to increase the sensitivity and specificity for multivariate regression with RNA-SEQ data [28]. Age and sex are also included in every model, and when a cancer contained strong histological grade information, grade was also included. If a patient was missing any of this information they were excluded from the analysis, and only primary tumors were considered, with the exception of SKCM, where metastatic tumors make up a large proportion of the patients.

bioRxiv preprint doi: https://doi.org/10.1101/030924; this version posted November 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A Cox model provides a p-value for each term in the model, indicating the significance of its association with the clinical outcome, and I recorded the p-values for every gene analyzed for the 16 different cancers. As can be seen in Table 1, there is a wide distribution among cancers in the number of genes that reached a Benjamini-Hochberg False Discovery Rate (FDR) adjusted p-value of less than or equal to .05. This can also be seen by looking at a distribution of the raw p-values for the different cancers (Fig. 1a and Fig. S1). This has important implications for understanding the significance of a gene being associated with survival in a specific cancer. For example, selecting a gene at random a researcher studying LGG has a 50% chance of being able to claim the gene is associated with survival, while a researcher studying STAD only has an 8% chance (using raw p-values).

Two factors that are known to be associated with power of a Cox model are sample size and number of events (deaths); however looking through Table 1 it is difficult to find a pattern that can explain why certain cancers have more significant genes than other cancers. For example, BRCA has around twice the number of patients of any other cancer, but only has 30 genes that meet a FDR cutoff, which can be considered expressionally prognostic genes (EPGs). In contrast, KIRP has a fourth the number of patients of BRCA but has 2,415 EPGs. In addition, LUAD and LUSC have similar numbers of patients, median survivals, and events, yet have a large difference in EPGs. Interestingly, it has been shown that the number of prognostic genes for a cancer can be significantly different depending on whether microarray data or RNA-SEQ data is used [29]. It is possible that the different numbers of EPGs between cancers are due to intra-disease heterogeneity and/or treatment differences that are not accounted for in the Cox model and are acting as confounding variables, or differences in the amount of transcriptional dysregulation between cancers.

Protective and harmful genes display opposite expression patterns

The Cox model also provides a coefficient for each term, which is related to its contribution to the hazard ratio. A positive coefficient indicates that the gene increases the hazard ratio, i.e. high expression of the gene correlates with earlier patient death, while a negative coefficient indicates that expression of the gene is protective. Using the cancer with the highest number of EPGs (LGG), I clustered patients with the 100 most significant genes which were harmful and the 100 most significant genes which were protective, and this revealed two broad clusters of patients: (1) those with high expression of harmful genes and low expression of protective genes, and (2) those with high expression of protective genes and low expression of harmful genes (Fig. 1b). Not surprisingly, performing a Kaplan Meier analysis with these two groups revealed that cluster 2 has a much higher survival than cluster 1 (Fig. 1c). This result has important implications for trying to find gene sets which can most accurately predict patient survival. The similar expression patterns indicate that there are numerous combinations of genes that would only differ slightly in their ability to predict survival, making the identification of a 'best' set of genes somewhat meaningless. In addition, given that each gene individually had a p-value less than or

bioRxiv preprint doi: https://doi.org/10.1101/030924; this version posted November 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

equal to 1.4E-8, it is unlikely these patterns are due to chance but rather are the result of some underlying gene regulation and these genes may be members of known pathways.

Unlike LGG, some of the other cancers in my analysis yielded a much lower number of EPGs. While it might be tempting to disregard the results in these cancers, I decided to check if I could observe patterns of expression in the most significant good and bad genes like I had observed for cancers with a high number of EPGs. Clustering of the patients of STAD, which has one of the lowest numbers of EPGs, with the 100 most significant harmful genes, and the 100 most significant protective genes, again divided patients into two broad clusters. Surprisingly, a Kaplan Meier analysis on these two groups showed a very significant difference with a p-value of 2.73E-6 (Fig. 1c). This indicates that despite the fact that none of the genes in STAD meet a 5% FDR cutoff, they still contain important biological information. As a result, I decided to include all cancers in further analyses regardless of their numbers of EPGs.

