PeerJ Computer Science

OncoLnc: linking TCGA survival data to mRNAs, miRNAs, and lncRNAs

Jordan Anaya

Omnesres.com, Charlottesville, United States

ABSTRACT

OncoLnc is a tool for interactively exploring survival correlations, and for downloading clinical data coupled to expression data for mRNAs, miRNAs, or long noncoding RNAs (lncRNAs). OncoLnc contains survival data for 8,647 patients from 21 cancer studies performed by The Cancer Genome Atlas (TCGA), along with RNA-SEQ expression for mRNAs and miRNAs from TCGA, and lncRNA expression from MiTranscriptome beta. Storing this data gives users the ability to separate patients by gene expression, and then create publication-quality Kaplan-Meier plots or download the data for further analyses. OncoLnc also stores precomputed survival analyses, allowing users to quickly explore survival correlations for up to 21 cancers in a single click. This resource allows researchers studying a specific gene to quickly investigate if it may have a role in cancer, and the supporting data allows researchers studying a specific cancer to identify the mRNAs, miRNAs, and lncRNAs most correlated with survival, and researchers looking for a novel lncRNA involved with cancer lists of potential candidates. OncoLnc is available at http://www.oncolnc.org.

Subjects Bioinformatics, Computational Biology, Databases
Keywords TCGA, Cancer, RNA-SEQ, mRNA, miRNA, lncRNA, Kaplan-Meier, Cox regression, Database

Submitted 23 February 2016
Accepted 13 May 2016
Published 13 June 2016

Corresponding author
Jordan Anaya,
omnesresnetwork@gmail.com

Academic editor
Hsueh-Fen Juan

Additional Information and Declarations can be found on page 11

DOI **10.7717/peerj-cs.67**

INTRODUCTION

The Cancer Genome Atlas (TCGA) provides researchers with unprecedented amounts of molecular data along with clinical and histopathological information (http://cancergenome.nih.gov/). This data set has not only led to increases in our understanding of cancer (*Ciriello et al., 2013*; *Hoadley et al., 2014*), but its scale has also allowed for previously impossible projects such as a comprehensive cataloguing of the human transcriptome (*Han et al., 2014*; *Iyer et al., 2015*). However, the size and complexity of this unique data set makes it difficult for cancer researchers to access and fully utilize.

Multiple resources exist to help researchers download or explore TCGA data (*Cerami et al., 2012*; *Győrffy et al., 2013*; *Koch et al., 2015*). However, a tool does not exist that focuses on survival analyses with TCGA data. Although cBioPortal (http://www.cbioportal.org/) does allow for simple Kaplan-Meier analyses for a range of TCGA cancer studies, it does not allow for more rigorous analyses with Cox regression. More importantly, the p-value obtained from a single analysis can be misleading, and it may be more informative to consider the relative strength of the correlation (*Anaya et al., 2016*).

Furthermore, current tools for survival analyses only allow users to view the results from one cancer at a time. This makes it difficult for a researcher to perform a comprehensive survival study with their gene of interest, and makes it possible that researchers will miss an interesting correlation. Current tools also do not allow for generation of Kaplan-Meier plots with user defined upper and lower groups, and do not allow for simple download of the survival data coupled to the expression data.

It is also important to note that the gene names used in TCGA data are outdated. While many online portals use current mRNA definitions, there is not an online data portal that uses modern miRNA definitions. This is because the TCGA Tier 3 read counts for the 5p and 3p arms are aggregated into one count for the stem-loop, which makes it difficult for researchers who want to obtain information for a specific mature miRNA.

In addition, although the role of long noncoding RNAs (lncRNAs) in cancer is beginning to be appreciated (*Yarmishyn & Kurochkin, 2015*), the Tier 3 TCGA mRNA files contain expression data for only the limited number of lncRNAs that were known at the initiation of the TCGA project. As a result, tools for exploring TCGA data will not contain many lncRNAs currently being studied. Although a platform has already been developed to fill this gap (*Li et al., 2015*), to help the scientific community study lncRNAs OncoLnc incorporates analyses and data for MiTranscriptome beta lncRNAs, http://mitranscriptome.org/, in addition to Tier 3 TCGA mRNAs and miRNAs.

MATERIALS AND METHODS

Code, files, and software

All of the code necessary to reproduce Tables 1–3 and S1–S3, and therefore the data in OncoLnc, along with a limited amount of raw data and intermediate files is located at https://github.com/OmnesRes/onco_lnc. The rest of the raw data can be downloaded from https://tcga-data.nci.nih.gov/tcga/ and http://mitranscriptome.org/. This code was run with Python 2.7.5, NumPy 1.7.1, and rpy2 2.5.6, and can require upwards of 6GB of RAM. OncoLnc runs on Django 1.8.2, Python 2.7, matplotlib 1.2.1, NumPy 1.7.1, rpy2 2.5.6, uses the SQLite3 database engine, and utilizes Bootstrap CSS and JavaScript, and Font Awesome icons.

