 doi:10.20944/preprints201912.0183.v1

Investigating strength that comprehensive mRNA expression level of prognostic genes influences on patient survival for every cancer type

Minhyeong Lee, PhD candidate

minhyeong1022@gmail.com

Cancer Research Institute, Seoul National University College of Medicine

101, Daehak-ro, Jongno-gu

Seoul 03080, Republic of Korea

Abstract

Purpose

This study aimed to rank cancers by the strength of relationship between comprehensive mRNA expression of the most harmful or protective genes and patient survival.

Materials and Methods

Using TCGA dataset including RNA-SEQ and clinical data, we investigated not only gene specific prognostic availability, but also comprehensive prognostic availability of prognostic genes filtered by cox coefficient, and ranked cancers by specially designed prognostic indicator.

Results

Through Kaplan-Meier plots, we checked that cancers vary in the strength of influence of prognostic genes, and they follow as the rank.

Conclusion

Developing treatment with method to reduce or increase expression of biomarkers for specific cancer which ranked bottom, it would be not efficient in high probability. The results of this study can be a scientific evidence for that.

Keywords

Bioinformatics, Cancer, Genomics, Computational Biology, RNA sequencing, TCGA

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 December 2019 doi:10.20944/preprints201912.0183.v1

Introduction

In order to perform the most comprehensive research possible, we used TCGA dataset including clinical information, 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 [1-11].

There are various approaches to cure cancers such as surgery, radiotherapy, immunotherapy or chemotherapy. Usually, most of chemotherapeutic methods have been the way to reduce or increase expression of protein or gene related to patient survival. However, while in some cancers, many biomarkers are efficient in treatment, in some other cancers, they are not. This research produces advantage of availability to provide scientific evidence of which cancers, comparing to another cancers, have significant relationship between clinical outcome and comprehensive mRNA expression level of prognostic genes. For the method, we formulated an indicator called prognostic score. The results of this study can be a kind of guide how to approach by cancers to develop cure for them.

Materials & Methods

1. Code and files

All of files and R codes generated for this study, including figures and tables can be downloaded from author's github page https://github.com/Minhyeong-1022/TCGA_mRNA-expression_survival_correlation_research. The scripts were run on a Samsung laptop 9 with i5 processor and 8GB of RAM running R 3.5.1 on Windows 10.

2. Construction of prognostic indicator model

To pull RNA-SEQ and clinical data into R source code, we used getFirehoseData function from R library named RTCGAToolbox, which pulls TCGA data from https://tcga-data.nci.nih.gov/tcga/, in September 2019. All of clinical follow up information were extracted from "clinical" files for each cancer, and only patient information whose clinical events occurred were included in this analysis. TCGA provides two types of mRNA expression data which are RNA-Seq and RNA-SeqV2. We used RNA-SeqV2 because it is the most recent version. "RNASeq2GeneNorm" files formed in RSEM which is one of mRNA expression reporting method and were outputted by the RSEM software [12]. For each cancer, only gene which has expression value greater than 1 included in this model. The prognostic score is determined by R2 and ratio of genes which have p-value of less than or equal to 0.01 generated from each term of cox model. R2 values are the Pearson correlation coefficient of R1 and cox coefficient. They were generated for each cancer. R1 values are Pearson correlation coefficient of patient specific survival days to death and mRNA expression values (RSEM) for each gene. They were generated for every gene. The information of TCGA dataset including number of genes selected by p-value generated from cox model are represented on Table.1. Prognostic scores were generated for each cancer. The comprehensive prognostic indicator formula is represented below.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 December 2019 doi:10.20944/preprints201912.0183.v1

$$R_1 = Pearson\ cor(mRNA\ expression\ values, Days\ to\ death) \rightarrow For\ each\ gene$$

$$R_2 = Pearson\ cor(Cox\ coefficient, R_1) \rightarrow For\ each\ cancer$$

$$Prognostic\ score = \frac{R_2(Number\ of\ selected\ genes)}{(Number\ of\ total\ genes)}$$

3. Information of multivariate cox model used in analysis

Gene specific cox coefficient and p-value for 16 cancers were generated in previous study, published in February 2016, by Jordan Anaya at University of Virginia [13,14]. We used the data file "All genes, p-values, cox coefficients for each cancer", and it can be downloaded via https://peerj.com/articles/1499/#supp-1 .

Results

We decided to analyze using TCGA data which had enough number of patients with RNA-SEQ data and mature clinical follow up information. Total of 14 cancers were used in this study which are acute myeloid leukemia (LAML), bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), 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), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), kidney renal papillary cell carcinoma (KIRP), liver hepatocellular carcinoma (LIHC), skin cutaneous melanoma (SKCM), and stomach adenocarcinoma (STAD).

This study has purpose on ranking cancers by the sensitivities to prognostic gene expression level using specially formulated indicator called prognostic score. To achieve this multivariate Cox proportions hazards model [15], which is a standard regression method to research clinical data [16-19] was needed. Data table in previous study [13,14] including gene specific cox coefficient and p-value for 16 cancers was used to minimize the efforts.