Cancers do not share prognostic genes, but do share gene sets

I was interested to see if the most significantly prognostic genes were shared across cancers. However, looking at the overlap of the 100 most significant genes across the 16 cancers revealed that there is very little overlap among the 16 cancers, consistent with previous results obtained from an analysis of four cancers [29] (Fig. 2a). Given the apparent co-regulation of the most significant genes in each cancer, I reasoned that although individual genes were not shared, maybe the genes were a part of gene sets which were shared between cancers. In addition, given that the harmful genes had an opposite pattern of expression from the protective genes, I hypothesized that they are regulated differently and would be enriched in different gene sets. To investigate this I took the 250 most significant harmful genes and 250 most significant protective genes in each cancer, and separately found the 100 most enriched gene sets through MSigDB [30]. Consistent with my idea that harmful and protective genes are regulated differently, there was very little overlap within a cancer between the 100 gene sets found with 250 harmful genes and the 100 gene sets found with 250 protective genes (Fig. 2b). In addition, the fact that even the protective and harmful gene sets from cancers with a low number of EPGs show almost no overlap reinforces the idea that prognostic genes in these cancers still contain biologically significant information.

Next I assessed the extent to which these protective and harmful gene sets overlapped between the different cancers. I investigated the overlaps separately for the 100 harmful gene sets and 100 protective gene sets (Fig. 2c, 2d). Overall there was more overlap between the harmful gene sets, and there were three cancers which clearly shared a high number of harmful gene sets, LUAD, LIHC, and KIRP. Investigating these overlaps further showed that the three cancers shared 58 gene sets, and LUAD and KIRP shared 85 gene sets (Fig. 2e). Looking at the overlaps of the protective gene sets, the largest overlap was between COAD and LUSC, and these cancers also shared gene sets with GBM (Fig. 2f).

bioRxiv preprint doi: https://doi.org/10.1101/030924; this version posted November 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

I next asked what are the most common harmful and protective gene sets across cancers. Table S2 shows frequency of every gene set, with gene sets that were shared between harmful and protective sets within a single cancer marked in bold as they may be nonspecific. As might be expected, the most common gene sets observed for harmful genes were associated with poor differentiation and metastasis. In contrast, the protective gene sets were enriched for apoptosis and good differentiation. Although when possible the grade of the tumor was included in the Cox model, and therefore should not be a confounding variable, it is possible that histological grade does not completely account for the differentiation of a tumor, indicating the importance of genomics for accurate profiling.

Cancers can be clustered by gene and gene program Cox coefficients

To date different cancers have been compared to each other through mRNA levels, miRNA levels, protein levels, networks, copy number alterations, DNA methylation, somatic mutations or some combination of these [20,31-35]. The Cox coefficients in my analysis contain a level of information not present in any of these data types, and consequently can potentially reveal similarities or differences between cancers that were not appreciated before. Therefore I sought to attempt to cluster cancers using Cox coefficients of genes instead of expression levels. Because the Cox models for the different cancers contain different numbers of covariates, and different strengths of gene expression correlation to survival, the range of values of the Cox coefficients vary between cancers. To correct for this, I normalized the coefficients for each cancer using a sigmoidal function which robustly scaled both negative and positive coefficients to their 95th percentile values (see methods). In addition, whereas every gene has an expression value, only significant prognostic genes have Cox coefficients appreciably above or below 0. Performing clustering with large numbers of nonsignificant genes which all have very similar values for every cancer will only add noise to the clustering. As a result, I limited my clustering to genes which had a FDR less than or equal to .05 in at least four of the sixteen cancers.

Hierarchical clustering of the 16 cancers was performed with the sigmoidal normalized Cox coefficients of this set of genes (Fig. 3). The clustering grouped LIHC, LUAD, and KIRP together, which were the same cancers that shared the highest number of harmful gene sets. In addition, GBM, COAD and LUSC clustered together, which were the cancers that had the highest number of protective gene sets overlap. The fact that two separate methods, using different sets of genes, were able to find these groupings of cancers gives me confidence that this finding is robust and biologically significant.

I next wanted to know if there were established pathways that distinguished my groupings of cancers from each other. Using a list of nonredundant gene programs that have been shown to distinguish cancers from one another on the basis of expression levels [20], I sought to distinguish cancers using Cox coefficients of pathways. For each pathway the average sigmoidal normalized Cox coefficient was calculated in each cancer. Because a Cox coefficient can be positive or negative, if a pathway has some genes which are protective and some genes which are harmful,

bioRxiv preprint doi: https://doi.org/10.1101/030924; this version posted November 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the average Cox score will be near zero. In addition, if a pathway only contains genes which are not prognostic, all of those Cox scores will be near zero and the pathway score will be near 0. The only way for a pathway to have a positive or negative score is for it to contain prognostic genes which are either consistently protective or consistently harmful.

