OncoRank: A pan-cancer method of combining survival correlations and its application to mRNAs, miRNAs, and lncRNAs

Jordan Anaya[1]

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

Corresponding Author:

Jordan Anaya[1]

Charlottesville, VA, US

Email address: omnesresnetwork@gmail.com

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2574v1 | CC BY 4.0 Open Access | rec: 3 Nov 2016, publ:

OncoRank: A pan-cancer method of combining survival correlations and its application to mRNAs, miRNAs, and lncRNAs

Jordan Anaya[1]

[1]**Omnes Res**

ABSTRACT

OncoRank adopts a method for finding recurrent miRNA-target interactions to find genes with consistent relationships to patient survival across cancers. Genes are first ranked in each cancer by their Cox coefficients, and these ranks are then combined by applying Fisher's method. Using ranks instead of the raw coefficients or p-values allows each cancer to be weighted equally and prevents bias from cancers with large numbers of patients. OncoLnc is a newly available resource for Cox coefficients and utilizes data from 21 cancers in The Cancer Genome Atlas. Using this resource I applied OncoRank to mRNAs, miRNAs, and lncRNAs and in each case found consistently harmful or protective genes. These genes may be members of central cancer pathways and should be of interest to cancer researchers.

Keywords: cancer, mRNA, miRNA, lncRNA, survival, oncogene

MOTIVATION

Identifying the gene mutations which lead to cancer or result in aggressive tumors is a well understood problem. There are many spurious mutations in cancers, but there are specific mutations which occur more often than chance and this distinguishes driver mutations from passenger mutations. However, when a gene does not show consistent molecular alterations it is more difficult to implicate it in cancer pathogenesis. In this case the prevailing strategy is to correlate the gene's expression with patient survival via Cox regression (Cox, 1972).

Thanks to The Cancer Genome Atlas (TCGA) it is possible to correlate gene alterations to patient survival at an unprecedented scale. I previously took advantage of this resource to correlate expression with survival for Tier 3 mRNAs, Tier 3 miRNAs, and MiTranscriptome beta lncRNAs, and made the data searchable with OncoLnc and available to download (Anaya, 2016). This resource is meant for identifying which genes are most highly correlated with survival in each cancer, but it is interesting to think how the correlations in OncoLnc can be combined across cancers to identify the genes which are consistently correlated to survival.

One method might be to simply look at the genes which are most highly correlated to survival in each cancer and see which genes show up most often. This approach has several issues. As seen in Figure 1 the most correlated genes are not shared between cancers. I took the top 5% of genes in each cancer based on p-value and recorded how often the genes were shared for all possible combinations of 2, 3, 4, or 5 cancers. For example, for mRNAs I looked at the top 930 genes in each cancer. When looking at two random cancers on average 66 genes (7% of 930) are shared. Although this number is significant by a hypergeometric test (18,616 genes in population, 930 successes in population, sample size 930, 66 successes in sample, p=.0025), the number of genes shared by all cancers in a random combination quickly approaches zero as more cancers are added to the combination.

Another issue with using p-values from the Cox regressions is that there are two ways to obtain a small p-value, either significantly increase the hazard ratio or significantly decrease the hazard ratio. Although it is entirely possible

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2574v1 | CC BY 4.0 Open Access | rec: 3 Nov 2016, publ:

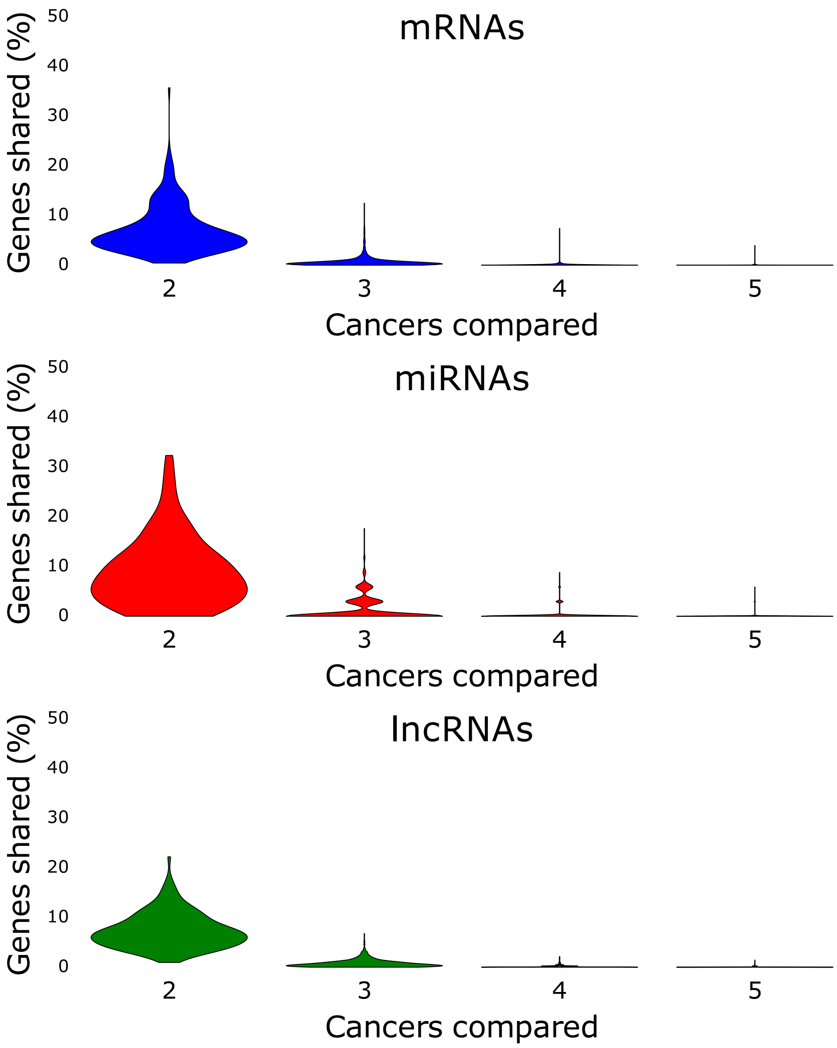


Figure 1. The top 5% genes (based on p-value) of each cancer were compared. For each sample size all possible combinations of cancers were enumerated and the number of genes shared by all cancers in each combination were recorded.

a gene may positively affect survival in the milieu of one cancer and negatively affect survival in the environment of a different cancer, it is unclear if these types of genes are likely to be part of central cancer pathways and how they should be scored.

Under the hypothesis that genes part of core cancer pathways should consistently impact the hazard ratio across cancers we can look for genes which have consistent Cox coefficients. In the field of small RNA biology a method was developed for finding consistent miRNA-mRNA correlations across cancers and the results were supported by both experimental and computational results (Jacobsen et al., 2013). This method ranked miRNAs by coefficients obtained from multiple linear regressions and then combined the ranks with Fisher's method, resulting in a recurrence (REC) score. The Cox coefficients from the multivariate Cox regressions in OncoLnc are analogous to the coefficients from multiple linear regressions. As a result, to find genes consistently correlated to survival I decided to follow the methodology used to calculate REC scores with the Cox coefficients from the OncoLnc GitHub repository.

METHODOLOGY

The Cox coefficients in OncoLnc are from multivariate Cox regressions that include age, grade, and sex as multivariates whenever possible. For each species (mRNA, miRNA, lncRNA), expression cutoffs were chosen since it is meaningless to perform correlations with genes that are not expressed. In addition, RNA-SEQ was used as a measure of expression in every case except for GBM miRNAs, which I am excluding from this analysis. To prevent extreme RNA-SEQ values from affecting the Cox regressions all expression data was inverse normal transformed prior to running the Cox regressions.

Each multivariate in a Cox model has a coefficient which describes that variable's contribution to the hazard ratio. A positive coefficient indicates the variable increases the hazard ratio. For example, age should be multiplied by a positive coefficient since older patients have a higher chance of death. All of the coefficients in OncoLnc are the coefficients from the gene term of the model, and when I refer to Cox coefficients I am specifically referring to the gene Cox coefficient. A positive Cox coefficient indicates that a higher expression of that gene increases the chance of an event (in this case death).