Cox models

Multivariate Cox regressions were performed with the coxph function from the R survival library. For each cancer and data type an attempt was made to construct a model with gene expression, sex, age, and grade or histology as multivariates. However, not every cancer contained complete grade information, and some cancers did not contain different grades. As a result, grade was not able to be included in every model. Cancers which have grade in the model are BLCA, BRCA, CESC, ESCA, HNSC, KIRC, LGG, LIHC, OV, PAAD, STAD, and UCEC. Cancers with grade missing from the model include COAD, GBM, KIRP, LAML, LUAD, LUSC, READ, SARC, and SKCM. When grade is available each grade is listed as a separate term in the models and is either 1 or 0. Similarly, sex is also listed in the models as 1 or 0. Gene expression is inverse normal transformed before inclusion into the models, except for the GBM miRNA data, which was normalized microarray data. Age is listed in the models as is. The multivariate model run for each cancer and each data type is listed at the top of Tables S1–S3, and the code for running all

Table 1 **Characteristics of the Tier 3 mRNA datasets in OncoLnc.** Age at diagnosis is in years, and is an average. The events indicate the number of deaths in the dataset. Median survival is in days and could not be calculated for COAD, KIRP, OV, READ, and UCEC.

Cancer	Patients	Male/ Female	Age at diagnosis	Events	Median survival	Genes in OncoLnc
BLCA	403	296/107	68.03	177	1,008	16,339
BRCA	1,006	11/995	58.34	135	3,941	16,602
CESC	264	0/264	48.23	59	4,086	16,330
COAD	440	235/205	66.58	85	NA	16,378
ESCA	144	126/18	60.51	59	801	16,790
GBM	152	99/53	59.84	27	1,426	16,783
HNSC	497	364/133	61.24	207	1,732	16,614
KIRC	523	341/182	60.56	167	2,764	16,638
KIRP	285	210/75	61.45	44	NA	16,399
LAML	151	81/70	54.40	92	577	15,227
LGG	510	282/228	43.02	124	2,835	16,781
LIHC	360	244/116	59.41	126	1,694	15,824
LUAD	492	225/267	65.32	176	1,492	16,748
LUSC	489	362/127	67.23	169	2,224	16,942
OV	294	0/294	59.19	42	NA	16,893
PAAD	175	96/79	64.37	92	607	17,177
READ	159	88/71	64.58	22	NA	16,472
SARC	259	118/141	60.71	98	1,991	16,197
SKCM	459	284/175	58.14	215	2,454	16,030
STAD	379	247/132	65.49	146	1,043	16,885
UCEC	541	0/541	63.95	90	NA	16,670

of the Cox regressions is present in the GitHub repository: https://github.com/OmnesRes/onco_lnc.

All the clinical data was downloaded from https://tcga-data.nci.nih.gov/tcga/ January 5th and 6th, 2016. Cancers were chosen for the study based on the quality and amount of overall survival data. It is possible for a cancer such as PRAD to have a large amount of survival information, but a low number of events (deaths), making a survival analysis difficult. The Tier 3 mRNA and miRNA data was also downloaded from https://tcga-data.nci.nih.gov/tcga/, while the MiTranscriptome data was downloaded from http://mitranscriptome.org/. Definitions for miRNAs were downloaded from http://www.mirbase.org/.

For each cancer, only patients who contained all the necessary clinical information were included in the analysis. In addition, patients had to have a follow up time or time to death greater than 0 days. For each cancer, only genes which met an expression cutoff were included in the analysis (see below for more details). In general, only primary solid tumors were included in analyses, and this is implemented by only using samples with "01" in the patient barcode. The exceptions are LAML, which is a blood derived cancer, and therefore has the designation "03," and SKCM, which contains primarily metastatic

Table 2 **Characteristics of the Tier 3 miRNA datasets in OncoLnc.** Age at diagnosis is in years, and is an average. The events indicate the number of deaths in the dataset. Median survival is in days and could not be calculated for COAD, KIRP, READ, and UCEC.

Cancer	Patients	Male/ Female	Age at diagnosis	Events	Median survival	Genes in OncoLnc
BLCA	404	297/107	68.02	177	1,036	512
BRCA	988	11/977	58.35	131	3,941	479
CESC	267	0/267	48.27	59	4,086	501
COAD	426	226/200	66.48	84	NA	476
ESCA	144	125/19	60.61	59	801	494
GBM	561	343/218	57.94	67	2,648	507
HNSC	501	363/138	61.30	208	1,732	514
KIRC	506	331/175	60.48	165	2,764	448
KIRP	286	210/76	61.52	44	NA	430
LAML	164	88/76	54.05	100	518	374
LGG	506	278/228	43.07	123	2,660	486
LIHC	362	248/114	59.41	125	1,791	485
LUAD	490	226/264	65.35	175	1,492	493
LUSC	467	346/121	67.43	160	2,224	519
OV	470	0/470	59.85	92	3,128	467
PAAD	175	96/79	64.37	92	607	494
READ	154	84/70	64.23	22	NA	495
SARC	259	119/140	60.85	98	1,991	455
SKCM	438	271/167	58.01	207	2,470	535
STAD	400	260/140	65.54	155	1,043	495
UCEC	534	0/534	63.91	87	NA	518

tumors, and therefore designations "01" and "06" were allowed for SKCM analyses. It is possible for a patient to have more than one sequencing file, and in these cases the counts were averaged.