Using the Cox scores for these 22 gene programs, I again performed hierarchical clustering (Fig. 4). I column scaled the values to highlight which gene programs are most important for each cancer. Overall the same groupings I had seen with gene sets and individual genes were recapitulated, with LUAD, KIRP, and LIHC again forming a cluster and COAD, LUSC, and GBM grouped together. In the LUAD/KIRP/LIHC group poor prognosis is associated with high proliferation rates and glycolysis, while good prognosis is associated with apoptosis and a dependency on oxidative phosphorylation. In contrast, for GBM/LUSC/COAD, proliferation is protective while EGF response predicts poor survival.

I also found cancer specific protective/harmful pathway enrichments that are consistent with known cancer biology. For example, in KIRC the highest intensity gene program is "fatty acid oxidation", and KIRC is a cancer that is known to depend on dysregulation of metabolism and is a classic example of the "Warburg effect" [36]. My results show that patients with high expression of genes utilizing oxygen survive longer, which underscores the importance of a metabolic shift in this cancer. As another example, EGFR is the most commonly mutated gene in GBM [3], and I observed that increased EGFR activity is associated with poorer outcomes. And BLCA and SKCM, which are known for being responsive to immunotherapy, both benefit from increased interferon response and an immune cell signature which is likely a proxy for immune cell infiltration.

Discussion

Cancer researchers are increasingly looking to focus on factors which have clinical significance, and many different resources now allow researchers to identify if a protein of interest has clinical implications, including OMIM, dbSNP, ClinVar, cBioPortal, FINDbase, and others [37-40]. Despite this, it currently is not possible to find comprehensive lists of genes which are associated with survival in different cancers. Using recently available RNA-SEQ and clinical data from the TCGA for 6,495 patients, I correlated every expressed annotated gene to survival in 16 different cancers, providing the scientific community with thousands of highly significant genes for further study.

There is an unexpectedly large variation between cancers in the number of statistically significant prognostic genes, which should be used to inform our evaluation of prognostic genes from different cancers. For example, a significant p-value for a gene from a cancer such as LGG or KIRC should not be surprising, given the thousands of genes that survive a stringent p-value cutoff in these tumors (Table 1, Fig. 1a, Fig. S1). In contrast, weaker p-values for predicting

bioRxiv preprint doi: https://doi.org/10.1101/030924; this version posted November 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

prognosis in cancers such as STAD or COAD are still biologically important although they have no genes that pass a stringent p-value threshold for biological significance (Table 1, Fig. 1a,c).

RNA-SEQ is a relatively new technology, and its ability to identify prognostic genes in many cancers has not been explored. Although the number of expressionally prognostic genes (EPGs) varied among cancers, regardless of the cancer I identified expression profiles which significantly separated patients into high risk and low risk groups. One of the main advantages of RNA-SEQ over microarrays is the ability to identify unannotated transcripts. In fact, recent studies have investigated the expressions of pseudogenes and long noncoding RNAs in large numbers of TCGA RNA-SEQ data sets [41,42]. It would be interesting to see if these transcripts show the same trends as protein coding genes across these cancers.

This comprehensive analysis of prognostic genes allowed me to explore the ability of the prognostic genes themselves, enriched gene sets, and Cox coefficients (a measure of strength of correlation to better or worse survival) to identify similarities and differences among cancers. I found that the most prognostically significant genes were not shared between cancers. However, I found that protective genes and harmful genes are enriched in very different gene sets, and there were large overlaps of these gene sets for LUAD, LIHC, and KIRP, and for COAD, LUSC, and GBM. I was able to recapitulate these findings by clustering with both Cox coefficients of individual genes, and average Cox coefficients of gene programs, suggesting that these findings are biologically significant and that I have identified a paradigm for incorporating genomic and clinical data to compare cancers.