To compute REC scores for the genes in OncoLnc I calculate the scores for the three data types, mRNAs, miRNAs, lncRNAs, separately. The first step in computing REC scores is to sort the genes in each cancer by their coefficients and compute a relative rank:

$$rr_{u,k,j} = \frac{r_{u,k,j}}{|L_{k,j}|} - \frac{1}{2|L_{k,j}|}$$

In this formula $r_{u,k,j}$ is the rank of gene u in cancer k for data type j. The list of all ranks for a data type and cancer is specified by $L_{k,j}$ and $|L_{k,j}|$ is the length of that list. As mentioned above, due to the expression cutoff used $|L_{k,j}|$ will be different for each cancer. The last term ensures the values are less than one but still more than 0.

Relative ranks can then be evaluated using Fisher's method:

$$-2\sum_{k} ln(p_k) \approx \chi^2_{2n}$$

The natural log of the p-values for the rank in each cancer are summed and evaluated with a chi-squared statistic with 2n degrees of freedom where n is the number of cancers that were summed over. And it just so happens that the relative rank, $rr_{u,k,j}$, is a p-value. This can be seen by imagining that we are looking at 100 genes. What is the chance of obtaining the first rank or better? One out of a hundred. What is the chance of obtaining the second rank or better? Two out of a hundred.

The null hypothesis is that genes do not show consistent correlations across the different cancers. If the null were true then a p-value distribution of the p-values from Fisher's method would result in a flat distribution. To see if I could obtain results above a random distribution I applied Fisher's method to each data type with the genes ranked from the most negative Cox coefficient to the most positive Cox coefficient in each cancer, and required that a gene be expressed in at least 8 cancers to be included in the analysis.

As can seen in Figure 2 none of the distributions are flat. Interestingly, both low p-values and large p-values are enriched. A low p-value would correspond to a gene that consistently had negative Cox coefficients since genes were ranked from negative to positive Cox coefficients. A high p-value indicates that a gene consistently had positive Cox coefficients. From the plots it looks like genes are more likely to consistently increase the hazard ratio than consistently decrease the hazard ratio.