RESULTS

Overview of OncoLnc

OncoLnc stores over 400,000 analyses, which includes Cox regression results as well as mean and median expression of each gene. For the Cox regression results, in addition to p-values, OncoLnc stores the rank of the correlation. Different cancers contain very different p-value distributions (*Anaya et al., 2016*; *Yang et al., 2014*), and it is unclear what causes this difference. As a result, using one p-value cutoff across cancers is not possible, and the rank of the correlation is a simple way to measure the relative strength of the correlation. The rank is calculated per cancer, per data type. Tables 1–3 contain information about how many genes there are for each cancer and each data type.

The mRNA and miRNA identifiers used by TCGA are out of date, and the identifiers in OncoLnc have been manually curated using NCBI Gene: http://www.ncbi.nlm.nih.gov/gene, and recent miRBase definitions: http://www.mirbase.org/. Over 2,000

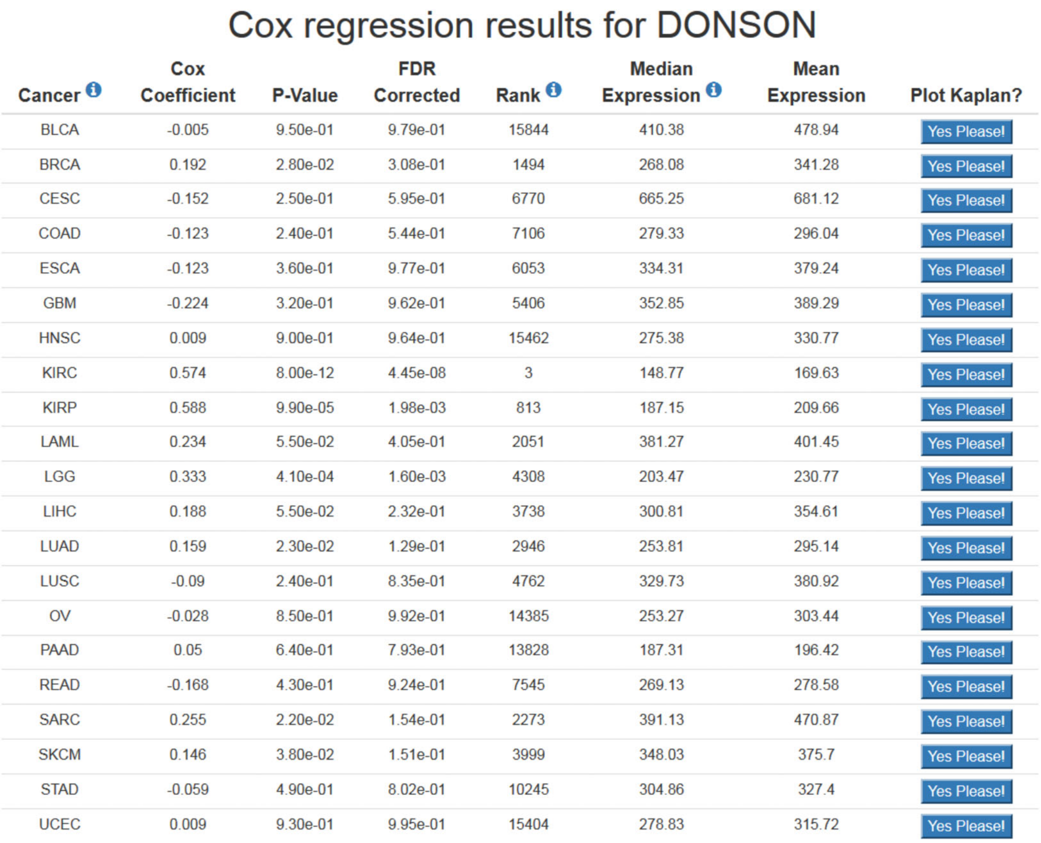
Table 3 **Characteristics of the MiTranscriptome beta analyses in OncoLnc.** Age at diagnosis is in years, and is an average. The events indicate the number of deaths in the dataset. Median survival is in days and could not be calculated for COAD, KIRP, and UCEC.

Cancer	Patients	Male/ Female	Age at diagnosis	Events	Median survival	Genes in OncoLnc
BLCA	120	86/34	67.37	61	706	4,322
BRCA	766	8/758	58.03	111	3,941	4,708
CESC	106	0/106	48.22	26	3,046	4,493
COAD	117	52/65	69.64	24	NA	3,302
GBM	144	94/50	59.56	24	1,426	4,524
HNSC	288	211/77	61.40	133	1,762	4,314
KIRC	457	299/158	60.75	156	2,764	5,191
KIRP	73	51/22	59.78	17	NA	4,627
LAML	20	15/5	54.75	10	580	3,940
LGG	217	123/94	42.82	65	2,660	4,875
LIHC	65	40/25	60.97	41	1,005	3,610
LUAD	320	148/172	65.72	118	1,357	4,636
LUSC	330	244/86	67.16	112	2,284	4,979
OV	369	0/369	59.63	69	3,128	4,901
READ	42	22/20	66.67	8	1,581	3,310
SKCM	255	159/96	56.82	148	2,192	3,893
STAD	148	93/55	65.72	56	940	4,619
UCEC	274	0/274	63.12	39	NA	3,706

mRNA symbols were updated, and these are listed in Table S4. Genes which have had their Entrez Gene ID removed from NCBI Gene, or could not be confidently mapped to a single identifier, are not included in OncoLnc but are still included in Table S1.

