Combination the Gene Expression of Adjacent Normal and Tumor Makes more Robust Gene Signature as Cancer Prognosis Biomarker

Wei Ma (langmawei@bjmu.edu.cn)
Chinese PLA General Hospital https://orcid.org/0000-0003-1044-5612
Dandan Li
Sun Yat-Sen University 2nd Affiliated Hospital: Sun Yat-Sen Memorial Hospital
Changjian Zhang
Sixth Medical Center of PLA General Hospital
Ming Xiong
Sixth Medical Center of PLA General Hospital
Yuanyuan Qiao
Sixth Medical Center of PLA General Hospital

Research article

Keywords: Prognosis biomarker, Gene signature, Cancer

Posted Date: December 22nd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-132449/v1

Abstract

Purpose: We tried to explore new gene signature via the combination of tumor-derived expression profile and the adjacent normal-derived expression profile to find more robust cancer biomarker.

Methods: Log2 transformed ratio of tumor tissue and the adjacent normal tissue (Log2TN) expression, tumor-derived expression, and normal-derived expression were used to do univariate Cox regression in The Cancer Genome Atlas (TCGA) lung squamous cell carcinoma (LUSC) respectively. Then, we used factor analysis and least absolute shrinkage and selection operator Cox (LASSO-Cox) to select gene signature in TCGA LUSC for Log2TN, tumor, and adjacent normal respectively.

Results: By comparing Log2TN with tumor and adjacent normal in LUSC, we found that genes derived from Log2TN show more robust ($p = 0.006$ and $p = 0.001$) and have lower p-values ($p < 0.001$). Gene signature selected from Log2TN shows the best generalization in the three GEO datasets even though only tumor-derived expression profiles were available in the three datasets. Enrichment analysis showed that the tumor cells mainly focus on proliferation with losing functional of metabolism.

Conclusions: These results indicate that (1) Log2TN could get more robust genes and gene signature than tumor-derived expression profiles used traditionally; (2) the adjacent-normal tissue may also play an important role in the progress and outcome of the tumor.

Implications for Cancer Survivors: By combined of tumor-derived expression profile and the adjacent normal-derived expression profile, we could find more robust gene signature than traditionally method. Using these robust gene signatures, robust cancer biomarkers could be constructed and will do great help to improve cancer prognosis.

Introduction

Cancer is the second leading cause of disease-related death worldwide. In 2018, about 9.6 million people died because of cancer according to the GLOBOCAN 2018 estimates of cancer incidence and mortality produced by the International Agency for Research on Cancer[1]. Lung cancer is the most commonly diagnosed cancer and leading cause of cancer death[1, 2]. One reason for the high mortality of cancer is that when cancer is diagnosed, it has been the advanced stages[3, 4]. Another reason is recurrence and metastasis after drug-treatment or surgery, especially in patients with advanced tumor stages[5–7]. As time goes on, nowadays, cancer is no longer a type of aged disease, people who get cancer become younger and younger[1]. Early diagnosis, choice of appropriate therapeutic strategies, and efficient monitoring can have a pivotal role in reducing cancer-related mortalities. Diagnosis biomarker and prognosis biomarker could help to diagnose cancer in the early stage and predict progression in the future, thus patients could get appropriate therapy in the beginning.

Nowadays, a lot of biomarkers were developed for types of cancer. Before the high-throughput sequencing developed, it was majorly single molecule-based biomarkers. For example, prostate-specific