In few earlier studies, cox coefficient and p-value provided from each term of cox model were used to analyze how much mRNA expression of each gene influence on patient survival. There are significant deviations of patient survival patterns by expression level of genes which have extreme cox coefficient, and we can check this through Kaplan-Meier plot for all cancers (Fig.1d-g). Patient survival in some cancer types are changed significantly by their gene expression level, but the patient survival in few another cancer type is not. It is because cancers vary in sensitivities to gene expression level. We calculated prognostic scores for all cancer types used in this study. Although the main factor of prognostic indicator is R2, not only R2 but also ratio of selected genes influences on patient survival. Cancers differ in prognostic score, and a bar-chart showing the rank displayed on Fig1. Expression levels of prognostic genes in top ranked cancer types are more influential on the patient clinical outcomes in those cancers,

 doi:10.20944/preprints201912.0183.v1

but such influences of bottom ranked are less than top ranked.

In the case of KIRC which has highest prognostic score among the 14 cancers, squared regression (R2) (Fig.1b) is more significant than R2 of OV (Fig.1c) which located bottom in the rank. KIRC and OV markedly differ in patient survival patterns by gene expression level. As following Fig1, KIRC is very sensitive to gene expression, but OV is not. And also, in the cases of another cancers, the prognostic sensitivities to gene expression level follow as the rank (Fig.1a). Supplemental materials including Kaplan-Meier plots and scatter plots for other cancers can be downloaded via Git-hub page.

Discussion

In almost of cancers, mRNA expression level of protective or harmful genes filtered by cox coefficient influence on patient survival. But the performances differ by cancers. To rank cancers by how much prognostic gene expression levels influence on clinical outcome, specially formulated indicator called prognostic score was used in this study. In conclusion, it successfully predicted performances as the ranking. When researchers who develop medicine for cancers consider the cancer sensitivities to gene expression level, it can be a scientific evidence for their considering. For example, according to Fig1, while developing treatment for OV with chemotherapeutic method: reducing or increasing mRNA and protein level is not going to be efficient in high probability, research for KIRC is going to be efficient.

References

1. Brennan CW, Verhaak RG, McKenna A, Campos B, Noushmehr H, Salama SR, et al. The somatic genomic landscape of glioblastoma, 2013. Cell. 155:462-477 doi:10.1016/j.cell.2013.09.034

2. Muzny, D., Bainbridge, N., Chang, K., Dinh, H., Drummond, J., Fowler, G., et al. Comprehensive molecular characterization of human colon and rectal cancer, 2012. Nature. 487:330-337 doi:10.1038/nature11252

3. Koboldt, D., Fulton, R., McLellan, M., Schmidt, H., Kalicki-Veizer, J., McMichael, J., et al. Comprehensive molecular portraits of human breast tumours, 2012. Nature 490:61–70 doi:10.1038/nature11412

4. Bell, D., Berchuck, A., Birrer, M., Chien, J., Cramer, D., Dao, F., et al. Integrated genomic analyses of ovarian carcinoma, 2011. Nature. 474:609–615 doi:10.1038/nature10166

5. Hammerman, P., Lawrence, M., Voet, D., Jing, R., Cibulskis, K., Sivachenko, A., et al. Comprehensive genomic characterization of squamous cell lung cancers, 2012. Nature. 489:519–525 doi:10.1038/nature11404

6. Creighton, C., Morgan, M., Gunaratne, P., Wheeler, D., Gibbs, R., Muzny, D., et al.

Comprehensive molecular characterization of clear cell renal cell carcinoma, 2013. Nature. 499:43–49 doi:10.1038/nature12222

7. Ley, TJ., Miller, C., Ding, L., Raphael, BJ., Mungall, AJ., Robertson, AG., et al. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia, 2013. New England Journal of Medicine. 368:2059-2074 doi:10.1056/NEJMoa1301689

8. Bass, A., Thorsson, V., Shmulevich, I., Reynolds, S., Miller, M., Bernard, B., et al. Comprehensive molecular characterization of gastric adenocarcinoma, 2014. Nature. 513:202–209 doi:10.1038/nature13480

9. Weinstein, J., Akbani, R., Broom, B., Wang, W., Verhaak, R., McConkey, D., et al. Comprehensive molecular characterization of urothelial bladder carcinoma, 2014. Nature. 507:315–322 doi:10.1038/nature12965

10. Collisson, E., Campbell, J., Brooks, A., Berger, A., William, L., Chmielecki, J., et al. Comprehensive molecular profiling of lung adenocarcinoma, 2014. Nature. 511:543–550 doi:10.1038/nature13385

11. Agrawal, N., Akbani, R., Aksoy, B., Ally, A., Arachchi, H, Asa, S., et al. Integrated genomic characterization of papillary thyroid carcinoma, 2014. Cell. 159:676-690 doi:10.1016/j.cell.2014.09.050

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

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

14. Anaya J. OncoLnc: linking TCGA survival data to mRNAs, miRNAs, and lncRNAs, 2016. PeerJ Computer Science. 2:e67 doi:10.7717/peerj-cs.67

15. Cox, DR. Regression models and life-tables, 1972. Journal of the Royal Statistical Society: Series B (Methodological). 34:187-220 doi:10.1111/j.2517-6161.1972.tb00899.x