Using OncoLnc is very straightforward. The preferred method of using OncoLnc is to submit a gene at the home page, and this submission is not case sensitive. If a user submits a gene not in the database, they will be notified and provided with links to all the possible gene names and IDs. Submission of a valid gene identifier will return correlation results for up to 21 cancers for mRNAs and miRNAs, or 18 cancers for MiTranscriptome beta lncRNAs (Fig. 1). If a gene does not meet the expression cutoff for the analysis, it will not be present in the database, and therefore a user may receive less than the maximum possible number of results. For users using OncoLnc on smaller devices, it is possible to perform a single cancer search. The link for this search is on the home page, and the user must submit the TCGA cancer abbreviation along with the gene of interest.

At the results page is a link to perform a Kaplan-Meier analysis for each cancer (Fig. 1). The user will be asked how they would like to divide the patients. Patients can be split into any non-overlapping upper and lower slices, for example upper 25 percent and lower 25 percent. Upon submission users will be presented with a PNG Kaplan-Meier plot, a logrank p-value for the analysis, and text boxes with the data that was plotted (Fig. 2).



Cox regression results for DONSON

Cancer	Cox Coefficient	P-Value	FDR Corrected	Rank	Median Expression	Mean Expression	Plot Kaplan?
BLCA	-0.005	9.50e-01	9.79e-01	15844	410.38	478.94	Yes Please!
BRCA	0.192	2.80e-02	3.08e-01	1494	268.08	341.28	Yes Please!
CESC	-0.152	2.50e-01	5.95e-01	6770	665.25	681.12	Yes Please!
COAD	-0.123	2.40e-01	5.44e-01	7106	279.33	296.04	Yes Please!
ESCA	-0.123	3.60e-01	9.77e-01	6053	334.31	379.24	Yes Please!
GBM	-0.224	3.20e-01	9.62e-01	5406	352.85	389.29	Yes Please!
HNSC	0.009	9.00e-01	9.64e-01	15462	275.38	330.77	Yes Please!
KIRC	0.574	8.00e-12	4.45e-08	3	148.77	169.63	Yes Please!
KIRP	0.588	9.90e-05	1.98e-03	813	187.15	209.66	Yes Please!
LAML	0.234	5.50e-02	4.05e-01	2051	381.27	401.45	Yes Please!
LGG	0.333	4.10e-04	1.60e-03	4308	203.47	230.77	Yes Please!
LIHC	0.188	5.50e-02	2.32e-01	3738	300.81	354.61	Yes Please!
LUAD	0.159	2.30e-02	1.29e-01	2946	253.81	295.14	Yes Please!
LUSC	-0.09	2.40e-01	8.35e-01	4762	329.73	380.92	Yes Please!
OV	-0.028	8.50e-01	9.92e-01	14385	253.27	303.44	Yes Please!
PAAD	0.05	6.40e-01	7.93e-01	13828	187.31	196.42	Yes Please!
READ	-0.168	4.30e-01	9.24e-01	7545	269.13	278.58	Yes Please!
SARC	0.255	2.20e-02	1.54e-01	2273	391.13	470.87	Yes Please!
SKCM	0.146	3.80e-02	1.51e-01	3999	348.03	375.7	Yes Please!
STAD	-0.059	4.90e-01	8.02e-01	10245	304.86	327.4	Yes Please!
UCEC	0.009	9.30e-01	9.95e-01	15404	278.83	315.72	Yes Please!