Although it is important not to mistake a correlation for causation, my analysis suggests intriguing insights into the pathogenesis of different cancers. For example, currently EGFR inhibitors are recommended for LUAD patients with EGFR mutations, but EGFR mutations are rare in LUSC and patients with mutations do not respond well to tyrosine kinase inhibitors [43]. Despite this, response rates to EGFR inhibitors for LUSC studies are threefold higher than expected [43], suggesting that although EGFR itself may not be mutated, responders may still have a cancer which is dependent on EGFR signaling. This is consistent with my gene program analysis, where EGFR response was most strongly associated with poor survival in LUSC, and LUSC was consistently associated with GBM, which is a cancer known for EGFR dysreguation. This suggests that using a measure of EGFR activity other than mutational status could be used to find LUSC patients that would benefit from a tyrosine kinase inhibitor. In addition, my analysis may be used to suggest treatments for cancers which are not well studied. For example, KIRP does not have successful treatments and there is a current search for drugs which may be of benefit [44]. My analysis suggests that the pathogenesis of KIRP is very similar to LIHC and LUAD, indicating that treatments currently used for those cancers may be able to be co-opted for KIRP.

This analysis is among the first attempts at using clinical correlations to compare cancers. Although I utilized the most up to date information possible, well established statistical

bioRxiv preprint doi: https://doi.org/10.1101/030924; this version posted November 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

techniques, and obtained robust findings, there are many ways this type of analysis can be improved. For example, it is now being recognized that cancer is not a single disease, but rather a group of molecularly and clinically distinct diseases which share a tissue of origin. Through a combination of genomic measurements, the TCGA Research Network has divided individual cancers into four or five subtypes, for example GBM has been divided into proneural, neural, classical, and mesenchymal subtypes [3]. Currently, clear subtypes have not been found for all 16 of the cancers in this study, and for many cancers dividing the cancers into the subtypes would result in a loss of power due to the limited number of patients. However, as these classifications are refined and the number of patient samples continues to grow, a natural extension of this study would be to repeat it for individual subtypes, which would potentially decrease the heterogeneity of the data. In addition, treatment is one the largest confounding variables in survival analyses, but the TCGA pharmacological data is currently incomplete making it impossible to incorporate this information into the model. Despite these current limitations, this study has shown that incorporating clinical information into pan-cancer analyses is capable of yielding insights into cancer pathogenesis that have thus far been unappreciated by other methods.

Additional Information and Declarations

Competing Interests

Jordan Anaya has no competing interests to report.

Author Contributions

This project was performed by Jordan Anaya. JA conceived the project, performed all analyses, wrote the paper, and maintains the GitHub repository and any associated material at omnesres.com.

Funding

Jordan Anaya has no funding to report.

Acknowledgments

I acknowledge the contributions of the TCGA Research Network and its TCGA Pan-Cancer Analysis Working Group led by J.M. Stuart, C. Sander and I. Shmulevich. Without their efforts this type of analysis would not be possible.

References

bioRxiv preprint doi: https://doi.org/10.1101/030924; this version posted November 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1 Cancer Genome Atlas, N. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70, doi:10.1038/nature11412 (2012).

2 Cancer Genome Atlas, N. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330-337, doi:10.1038/nature11252 (2012).

3 Brennan, C. W. *et al.* The somatic genomic landscape of glioblastoma. *Cell* **155**, 462-477, doi:10.1016/j.cell.2013.09.034 (2013).

4 Cancer Genome Atlas Research, N. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499**, 43-49, doi:10.1038/nature12222 (2013).

5 Cancer Genome Atlas Research, N. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543-550, doi:10.1038/nature13385 (2014).

6 Cancer Genome Atlas Research, N. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519-525, doi:10.1038/nature11404 (2012).

7 Cancer Genome Atlas Research, N. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609-615, doi:10.1038/nature10166 (2011).

8 Cancer Genome Atlas Research, N. Integrated genomic characterization of papillary thyroid carcinoma. *Cell* **159**, 676-690, doi:10.1016/j.cell.2014.09.050 (2014).

9 Cancer Genome Atlas Research, N. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507**, 315-322, doi:10.1038/nature12965 (2014).

10 Cancer Genome Atlas Research, N. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513**, 202-209, doi:10.1038/nature13480 (2014).

11 Cancer Genome Atlas Research, N. *et al.* Integrated genomic characterization of endometrial carcinoma. *Nature* **497**, 67-73, doi:10.1038/nature12113 (2013).

12 Cancer Genome Atlas Research, N. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *The New England journal of medicine* **368**, 2059-2074, doi:10.1056/NEJMoa1301689 (2013).

13 Vogelstein, B. *et al.* Cancer Genome Landscapes. *Science* **339**, 1546-1558, doi:10.1126/science.1235122 (2013).

14 Zhao, Q. *et al.* Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA. *Briefings in bioinformatics* **16**, 291-303, doi:10.1093/bib/bbu003 (2015).