antigen (PSA) has somewhat revolutionized the assessment of prostate cancer in 1987[8]. In 1998, it was found that %fPSA (free PSA) was an independent predictor of prostate cancer[9]. Prostate cancer antigen 3 (PCA3) was found in 1999 by Bussemakers *et al* and promoted the diagnosis of prostate cancer[10]. Then, it was found that PCA3 was also a prognosis biomarker for prostate cancer[11]. Glypican-1 (GPC-1) is a member of heparan sulfate proteoglycans. A lot of researches had found that GPC-1 is a prognosis biomarker in many cancers such as pancreatic cancer, colorectal cancer, and prostate cancer[12]. Human epidermal growth factor receptor 2 (HER2) is a diagnosis and prognosis biomarker for many cancers, especially for breast cancer[13]. With the development of high-throughput sequencing technology, multiple gene expression-based prognosis signatures were developed. MammaPrint was a 70-gene-based signature used for prediction recurrence risk of breast cancer[14]. MammaPrint is the first successful prognostic signature that was marketed by Agendia (the Netherlands). It was developed in 2002 and approved by US food and drug administration (FDA) in 2007. PAM50 is another gene-based signature for prediction recurrence risk of breast cancer, also approved by the FDA[15]. Yang et al. developed a 28 hypoxia-related gene-based recurrence prognosis signature for prostate cancer[16]. In addition to these gene expression-based signatures, other signatures based on micro RNA (miRNA), long non-coding RNA (lncRNA), and gene methylation were also very popular. For example, a combination of miR-331 and miR-21 could be a diagnostic and prognostic signature for gastric cancer[17]. Three lncRNAs including AI364715, GACAT1, and GACAT2 could be constructed as a signature for diagnostic and prognostic of gastric cancer[18]. Deleted in Split hand/Split foot 1 (*DSS1*) promoter hypomethylation could predict poor prognosis in melanoma and squamous cell carcinoma patients[19].

Signatures like described above are innumerable and couldn't be list here one by one. These genes (miRNA, lncRNA) expression-based prognosis signatures were developed based on gene (miRNA, lncRNA) expression in tumor tissue. This faced a big problem: there is no stander or baseline for expression. Different methods such as an array, sequencing, and Q-PCR get different expression values in different scales. Different instruments or operators also get different results. Another problem is that extrapolation and robustness of these signatures are not enough for widely using[20, 21]. Besides, the combination of expression in tumor tissue and adjacent-normal tissue maybe could get better performance.

In this study, we explored these questions in lung squamous cell carcinoma (LUSC) on recurrence prognosis issue. Instead of using expression profiles in tumor tissue, we used the log2 transformed ratio of tumor tissue and the adjacent normal tissue (Log2TN) to explore the prognosis signature. For a single gene, compared with the tumor tissue expression profile, Log2TN based was more significant. Log2TN based signature was also more significant. Furthermore, the signature explored from Log2TN was very robust in the test dataset even if using the tumor tissue expression profile.

Materials And Method

Data collection:

To investigate the Log2TN based recurrence-related prognosis signature in LUSC, the LUSC dataset of The Cancer Genome Atlas (TCGA) was downloaded from UCSC Cancer Browser (UCSC Xena; https://xenabrowser.net/datapages). Then we searched the Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) for LUSC datasets that fulfilled the following criteria: included samples were hybridized to the Affymetrix Human Genome U133 Plus 2.0 Array (GEO accession number GPL570) platforms; information on the recurrence event was available. The expression matrix was downloaded and the mRNA expression profiles were log2 transformed for further analysis.

Genes selection and Cox regression:

Univariate Cox regression was used to investigate the recurrence prognosis related genes. Factor analysis (FA) and least absolute shrinkage and selection operator (LASSO) Cox were both used to reduce the dimensionality and to select the most significantly relapse-related genes to build a recurrence prognostic model using the multivariate Cox regression method. Survival rates were calculated by the Kaplan–Meier method, and the significance of differences between survival curves was determined using the log-rank test. Uni- and multivariate analyses were performed using Cox proportional hazard models. Cox regression and survival curves were performed through the R language "survival" package. Factor analysis was performed through the R language "psych" package. LASSO-Cox was performed through the R language "SIS" packages.