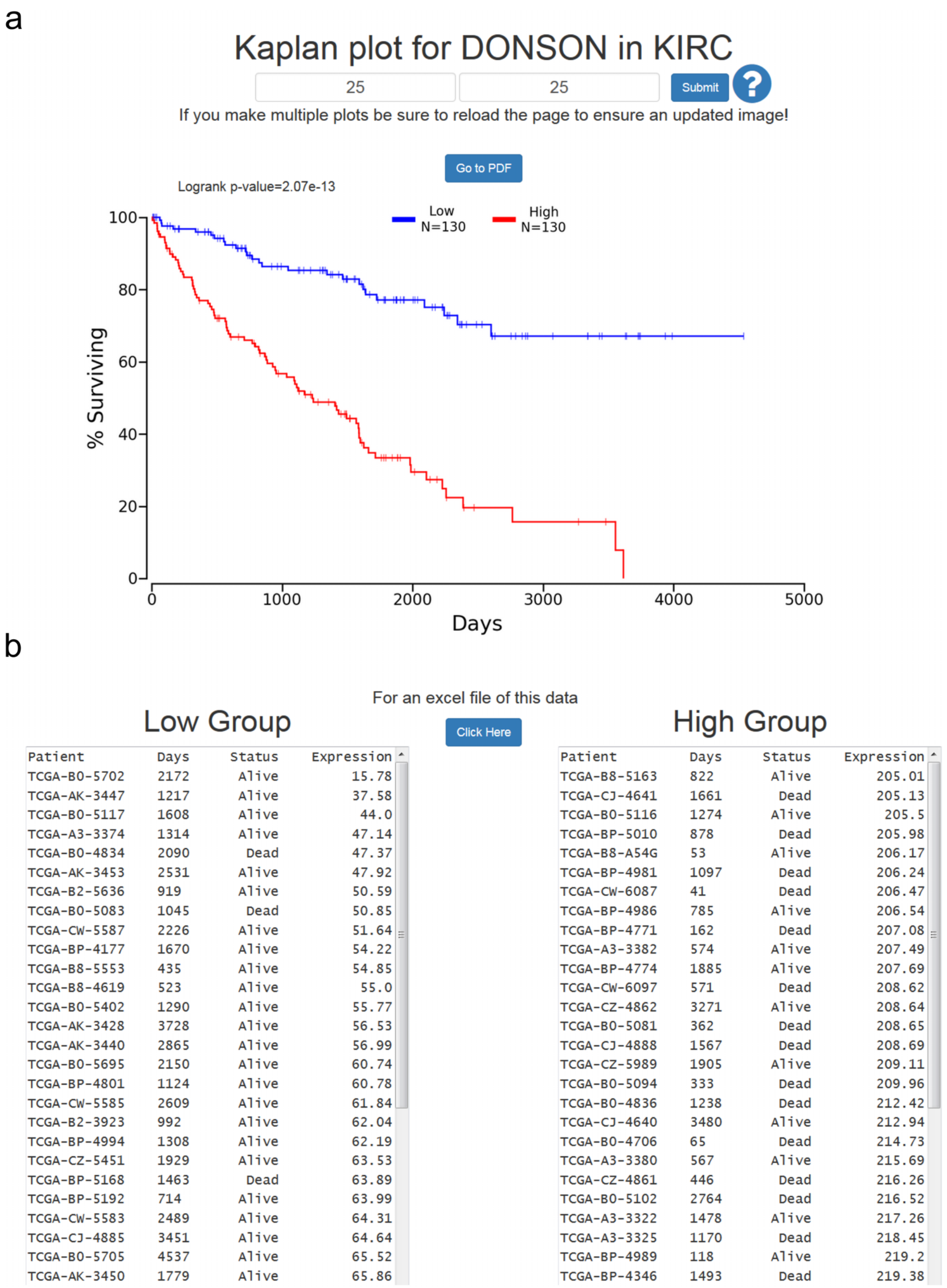
Figure 1 **Example of OncoLnc search results.** The Cox coefficient and p-value are from the gene term in precomputed multivariate Cox regressions. The FDR correction is performed per cancer analysis per data type, and in this example the correction would have involved around 16,000 genes for each cancer. The rank is also performed per cancer per data type. In this example DONSON is the 3rd most highly correlated gene in KIRC.

If a user simply wants all the data for that cancer and that gene, the user can submit 100 for "Lower Percentile," and 0 for "Upper Percentile."

Users then have the option to either go to a PDF of the Kaplan-Meier plot, or download a CSV file of the data plotted. In both cases the file name will be the cancer, gene ID, lower percentile, upper percentile, separated by underscores. Gene ID had to be used instead of gene name because there are multiple HUGO gene symbol conflicts between TCGA Tier 3 mRNAs and MiTranscriptome beta, as well as between TCGA mRNA HUGO gene symbols and updated mRNA HUGO gene symbols. In the case that a user performs a search for a name with a conflict, OncoLnc presents a warning message and instructs the user how to proceed.

mRNAs

Table 1 contains information about the patients for each Tier 3 mRNA study included in OncoLnc, and how many gene analyses are present in OncoLnc for each study. Tier 3 RNASeqV2 was used for all 21 cancers, and expression was taken from the "rsem.genes.normalized_results" files. As a result, the expression data in OncoLnc for



If you make multiple plots be sure to reload the page to ensure an updated image!

For an excel file of this data

Low Group

Patient	Days	Status	Expression
TCGA-B0-5702	2172	Alive	15.78
TCGA-AK-3447	1217	Alive	37.58
TCGA-B0-5117	1608	Alive	44.0
TCGA-A3-3374	1314	Alive	47.14
TCGA-B0-4834	2090	Dead	47.37
TCGA-AK-3453	2531	Alive	47.92
TCGA-B2-5636	919	Alive	50.59
TCGA-B0-5083	1045	Dead	50.85
TCGA-CW-5587	2226	Alive	51.64
TCGA-BP-4177	1670	Alive	54.22
TCGA-B8-5553	435	Alive	54.85
TCGA-B8-4619	523	Alive	55.0
TCGA-B0-5402	1290	Alive	55.77
TCGA-AK-3428	3728	Alive	56.53
TCGA-AK-3440	2865	Alive	56.99
TCGA-B0-5695	2150	Alive	60.74
TCGA-BP-4801	1124	Alive	60.78
TCGA-CW-5585	2609	Alive	61.84
TCGA-B2-3923	992	Alive	62.04
TCGA-BP-4994	1308	Alive	62.19
TCGA-CZ-5451	1929	Alive	63.53
TCGA-BP-5168	1463	Dead	63.89
TCGA-BP-5192	714	Alive	63.99
TCGA-CW-5583	2489	Alive	64.31
TCGA-CJ-4885	3451	Alive	64.64
TCGA-B0-5705	4537	Alive	65.52
TCGA-AK-3450	1779	Alive	65.86