15 Chen, H. Y. *et al.* A five-gene signature and clinical outcome in non-small-cell lung cancer. *The New England journal of medicine* **356**, 11-20, doi:10.1056/NEJMoa060096 (2007).

16 van de Vijver, M. J. *et al.* A gene-expression signature as a predictor of survival in breast cancer. *The New England journal of medicine* **347**, 1999-2009, doi:10.1056/NEJMoa021967 (2002).

17 Valk, P. J. *et al.* Prognostically useful gene-expression profiles in acute myeloid leukemia. *The New England journal of medicine* **350**, 1617-1628, doi:10.1056/NEJMoa040465 (2004).

18 Bange, B., Zwick, E. & Ullrich, A. Molecular targets for breast cancer therapy and prevention. *Nature Medicine* **7**, 548-552 (2001).

19 Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* **12**, 323, doi:10.1186/1471-2105-12-323 (2011).

20 Hoadley, K. A. *et al.* Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929-944, doi:10.1016/j.cell.2014.06.049 (2014).

21 Fan, C. *et al.* Building prognostic models for breast cancer patients using clinical variables and hundreds of gene expression signatures. *BMC medical genomics* **4**, 3, doi:10.1186/1755-8794-4-3 (2011).

22 Cancer Genome Atlas Research, N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics* **45**, 1113-1120, doi:10.1038/ng.2764 (2013).

bioRxiv preprint doi: https://doi.org/10.1101/030924; this version posted November 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

23 Cox, D. R. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)* **34**, 187-220 (1972).

24 Claus, E. B. *et al.* Survival and low-grade glioma: the emergence of genetic information. *Neurosurgical focus* **38**, E6, doi:10.3171/2014.10.FOCUS12367 (2015).

25 Gyorffy, B., Surowiak, P., Budczies, J. & Lanczky, A. Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung cancer. *PloS one* **8**, e82241, doi:10.1371/journal.pone.0082241 (2013).

26 Wu, G. & Stein, L. A network module-based method for identifying cancer prognostic signatures. *Genome biology* **13**, R112, doi:10.1186/gb-2012-13-12-r112 (2012).

27 Zhang, W. *et al.* Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment. *PLoS computational biology* **9**, e1002975, doi:10.1371/journal.pcbi.1002975 (2013).

28 Zwiener, I., Frisch, B. & Binder, H. Transforming RNA-Seq data to improve the performance of prognostic gene signatures. *PloS one* **9**, e85150, doi:10.1371/journal.pone.0085150 (2014).

29 Yang, Y. *et al.* Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nature communications* **5**, 3231, doi:10.1038/ncomms4231 (2014).

30 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).

31 Hamilton, M. P. *et al.* Identification of a pan-cancer oncogenic microRNA superfamily anchored by a central core seed motif. *Nature communications* **4**, 2730, doi:10.1038/ncomms3730 (2013).

32 Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333-339, doi:10.1038/nature12634 (2013).

33 Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nature genetics* **45**, 1127-1133, doi:10.1038/ng.2762 (2013).

34 Knaack, S. A., Siahpirani, A. F. & Roy, S. A pan-cancer modular regulatory network analysis to identify common and cancer-specific network components. *Cancer informatics* **13**, 69-84, doi:10.4137/CIN.S14058 (2014).

35 Akbani, R. *et al.* A pan-cancer proteomic perspective on The Cancer Genome Atlas. *Nature communications* **5**, 3887, doi:10.1038/ncomms4887 (2014).

36 Linehan, W. M., Srinivasan, R. & Schmidt, L. S. The genetic basis of kidney cancer: a metabolic disease. *Nature reviews. Urology* **7**, 277-285, doi:10.1038/nrurol.2010.47 (2010).

37 Papadopoulos, P. *et al.* Developments in FINDbase worldwide database for clinically relevant genomic variation allele frequencies. *Nucleic acids research* **42**, D1020-1026, doi:10.1093/nar/gkt1125 (2014).

38 Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research* **33**, D514-517, doi:10.1093/nar/gki033 (2005).

39 Smigielski, E. M., Sirotkin, K., Ward, M. & Sherry, S. T. dbSNP: a database of single nucleotide polymorphisms. *Nucleic acids research* **28**, 352-355 (2000).