Gene set enrichment analysis:

We used gene set enrichment analysis (GSEA) to investigate the potential mechanisms in the c2 (c2.cp.kegg.v7.0.symbols) of the molecular signatures database (MSigDB) using the JAVA program (http://software.broadinstitute.org/gsea/index.jsp). The pre-ranked method was used to perform GSEA analysis. The number of random sample permutations was set at 1000, and the gene set size was set from 5 to 500. The significance threshold was set at $p<0.05$.

Statistical analysis:

All statistical analyses including Fisher's exact test and student's t-test are based on R language 3.6.2 version and attached packages. R is a free software environment for statistical computing and graphics at https://www.r-project.org/.

Result

Data characteristics:

The TCGA LUSC contains 553 samples including 502 primary tumor samples and 51 solid adjacent normal tissue samples. Among the 502 tumor samples, 386 have recurrence event information. All the 51 solid adjacent normal tissue samples have matched tumor tissue samples with 40 samples have recurrence event information. We downloaded three datasets from GEO including GSE8894, GSE30219, and GSE37745. GSE8894 contains 133 samples including 75 lung squamous cell carcinoma samples and 63 other types of lung cancer samples. All of the 75 squamous cell carcinoma samples have recurrence event information. GSE30219 contains 61 lung squamous cell carcinoma samples, 232 other types of lung cancer samples, and 14 normal tissue samples. 60 lung squamous cell carcinoma samples have recurrence event information in GSE30219. GSE37745 contains 66 squamous cell carcinoma samples and 130 other types of lung cancer samples while only 30 of the 66 squamous cell carcinoma samples have recurrence event information.

Genes explored via Log2TN are more robust and significant:

In the TCGA LUSC dataset, we selected the 40 tumor-normal paired sample with recurrence event information and calculated Log2TN for every gene. Then univariate Cox regression analysis was performed for each gene using Log2TN, expression profile of tumor tissue, and expression profile of normal tissue within these 40 individuals respectively. For the three datasets from GEO, even if only the expression profile of the tumor is available, we performed univariate Cox regression analysis for each gene using the tumor-derived expression profile in each dataset. Significant genes were picked out using $P < 0.05$. Then we investigated the overlapping between the TCGA LUSC and the other three datasets. Significant genes in TCGA LUSC dataset in anyone of the three datasets of GEO are defined as overlapped genes. The overlapping of significant genes between the TCGA LUSC dataset and the other three datasets are shown in Table 1.

Table 1
Overlapping between TCGA LUSC and three datasets from GEO

Positive coefficient					
	All	**Remaining**	**Overlapped**	**Ratio**	P-**value**
Tumor	373	345	28	0.075	T-N: $p = 0.871$
Normal	214	199	15	0.070	T-L: $p = 0.757$
Log2TN	231	213	19	0.082	N-L: $p = 0.722$
Negative coefficient					
Tumor	196	189	7	0.036	T-N: $p = 0.622$
Normal	203	193	10	0.049	T-L: $p = 0$ .012
Log2TN	840	765	75	0.089	N-L: $p = 0.064$
All					
Tumor	569	494	75	0.132	T-N: $p = 0.641$
Normal	417	365	61	0.146	T-L: $p = 1.000$
Log2TN	1071	929	142	0.133	N-L: $p = 0.616$

T-N: Tumor versus Normal; T-L: Tumor versus Log2TN; N-L: Normal versus Log2TN; *p*: Fisher's exact test

For the genes with the Cox regression coefficient below zero, although the overlapping ratio of Log2TN is higher, it is not significant (Tumor versus Log2TN, $p = 0.757$; Normal versus Log2TN, $p = 0.722$, Fisher's exact test). For the genes with the Cox regression coefficient above zero, the overlapping ratio of Log2TN is significantly higher than the tumor ($p = 0.012$, Fisher's exact test). When contrasted with normal, the p-value is on the verge of significant ($p = 0.064$, Fisher's exact test). When combined with the two parts above, there is no difference. When compare the overlapping ratio of tumor and normal, there is no difference. Though there is no huge advantage on overlapping ratio for Log2TN, considering the Cox regression coefficient sign, the Log2TN is more robust, as shown in Table 2.