High Group

Patient	Days	Status	Expression
TCGA-B8-5163	822	Alive	205.01
TCGA-CJ-4641	1661	Dead	205.13
TCGA-B0-5116	1274	Alive	205.5
TCGA-BP-5010	878	Dead	205.98
TCGA-B8-A54G	53	Alive	206.17
TCGA-BP-4981	1097	Dead	206.24
TCGA-CW-6087	41	Dead	206.47
TCGA-BP-4986	785	Alive	206.54
TCGA-BP-4771	162	Dead	207.08
TCGA-A3-3382	574	Alive	207.49
TCGA-BP-4774	1885	Alive	207.69
TCGA-CW-6097	571	Dead	208.62
TCGA-CZ-4862	3271	Alive	208.64
TCGA-B0-5081	362	Dead	208.65
TCGA-CJ-4888	1567	Dead	208.69
TCGA-CZ-5989	1905	Alive	209.11
TCGA-B0-5094	333	Dead	209.96
TCGA-B0-4836	1238	Dead	212.42
TCGA-CJ-4640	3480	Alive	212.94
TCGA-B0-4706	65	Dead	214.73
TCGA-A3-3380	567	Alive	215.69
TCGA-CZ-4861	446	Dead	216.26
TCGA-B0-5102	2764	Dead	216.52
TCGA-A3-3322	1478	Alive	217.26
TCGA-A3-3325	1170	Dead	218.45
TCGA-BP-4989	118	Alive	219.2
TCGA-BP-4346	1493	Dead	219.38

Figure 2 **Example of OncoLnc Kaplan-Meier results.** (A) Submitting non-overlapping percentiles will return a logrank p-value for the analysis and a PNG image with the option to generate a PDF of the plot. (B) Below the Kaplan-Meier image will be the data that was plotted along with an option to download a CSV file.

Tier 3 mRNAs is in normalized RSEM values. Table 1 contains different numbers of genes for the different cancers because an expression cutoff was used to determine if a gene would be included in the analysis. For mRNAs this cutoff was a median expression greater than 1 RSEM, and less than a fourth of the patients with an expression of 0.

The results of every Tier 3 mRNA Cox regression performed are included in Table S1. The Tier 3 expression files contain both a HUGO gene symbol and Entrez Gene ID for each gene, but these IDs and gene symbols are not current. To update the gene symbols I downloaded every human gene from NCBI Gene, and updated any symbol for which the Entrez Gene ID was still current. For genes that had deleted or changed Entrez Gene IDs I had to manually curate the Gene IDs and gene symbols. Genes which I could not confidently assign to a modern ID are not included in OncoLnc, but are still included in Table S1. Table S1 includes the original TCGA IDs and symbols along with the updated names and symbols, and Table S4 lists genes which had either the symbol or ID changed. OncoLnc allows users to search mRNAs using either an updated HUGO gene symbol or Entrez Gene ID.

miRNAs

Table 2 contains information about the patients for each Tier 3 miRNA study included in OncoLnc, and how many gene analyses are present in OncoLnc for each study. Tier 3 miRNASeq was used for every cancer except GBM, which only had microarray data available. The results of every Cox regression performed are included in Table S2. Many of the miRBase IDs, or possibly read counts, present in Table S2 and OncoLnc will be different from the IDs and read counts in TCGA data files and available at other data portals for TCGA data. This is because I went through each expression file and updated the IDs and read counts.

The "isoform.quantification" files contain both miRBase IDs as well accession numbers. In these files the 5p and 3p arms of miRNAs are referred to with the same ID, for example hsa-let-7b-5p and hsa-let-7b-3p would both be listed as hsa-let-7b. In order to update the names and read counts for the Tier 3 miRNAs I used the read counts assigned to each accession number to obtain reads per million miRNAs mapped for each accession number, and updated the ID with the current miRBase ID. When an accession number was not available I used the genomic coordinates provided to identify the accession number, and therefore ID. GBM names were updated using the "aliases" file from the miRBase FTP site, and if an alias could not be confidently identified the miRNA was not included in OncoLnc, but is still in Table S2.

As a result, all expression values in Table S2 and in OncoLnc are reads per million miRNA mapped for every cancer except GBM, which are microarray normalized values. The numbers of miRNAs in Table 2 differ because the miRNA may not have been in the expression files for that cancer, or may not have met the expression cutoff. An expression cutoff of a median of .5 reads per million miRNA mapped, and less than one fourth of the patients with 0 expression was used. OncoLnc allows users to search for miRNAs with either a miRBase version 21 mature accession number or ID.

lncRNAs

Table 3 contains information about the patients for each MiTranscriptome beta lncRNA analysis, along with how many lncRNAs are included in OncoLnc for each cancer. Normalized lncRNA counts were downloaded from http://mitranscriptome.org/, and these were mapped to patient barcodes using the library information provided. MiTranscriptome beta contains over 8,000 of the most differentially expressed lncRNAs in the entire MiTranscriptome dataset, but the actual number of lncRNAs in OncoLnc for each cancer is far fewer due to the expression cutoff used: a median of .1 normalized counts, and less than a fourth of patients with 0 expression. Table S3 contains every lncRNA Cox regression performed, and these are all included in OncoLnc. OncoLnc allows users to search for MiTranscriptome beta lncRNAs using either a name or transcript ID.