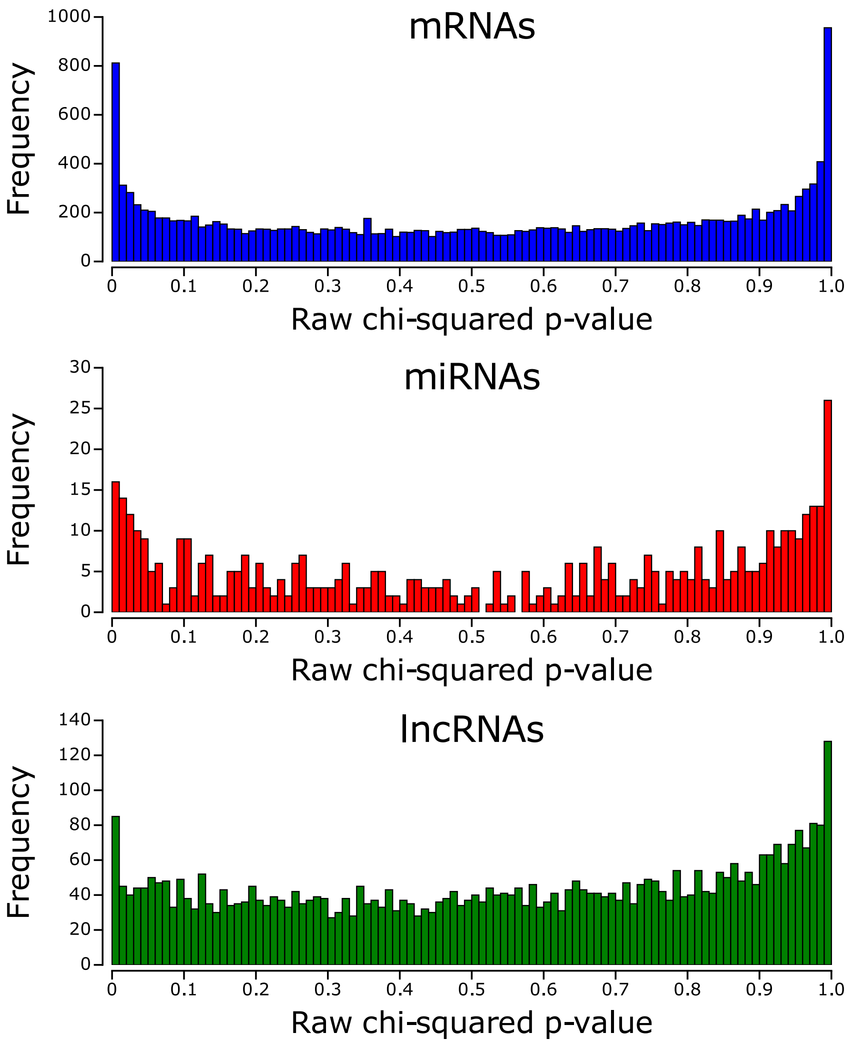


Figure 2. The p-value distributions for each data type are shown.

To account for the two tails of the distributions we actually need to apply Fisher's method twice. Once with the genes sorted negative to positive, and once with the genes sorted positive to negative. The p-values for each gene will then be compared and the smaller p-value will be kept. To account for the multiple testing the p-values will be multiplied by 2. Using these new p-values we can then calculate recurrence (REC) scores for genes in each data type:

$$REC_{u,j} = \begin{cases} log_{10}(2p_-) & \text{if } p_- < p_+ \\ -log_{10}(2p_+) & \text{if } p_+ < p_- \\ 0 & \text{if } p_- = p_+ \end{cases}$$

Where p_- are the p-values obtained from ranking genes from negative to positive and p_+ are the p-values obtained from sorting the genes from positive to negative. A large negative REC score indicates that the gene consistently has negative Cox coefficients and is protective across cancers. A large positive REC score indicates that the gene consistently has positive Cox coefficients and is harmful across cancers.

I again looked at the distributions after performing Fisher's method on the genes ranked in both directions and merging the results. As can be seen in Figure 3 there is now only a pileup of small p-values, which is what should happen if the null hypothesis is wrong.