Table 2
Overlapped genes of consistent and inconsistent sign

	All	**Consistent sign**	**Inconsistent sign**	***P*-value**
Tumor	75	35	40	T-N: $p = 0.602$
Normal	61	25	36	T-L: $p = 0.006$
Log2TN	142	94	48	N-L: $p = 0.001$

T-N: Tumor versus Normal; T-L: Tumor versus Log2TN; N-L: Normal versus Log2TN; *p*: Fisher's exact test

For the tumor, there are 28, 7 overlapped genes with the Cox regression coefficient above and below zero respectively. When combined with the genes with the coefficient above and below zero together, there are 75 genes, which means there are 40 genes with opposite Cox regression coefficients in TCGA LUSC and the three datasets. That means a good outcome with high (low) expression level of a gene in TCGA LUSC while poor outcome with high(low) expression level of that gene in the other three datasets. Log2TN has fewer genes of inconsistent compared with tumor and normal (Tumor versus Log2TN, $p = 0.006$; Normal versus Log2TN, $p = 0.001$, Fisher's exact test; Tumor versus Normal, $p = 0.602$). It is indicated that genes obtained via Log2TN are more robust.

Furthermore, we investigated the p-value of all genes between tumor, normal, and Log2TN. We found that *p-value* of Log2TN was lower than normal and tumor (Log2TN versus tumor, 0.495 versus 0.519, $p = 3.08 \times 10{-}15$; Log2TN versus normal 0.495 versus 0.512, $p = 4.01 \times 10 - 8$, Student's t-test). The *p-value* of normal was lower than that of the tumor (Normal versus tumor, 0.511 versus 0.519, $p = 0.013$, Student's t-test). Then we selected the top 100, 200, 400, and 800 low *p*-values from each of them and compared them using Student's t-test. As shown in supplementary Fig. 1, p-value of Log2TN was the lowest.

Cox regression model based on the consistent genes obtained via Log2TN is robust:

We acquired 35, 25, 94 consistent genes from tumor, normal and Log2TN respectively. The 40 tumor-normal paired samples in the TCGA LUSC dataset were used in the following analysis. For each of them, we selected the top 25 significant genes for further analysis. In the following analysis, we ran the same process for tumor, normal, and Log2TN using the corresponding expression value, for example, when using the 25 genes from normal, we used the expression profile from normal. First, factor analysis (FA) was performed to find out the nonredundant genes. Eigenvalues above zero was used to confirm the number of factors. Then, we selected the genes with the absolute value of factor loading above 0.5. At last, the LASSO was used to identify the prognostic genes and built the Cox regression model. For Log2TN, 10 factors were retained. In the 10 factors, there were 22 genes with the absolute value of factor loading above 0.5. Then after the LASSO selection, 14 genes (*ZNF275, PODXL2, SLCO4C1, POMZP3, ACAD11, MAPK4, BCAR3, DKKL1, FRK, TRPM3, NRP1, PAEP, KLHL13, HRSP12*) remained to build the Cox regression model. As shown in Table 3, the likelihood ratio test *p-value* of the model is 4.57e-05. For Tumor, at last, there were 12 genes (*TNFRSF10A, PYGB, OSBPL10, LRRN4, HHIP, ZBED2, CECR6, MBOAT2, ZNF45, SLC14A1, C4orf26, XRRA1*) kept for building the Cox model ($p = 0.01052$, likelihood ratio test). For normal, 11 genes (*ITGB6, UNC80, TMEM92, VPS37D, KCNB1, FHDC1, CNR1, TIGD6, SLC16A4, LIX1, ARMC9*) were left to build the Cox model ($p = 1.044e-08$, likelihood ratio test).