DISCUSSION

Depending on the researcher, OncoLnc should be used in different ways. If a researcher is studying a specific gene and looking for a cancer association, they should go to http://www.oncolnc.org and perform a search with their gene of interest. Instead of focusing on p-values, I would focus more on the rank of the correlations for the different cancers, and also on the sign of the Cox coefficients. A positive Cox coefficient indicates high expression of the gene increases the risk of death, while a negative Cox coefficient indicates the opposite. A gene with a high rank in multiple cancers (indicated by a low number, 1 being the best), and Cox coefficients with the same sign could be very interesting. It is also important to look at the level of expression of the gene. Different genes obviously require different levels of expression to exert their effects, but genes with expression near 0 should be dealt with caution. In addition, users can investigate the range of expression of the gene at the Kaplan-Meier plotting page. Genes that have large fold increases from the low expression to high expression group could be interesting candidates.

A researcher studying a specific cancer should download Tables S1–S3 to see which mRNAs, miRNAs, and lncRNAs are most correlated to survival for their cancer. Once they identify some genes of interest they can go to http://www.oncolnc.org to perform further analyses such as checking the range of expression of the gene, or if it is associated with survival in other cancers. Similarly, bioinformaticians looking to perform large scale analyses of prognostic genes can use these tables as a starting point, or if a user wants to change the Cox models they can use the GitHub code to alter the models.

The importance of the ability to perform survival correlations with lncRNAs must be emphasized. There are multiple techniques for identifying protein coding genes that are involved in cancer because mutations that occur in protein coding genes can result in missense mutations, and methods have been developed for identifying which of these mutations are drivers as opposed to simply passengers (*Carter et al., 2009*; *Kaminker et al., 2007*; *Youn & Simon, 2011*). In contrast, because it is unclear how mutations will affect lncRNA function, methods to identify lncRNAs involved in cancer must rely on lncRNA expression. As a result, OncoLnc is one of the few resources available for finding lncRNAs involved in cancer, and if a lncRNA researcher is searching for a novel lncRNA to study, Table S3 would be a good place to start.

Compared to microarrays, RNA-SEQ is a relatively unproven technology for identifying prognostic genes. However, I previously performed a pan-cancer analysis of prognostic genes that used RNA-SEQ data and provided consistent groupings of cancers and identified interesting cancer biology (*Anaya et al., 2016*). For example, in my analysis EGF signaling was strongly associated with survival in LUSC, and EGFR inhibitors have been shown to be effective in LUSC patients despite EGFR mutations being rare in this cancer (*Chiu et al., 2014*). Although examples such as this are assuring, when using OncoLnc it is important to remember that the correlations observed, regardless of p-value, are still only correlations. Perhaps the largest limitation of OncoLnc is that the Cox models do not account for intra-cancer subtypes. For example, GBM and BRCA both have well-established subtypes (*Brennan et al., 2013*; *Perou et al., 2000*). If the expression of a gene correlates with cancer subtypes, and those subtypes correlate with survival, subtype would be a confounding variable. As subtype definitions for the different cancers improve a future version of OncoLnc may be able to incorporate the subtypes in the Cox models.

An analysis is only as good as the data available, and the Tier 3 TCGA RNA-SEQ analyses were performed with outdated software and transcript information. There have been some attempts to reanalyze both the TCGA mRNA RNA-SEQ data and miRNA-SEQ data (*Kuo et al., 2015*; *Rahman et al., 2015*). In the event that TCGA or the scientific community releases a gold standard analysis of TCGA data, a future version of OncoLnc could incorporate this data.

Current data portals for TCGA data only allow users to view the results for one cancer at a time, may or may not offer Cox regression results, do not allow for complete control over separating patients during Kaplan-Meier analysis, and do not allow for download of the data used in the analysis. To my knowledge OncoLnc is the only online resource for TCGA data that includes these features, is the only resource that uses modern gene definitions for TCGA mRNA and miRNA data, and is the only resource for survival analysis of MiTranscriptome beta lncRNAs. In addition, current methods for survival analysis rely on a p-value cutoff of .05 for significance, which may lead to either the study of genes not actually correlated with survival or missing genes that are correlated with survival depending on the cancer. By storing the results of the correlation for every gene, OncoLnc can provide a context for the significance of a correlation. As a result, used correctly OncoLnc can not only increase the sensitivity of finding genes involved in cancer, but also the specificity. This combination of ease of use, results for complex analyses, and tools for exploring and downloading data make OncoLnc an invaluable resource for cancer researchers.

ACKNOWLEDGEMENTS

This project was made possible by data generated by the TCGA Research Network: http://cancergenome.nih.gov/.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The author received no funding for this work.

Competing Interests

Jordan Anaya offers services at omnesres.com which can involve web development or analysis of TCGA data.

Author Contributions

- Jordan Anaya conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, performed the computation work, reviewed drafts of the paper, developed OncoLnc, maintains the GitHub repository and information at omnesres.com.

Data Deposition

The following information was supplied regarding data availability:

GitHub: https://github.com/OmnesRes/onco_lnc

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.67#supplemental-information.

REFERENCES

Anaya J, Reon B, Chen W-M, Bekiranov S, Dutta A. 2016. A pan-cancer analysis of prognostic genes. *PeerJ* **3**:e1499 DOI 10.7717/peerj.1499.