40 Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research* **42**, D980-985, doi:10.1093/nar/gkt1113 (2014).

41 Han, L. *et al.* The Pan-Cancer analysis of pseudogene expression reveals biologically and clinically relevant tumour subtypes. *Nature communications* **5**, 3963, doi:10.1038/ncomms4963 (2014).

42 Iyer, M. K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nature genetics* **47**, 199-208, doi:10.1038/ng.3192 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/030924; this version posted November 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

43 Chiu, C. H., Chou, T. Y., Chiang, C. L. & Tsai, C. M. Should EGFR mutations be tested in advanced lung squamous cell carcinomas to guide frontline treatment? *Cancer chemotherapy and pharmacology* **74**, 661-665, doi:10.1007/s00280-014-2536-3 (2014).

44 Schuller, A. *et al.* The MET inhibitor AZD6094 (Savolitinib, HMPL-504) induces regression in papillary renal cell carcinoma patient derived xenograft models. *Clinical cancer research : an official journal of the American Association for Cancer Research*, doi:10.1158/1078-0432.ccr-14-2685 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/030924; this version posted November 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Legends

Figure 1

Distinct expression patterns of protective and harmful prognostic genes

(A) Raw gene p-value distributions from multivariate Cox models for a cancer with high number of expressionally prognostic genes (EPGs; LGG, left), and a cancer with low number of EPGs (STAD, right). Distributions for the other 14 cancers are displayed in Fig. S1. (B) Unsupervised hierarchical clustering (Pearson correlation distance metric) of patients using the inverse normal transformed expression values from the 100 most significant protective genes and 100 most significant harmful gene for LGG, left, and STAD, right. (C) Kaplan Meier plots comparing survival times for the two broad clusters of patients identified in B and logrank p-values for LGG, left, and STAD, right.

Figure 2

Overlaps of prognostic genes and gene sets

(A) Heatmap displaying the overlaps between cancers of the 100 most significant genes of each cancer. (B) Overlaps within cancers of the 100 most significantly enriched gene sets for protective genes, and the 100 most significantly enriched gene sets for harmful genes. (C,D) Overlaps between cancers of the 100 most significantly enriched gene sets for harmful genes (C) and protective genes (D). (E) Venn diagram showing the overlaps of the 100 harmful gene sets for LIHC, LUAD, and KIRP. (F) Venn diagram showing the overlaps of the 100 protective gene sets for COAD, GBM, and LUSC.

Figure 3

Clustering of cancers using gene Cox coefficients

Clustering of genes and cancers using the sigmoidal normalized Cox coefficients of a list of genes that had an FDR less than or equal to .05 for at least four cancers. Pearson correlation distance metric was used for both row and column clustering, and Cox coefficients were row scaled (z-score).

Figure 4

bioRxiv preprint doi: https://doi.org/10.1101/030924; this version posted November 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

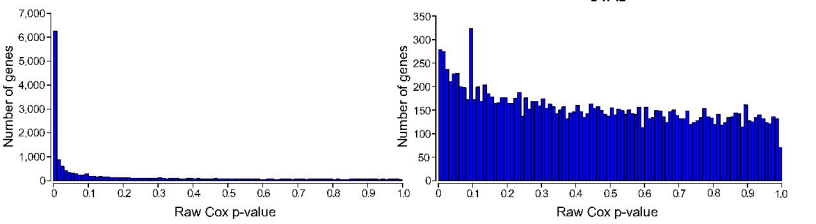
Clustering of cancers using gene programs

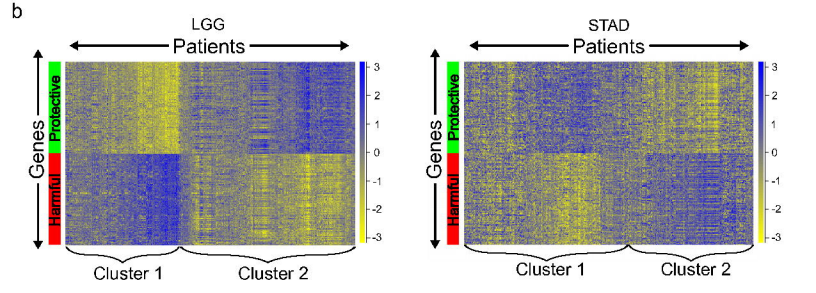
Using an established list of gene programs (see methods and Table S3), cancers and gene programs were clustered using the means of sigmoidal normalized Cox coefficients of the genes present in each program. Pearson correlation distance metric was used for both row and column clustering, and the average Cox coefficients were column scaled (z-score).