Table 3
Selected genes based Cox models in four datasets

	Factor number	Gene Number	TCGA LUSC	GSE8894	GSE30219	GSE37745
Log2TN	10	14	4.57×10^{-5}	0.00394	1.03×10^{-5}	0.0412
Tumor	10	12	0.0105	0.3082	0.0239	0.0528
Normal	8	11	1.04×10^{-8}	0.0171	0.0564	0.0154

Values in the last for columns were likelihood ratio test *p-value* of cox regression.

Then, we tested the selected genes in the three datasets from GEO for each of them. The selected genes were used to build the Cox regression model in each of the three datasets. The results were summarized in Table 3. Although the expression profile of tumor tissue was used to build the Cox model in the three datasets from GEO (only expression profile was obtained), we found that genes selected via Log2TN had the best performance. The Cox regression models based on these genes in the four datasets were all significant. In GSE8894 and GSE30129, Cox models based on these genes were the most significant. In GSE37745 and TCGA LUSC, the Cox models of normal based genes got the lowest p-value while the Log2TN based Cox models were the second. For each Cox model in each dataset, we calculated the prognostic index (PI) for each sample. We divided a dataset into two groups with the median of PI as the cutoff. Then, Kaplan-Meier curves were drawn and *p-values* were obtained by the log-rank test. The results of the three conditions in four datasets were shown in Fig. 1. The significant level of them was summarized in supplementary table 1. Except the TCGA LUSC, Log2TN based Cox models in the three datasets from GEO got the most difference between the two survival curves. It is unbelievable that tumor-derived genes from TCGA LUSC got the worst performance in the three datasets of GEO while the expression profile of the three datasets was tumor-derived. It seems that the adjacent-normal tissue of tumor plays an important role in the progress of tumor. Maybe even if it is adjacent-normal tissue, something has happened, which makes it not the same as the non-tumor normal tissue. Hence, Log2TN that calculated based on adjacent-normal and tumor could get the most robust genes.

In the previous analysis of this study, we used only the 40 tumor-normal paired samples with recurrence event information in TCGA LUSC because Log2TN needed paired samples to calculate the value. For comparable, these 40 tumor samples were used for tumor and normal too previously. To test the robustness of the selected genes via tumor, normal, and Log2TN, we built the Cox model with the 386 tumor samples that have recurrence information in the TCGA LUSC using tumor-derived expression profile based on the selected genes for them respectively. For normal-derived genes, it was not significant ($p = 0.461$, likelihood ratio test). It is reasonable because genes were selected based on the expression profile of normal while tested using tumor-derived expression profile. However, compared to the tumor, Log2TN derived genes got a more significant Cox model (tumor: $p = 0.0315$, Log2TN: $p = 0.0191$, likelihood ratio test). For each of the three models, we calculated the PI and divided the 386 samples into two groups using median of PI. Then, we drew Kaplan-Meier curves and assessed the difference between

two survival curves using log-rank test. As shown in Fig. 2, although it isn't significant in the Cox model using normal-derived genes, the survival curve is discriminating. The power of their distinguishing for survival is almost the same. Thus, not only the tumor tissue, but also the adjacent-normal tissue may influence the outcome.

Gene set enrichment analysis of Log2TN genes:

Log2TN derived genes show the best robust performance. Thus we investigated the biological functions of genes ranked according to Log2TN univariable Cox regression via gene set enrichment analysis of KEGG pathways. For each gene, we multiply the sign of coefficient by -log10 of the p-value as its prognosis score. For example, if the coefficient and p-value of the Log2TN univariable Cox regression for a gene is -0.35 and 0.01, the score for this gene − 1 * (-log(0.01, 10)) = -2. Then, we ranked genes with this score and performed GSEA analysis. As shown in Fig. 3, these positively correlated pathways include multiple biological pathways, such as cell cycle, DNA replication, homologous recombination, ribosome, primary immunodeficiency, and p53 signaling pathway. The high-level Log2TN of genes in these pathways results in worse outcomes. These negatively correlated pathways mainly include metabolism-related pathways such as complement and coagulation cascades, fatty acid metabolism, retinol metabolism, primary bile acid biosynthesis, tryptophan metabolism, drug metabolism cytochrome P450, valine leucine and isoleucine degradation, and glycine serine and threonine metabolism. It seems that the tumor cells mainly focus on proliferation with losing functional metabolism.