Brennan CW, Verhaak RG, McKenna A, Campos B, Noushmehr H, Salama SR, Zheng S, Chakravarty D, Sanborn JZ, Berman SH, Beroukhim R, Bernard B, Wu C-J, Genovese G, Shmulevich I, Barnholtz-Sloan J, Zou L, Vegesna R, Shukla SA, Ciriello G, Yung WK, Zhang W, Sougnez C, Mikkelsen T, Aldape K, Bigner DD, Van Meir EG, Prados M, Sloan A, Black KL, Eschbacher J, Finocchiaro G, Friedman W, Andrews DW, Guha A, Iacocca M, O'Neill BP, Foltz G, Myers J, Weisenberger DJ, Penny R, Kucherlapati R, Perou CM, Hayes DN, Gibbs R, Marra M, Mills GB, Lander E, Spellman P, Wilson R, Sander C, Weinstein J, Meyerson M, Gabriel S, Laird PW, Haussler D, Getz G, Chin L, Network TR. 2013. The somatic genomic landscape of glioblastoma. *Cell* **155(2)**:462–477 DOI 10.1016/j.cell.2013.09.034.

Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, Vogelstein B, Karchin R. 2009. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Research* **69(16)**:6660–6667 DOI 10.1158/0008-5472.can-09-1133.

Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, Antipin Y, Reva B, Goldberg AP, Sander C, Schultz N. 2012. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discovery* **2(5)**:401–404 DOI 10.1158/2159-8290.cd-12-0095.

Chiu C-H, Chou T-Y, Chiang C-L, Tsai C-M. 2014. Should EGFR mutations be tested in advanced lung squamous cell carcinomas to guide frontline treatment? *Cancer Chemotherapy and Pharmacology* **74(4):**661–665 DOI 10.1007/s00280-014-2536-3.

Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. 2013. Emerging landscape of oncogenic signatures across human cancers. *Nature Genetics* **45(10):**1127–1133 DOI 10.1038/ng.2762.

Győrffy B, Surowiak P, Budczies J, Lánczky A. 2013. Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung cancer. *PLoS ONE* **8(12):**e82241 DOI 10.1371/journal.pone.0082241.

Han L, Yuan Y, Zheng S, Yang Y, Li J, Edgerton ME, Diao L, Xu Y, Verhaak RGW, Liang H. 2014. The Pan-Cancer analysis of pseudogene expression reveals biologically and clinically relevant tumour subtypes. *Nature Communications* **5:**3963 DOI 10.1038/ncomms4963.

Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, Leiserson MDM, Niu B, McLellan MD, Uzunangelov V, Zhang J, Kandoth C, Akbani R, Shen H, Omberg L, Chu A, Margolin AA, Van't Veer LJ, Lopez-Bigas N, Laird PW, Raphael BJ, Ding L, Robertson AG, Byers LA, Mills GB, Weinstein JN, Van Waes C, Chen Z, Collisson EA, Cancer Genome Atlas Research N, Benz CC, Perou CM, Stuart JM. 2014. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158(4):**929–944 DOI 10.1016/j.cell.2014.06.049.

Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, Barrette TR, Prensner JR, Evans JR, Zhao S, Poliakov A, Cao X, Dhanasekaran SM, Wu Y-M, Robinson DR, Beer DG, Feng FY, Iyer HK, Chinnaiyan AM. 2015. The landscape of long noncoding RNAs in the human transcriptome. *Nature Genetics* **47(3):**199–208 DOI 10.1038/ng.3192.

Kaminker JS, Zhang Y, Watanabe C, Zhang Z. 2007. CanPredict: a computational tool for predicting cancer-associated missense mutations. *Nucleic Acids Research* **35(suppl 2):**W595–W598 DOI 10.1093/nar/gkm405.

Koch A, De Meyer T, Jeschke J, Van Criekinge W. 2015. MEXPRESS: visualizing expression, DNA methylation and clinical TCGA data. *BMC Genomics* **16(1):**636 DOI 10.1186/s12864-015-1847-z.

Kuo W-T, Su M-W, Lee YL, Chen C-H, Wu C-W, Fang W-L, Huang K-H, Lin W-C. 2015. Bioinformatic interrogation of 5p-arm and 3p-arm specific miRNA expression using TCGA datasets. *Journal of Clinical Medicine* **4(9):**1798–1814 DOI 10.3390/jcm4091798.

Li J, Han L, Roebuck P, Diao L, Liu L, Yuan Y, Weinstein JN, Liang H. 2015. TANRIC: an interactive open platform to explore the function of lncRNAs in cancer. *Cancer Research* **75(18):**3728–3737 DOI 10.1158/0008-5472.can-15-0273.

Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge Ø, Pergamenschikov A, Williams C, Zhu SX, Lønning PE, Børresen-Dale AL, Brown PO, Botstein D. 2000. Molecular portraits of human breast tumours. *Nature* **406(6797):**747–752 DOI 10.1038/35021093.

Rahman M, Jackson LK, Johnson WE, Li DY, Bild AH, Piccolo SR. 2015. Alternative preprocessing of RNA-Sequencing data in The Cancer Genome Atlas leads to improved analysis results. *Bioinformatics* **31(22):**3666–3672 DOI 10.1093/bioinformatics/btv377.

Yang Y, Han L, Yuan Y, Li J, Hei N, Liang H. 2014. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nature Communications* **5**:3231 DOI 10.1038/ncomms4231.

Yarmishyn AA, Kurochkin IV. 2015. Long noncoding RNAs: a potential novel class of cancer biomarkers. *Front Genet* **6**:145 DOI 10.3389/fgene.2015.00145.

Youn A, Simon R. 2011. Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics* **27(2):**175–181 DOI 10.1093/bioinformatics/btq630.