16. Claus, EB., Walsh, KM., Wiencke, JK., Molinaro, AM., Wiemels, JL., Schildkraut, JM. Survival and low-grade glioma: the emergence of genetic information, 2015. Neurosurgical Focus. 38(1):E6 doi:10.3171/2014.10.FOCUS12367

17. 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, 2013. PLoS ONE. 8:e82241 doi:10.1371/journal.pone.0082241

18. Wu, G., Stein, L. A network module-based method for identifying cancer prognostic signatures, 2012. Genome Biol. 13:R112 doi:10.1186/gb-2012-13-12-r112

19. Zhang, W., Ota, T., Shridhar, V., Chien, J., Wu, B., Kuang, R. Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment, 2013. PLoS Computational Biology. 9:e1002975 doi:10.1371/journal.pcbi.1002975

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 December 2019 doi:10.20944/preprints201912.0183.v1

Cancer type	Number of patients	Number of total genes	Number of selected genes (Raw P-value < 0.01)
OV	591	16923	115
BRCA	1097	16632	695
COAD	458	16408	963
BLCA	412	16367	1557
KIRP	291	16429	3305
CESC	307	16359	951
LUAD	522	16777	2028
STAD	443	16914	852
HNSC	528	16642	832
LUSC	504	16972	180
LAML	200	15251	599
SKCM	470	16058	2400
LIHC	377	15850	1766
KIRC	537	16665	5447

Table1: Characteristics of TCGA dataset and investigated numbers of selected genes

For all of 14 cancers used in this study, the numbers of total genes and the numbers of genes selected by raw p-value of less than or equal to 0.01 were represented on the table above. The ratio of these two would be factor of prognostic indicator.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 December 2019 doi:10.20944/preprints201912.0183.v1

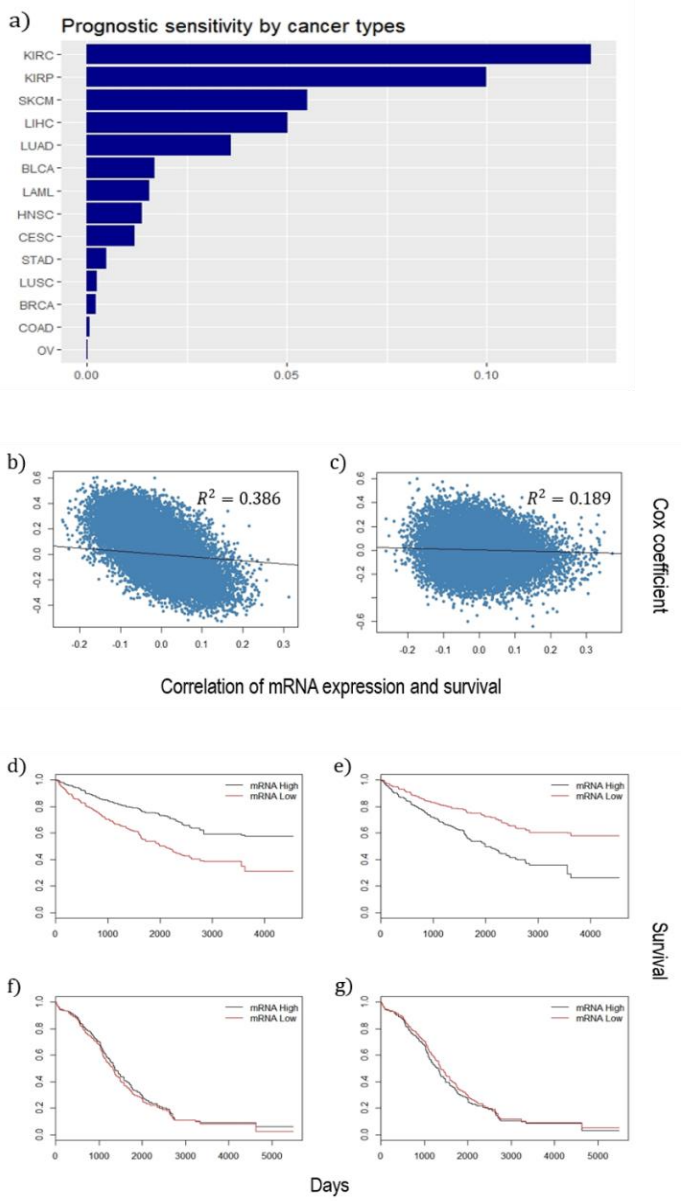

Figure 1: Distinct difference of survival patterns by prognostic score

(a) Bar-chart shows the rank of cancers by prognostic score. Prognostic score was mostly determined by R2 which is the squared regression of cox coefficient and the Pearson correlation coefficient(R1) of patient-specific survival days to death and mRNA expression for each gene. (b) Scatter plot related to R2 of KIRC which ranked top in bar-chart and (c) R2 of OV which ranked bottom in bar-chart. (d), (e), (f), (g) Kaplan-Meier plots comparing survival days of KIRC to survival days of OV patients for the 100 most protective genes and for the 100 most harmful genes.