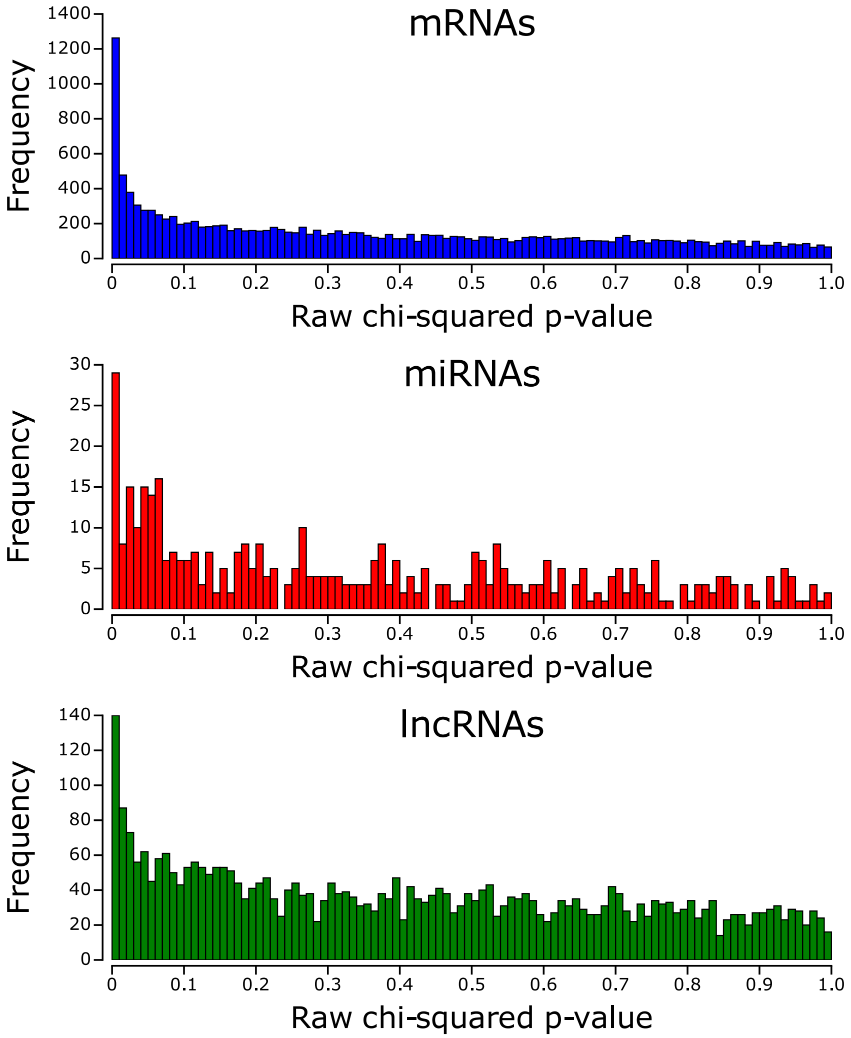
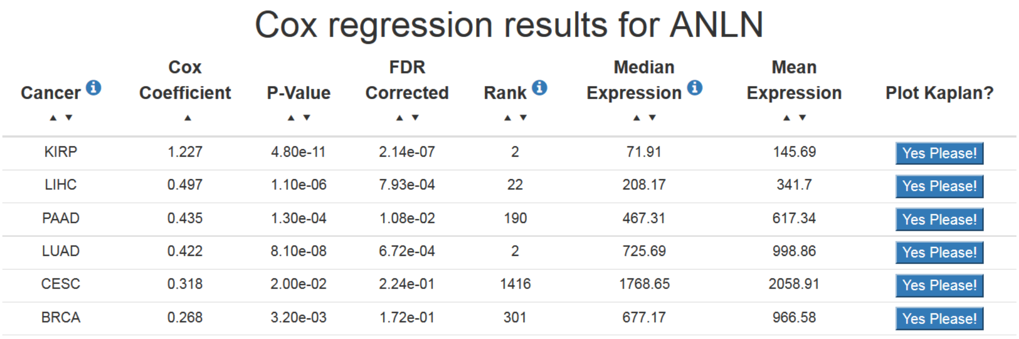


Figure 3. The p-value distributions for each data type are shown after merging of p-values.

To account for multiple testing I used the Benjamini-Hochberg method to calculate a FDR corrected p-value for each gene. All of the REC scores calculated for the three data types are available in Table S1 and are sorted by the FDR corrected p-value.

DISCUSSION

The REC scores in Table S1 can be explored using OncoLnc. For example, inputting the highest scored mRNA, ANLN, gives the output shown in Figure 4. The REC score for ANLN is 7.76, indicating it should consistently be harmful and have positive Cox coefficients across cancers. As can be seen in Figure 4, ANLN does indeed have positive Cox coefficients and low ranks (the ranks in OncoLnc are based on p-value). A quick literature search on this gene shows that it is likely to play a key role in cancer pathogenesis. Anillin interacts with the cytoskeleton and is important for cell division Piekny and Glotzer (2008), and it is easy to see how this gene may increase the hazard ratio by increasing cell division or metastasis. Indeed, there is experimental evidence in several cancers that anillin does just that (Suzuki et al., 2005; Wang et al., 2015; Zhou et al., 2015). Perhaps confirming its role in metastasis is the fact that its most negative Cox coefficient occurs in LAML, a blood cancer.



Cox regression results for ANLN

Cancer	Cox Coefficient	P-Value	FDR Corrected	Rank	Median Expression	Mean Expression	Plot Kaplan?
KIRP	1.227	4.80e-11	2.14e-07	2	71.91	145.69	Yes Please!
LIHC	0.497	1.10e-06	7.93e-04	22	208.17	341.7	Yes Please!
PAAD	0.435	1.30e-04	1.08e-02	190	467.31	617.34	Yes Please!
LUAD	0.422	8.10e-08	6.72e-04	2	725.69	998.86	Yes Please!
CESC	0.318	2.00e-02	2.24e-01	1416	1768.65	2058.91	Yes Please!
BRCA	0.268	3.20e-03	1.72e-01	301	677.17	966.58	Yes Please!