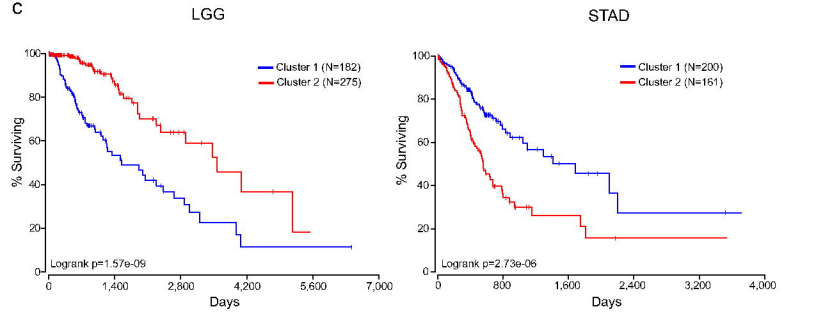
Table 1

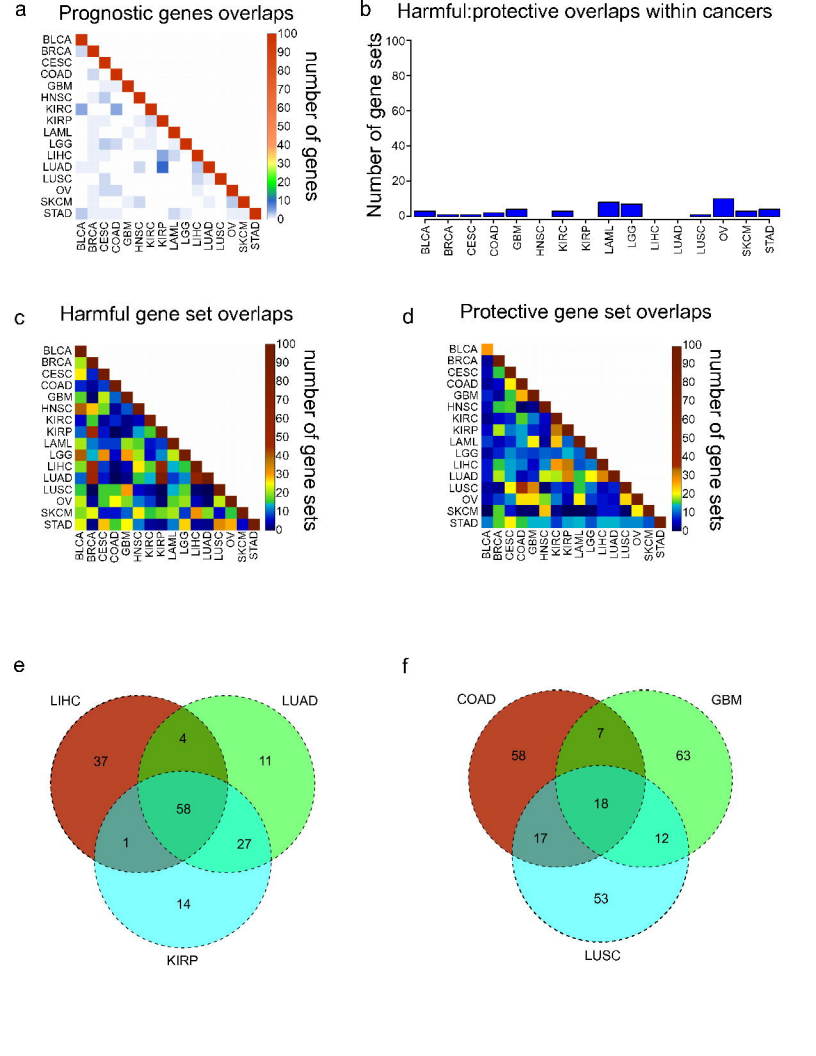
Characteristics of datasets and patients included in this study

Events are the number of deaths in the data set. Age is the average age and is in years. Median survival is in days. The median survival for KIRP could not be calculated.