Figure 4. OncoLnc search results for ANLN sorted by Cox coefficients.

Looking through Table S1 or examining Figure 3 reveals that there are more mRNAs with significant REC scores than miRNAs and lncRNAs. This is largely due to there being more Tier 3 mRNAs than Tier 3 miRNAs or MiTranscriptome beta lncRNAs. There are 18,616 Tier 3 mRNAs in OncoLnc, but only 684 Tier 3 miRNAs (excluding GBM), and only 8,296 MiTranscriptome beta lncRNAs. In addition, lncRNAs are less likely to meet the cutoff of being expressed in at least 8 cancers. 16,966 out of 18,616 mRNAs met the cutoff, but only 4,408 out of 8,296 lncRNAs met the cutoff (there is some overlap in the genes that are Tier 3 mRNAs and the MiTranscriptome beta dataset). A clear natural extension of OncoRank would be to apply it on a larger gene pool. For example, the entire MiTranscriptome dataset contains over 90,000 genes (Iyer et al., 2015). If this dataset is released OncoRank could be used to prioritize the study of these novel transcripts. Additionally, miRBase lists over 2,500 mature miRNAs. If read counts for these miRNAs in the TCGA dataset become available OncoRank could be run on these as well.

An argument could be made that OncoRank should only be run on a group of cancers that have similar pathogenesis. As mentioned above, if a gene is involved in metastasis it might not make sense to include blood derived cancers in the analysis. On the other hand, if you don't know what you are looking for, or are looking for genes of unknown function, it may not be clear what set of cancers should be used in the analysis.

CONCLUSIONS

OncoRank is able to identify which genes show consistent relationships with survival across cancers. Because the genes have the same effect on survival across cancers these genes may reveal the pathways which are shared by aggressive tumors, or alternatively reveal which pathways decrease pathogenicity. OncoRank also shows the potential of pan-cancer analyses and how databases such as OncoLnc can be utilized.

METHODS

All the code for OncoRank and for generating the figures in this publication is available at the GitHub repository. The code was run with Python 2.7 and requires SciPy and rpy2.

ACKNOWLEDGMENTS

This project was made possible by data generated by the TCGA Research Network.

REFERENCES

Anaya, J. (2016). Oncolnc: linking tcga survival data to mrnas, mirnas, and lncrnas. *PeerJ Computer Science*, 2:e67.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34:187–220.

Iyer, M. K., Niknafs, Y. S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T. R., Prensner, J. R., Evans, J. R., Zhao, S., Poliakov, A., Cao, X., Dhanasekaran, S. M., Wu, Y. M., Robinson, D. R., Beer, D. G., Feng, F. Y., Iyer,

H. K., and Chinnaiyan, A. M. (2015). The landscape of long noncoding rnas in the human transcriptome. *Nat Genet*, 47(3):199–208.

Jacobsen, A., Silber, J., Harinath, G., Huse, J. T., Schultz, N., and Sander, C. (2013). Analysis of microrna-target interactions across diverse cancer types. *Nat Struct Mol Biol*, 20(11):1325–32.

Piekny, A. J. and Glotzer, M. (2008). Anillin is a scaffold protein that links rhoa, actin, and myosin during cytokinesis. *Curr Biol*, 18(1):30–6.

Suzuki, C., Daigo, Y., Ishikawa, N., Kato, T., Hayama, S., Ito, T., Tsuchiya, E., and Nakamura, Y. (2005). 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(24):11314–25.

Wang, S., Mo, Y., Midorikawa, K., Zhang, Z., Huang, G., Ma, N., Zhao, W., Hiraku, Y., Oikawa, S., and Murata, M. (2015). The potent tumor suppressor mir-497 inhibits cancer phenotypes in nasopharyngeal carcinoma by targeting anln and hspa4l. *Oncotarget*, 6(34):35893–907.

Zhou, W., Wang, Z., Shen, N., Pi, W., Jiang, W., Huang, J., Hu, Y., Li, X., and Sun, L. (2015). Knockdown of anln by lentivirus inhibits cell growth and migration in human breast cancer. *Mol Cell Biochem*, 398(1-2):11–9.