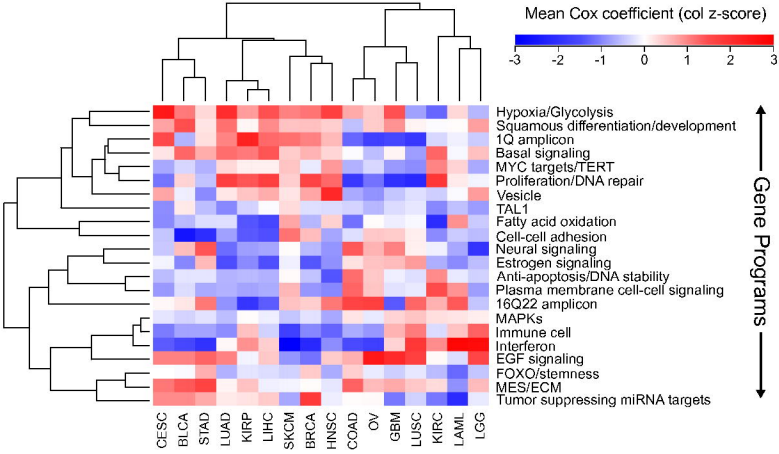

Number of genes
Raw Cox p-value
Number of genes
Raw Cox p-value


b
LGG
Patients
Genes
Protective
Harmful
Cluster 1
Cluster 2
STAD
Patients
Genes
Protective
Harmful
Cluster 1
Cluster 2


c
LGG
Cluster 1 (N=182)
Cluster 2 (N=275)
% Surviving
Logrank p=1.57e-09
Days
STAD
Cluster 1 (N=200)
Cluster 2 (N=161)
% Surviving
Logrank p=2.73e-06
Days


a
Prognostic genes overlaps
number of genes
BLCA
BRCA
CESC
COAD
GBM
HNSC
KIRC
KIRP
LAML
LGG
LIHC
LUAD
LUSC
OV
SKCM
STAD
b
Harmful:protective overlaps within cancers
Number of gene sets
c
Harmful gene set overlaps
number of gene sets
d
Protective gene set overlaps
number of gene sets
e
LIHC
LUAD
KIRP
37
4
11
58
1
27
14
f
COAD
GBM
LUSC
58
7
63
18
17
12
53

Prognonstic Genes
CESC
STAD
LGG
BLCA
HNSC
BRCA
LIHC
KIRP
LUAD
KIRC
SKCM
LAML
GBM
LUSC
COAD
OV
Cox coefficient (row z-score)
2
1
0
-1
-2


Mean Cox coefficient (col z-score)
-3
-2
-1
0
1
2
3
Hypoxia/Glycolysis
Squamous differentiation/development
1Q amplicon
Basal signaling
MYC targets/TERT
Proliferation/DNA repair
Vesicle
TAL1
Fatty acid oxidation
Cell-cell adhesion
Neural signaling
Estrogen signaling
Anti-apoptosis/DNA stability
Plasma membrane cell-cell signaling
16Q22 amplicon
MAPKs
Immune cell
Interferon
EGF signaling
FOXO/stemness
MES/ECM
Tumor suppressing miRNA targets
Gene Programs
CESC
BLCA
STAD
LUAD
KIRP
LIHC
SKCM
BRCA
HNSC
COAD
OV
GBM
LUSC
KIRC
LAML
LGG

Cancer	Patients	Median Survival	Events	Age at Diagnosis	Male/Female	RNASeqV2	Genes in study	FDR<=.05
BLCA	347	1064	139	67.9	254/93	YES	16385	532
BRCA	981	3669	116	58.4	10/971	YES	16649	30
CESC	259	3097	60	48.1	0/259	YES	16358	146
COAD	434	2475	89	66.6	232/202	YES	16414	0
GBM	152	406	119	59.8	98/54	YES	16833	0
HNSC	484	1671	190	61.2	353/131	YES	16652	45
KIRC	516	2386	167	60.7	335/181	YES	16677	5785
KIRP	247	NA	32	60.8	181/66	YES	16430	2415
LAML	149	577	92	54.7	80/69	YES	15255	4
LGG	457	2875	91	43.1	255/202	YES	16818	7186
LIHC	324	2116	105	60.1	217/107	YES	15855	2
LUAD	486	1379	146	65.3	224/262	YES	16784	1179
LUSC	471	1655	180	67.4	352/119	YES	16979	0
SKCM	427	2889	195	57.5	264/163	YES	16067	1548
OV	401	1321	224	59.6	0/401	NO	15748	0
STAD	361	874	130	65.2	238/123	NO	15560	0