Discussion

With the development of high-throughput technology, tremendous data of biomedicine accumulated. It provides the opportunity to explore new biomarkers of diagnosis and prognosis. Biomarkers evolved from a single gene or few genes before to multiple genes now. Maybe it would develop towards entire omics or even the combination of multi-omics in the future with reducing cost. These genes that are aberrantly expressed in cancer tissue have attracted substantial interest in the construction of biomarkers using expression profiles of tumors[13, 22–25]. However, it assumed that all the difference between tumor and adjacent-normal comes from tumor tissue and the adjacent-normal tissue has nothing changed during the tumor development. If an aberrantly expressed gene mainly results from the change of adjacent-normal tissue, when constructed the biomarker using tumor-derived expression profiles, it may not achieve good performance. Furthermore, it is hard to get the absolute expression profiles in tumor tissue. Different methods and technicians could also bring in error. In the present study, we didn't use the aberrantly expressed genes in cancer. Instead of it, we used the log2-transformed ratio between tumor and adjacent-normal to do Cox regression and filter prognosis related genes in the TCGA LUSC. The significantly prognosis related genes filtered via Log2TN had more overlapped genes with the three GEO datasets, even if the tumor-derived expression profiles were used. In the TCGA LUSC, Cox regression p-values via Log2TN were lower than the other two. Moreover, genes filtered via Log2TN show more robust with fewer inconsistent genes. These LASSO-Cox selected Genes that were used for the construction of

the Cox regression model also showed better performance for Log2TN in the three GEO datasets. From our study, we could conclude that: (1) combination of the expression profiles of tumor and adjacent-normal could obtain more robust recurrence prognosis related genes; (2) the adjacent-normal tissue may also play an important role in the progress and outcome of the tumor. This finding opens a door to construction of a more robust prognostic strategy from the combination of the tumor and normal. We hope that in this way, constructed biomarkers could be used in any way, such as Q-PCR, array, and RNAseq in the future because what it needed is the relative value.

Nevertheless, there are inevitably some limitations in our study that should be acknowledged. First, there are only 40 tumor-normal paired samples in TCGA LUSC could be used in our study for analyzing the Log2TN. It may not provide enough power to support our conclusion. Second, all of the three GEO datasets provide only the tumor-derived expression profile. Good verification couldn't be got via these three datasets. Third, although genes Log2TN show more robust, it isn't a huge gap between Log2TN and tumor. Log2TN doesn't have enough advantage. More researches are needed in the future to establish it.

Abbreviations

DSS1: Deleted in Split hand/Split foot 1; FDA: food and drug administration; fPSA: free prostate-specific antigen; GEO: Gene Expression Omnibus; GSEA: gene set enrichment analysis; GPC-1: glypican-1; HER2: human epidermal growth factor receptor 2; LASSO: least absolute shrinkage and selection operator; LUSC: lung squamous cell carcinoma; Log2TN: Log2 transformed ratio of tumor tissue and the adjacent normal tissue; miRNA: micro RNA; lncRNA: long non-coding RNA; MSigDB: molecular signatures database; PCA3: Prostate cancer antigen 3; PSA: prostate-specific antigen; TCGA: The Cancer Genome Atlas.

Declarations

Acknowledgements

This study was supported by the National Natural Science Foundation of China (Grant Nos. 31500756 to Yuanyuan Qiao).

Conflict of Interest:

The authors declare that there are no conflicts of interest.

Data Availability statement:

All data used in this study are openly available. LUSC is TCGA dataset downloaded at https://xenabrowser.net/datapages, and three GEO dataset were downloaded at

http://www.ncbi.nlm.nih.gov/geo/.

Ethics statement:

Not applicable.

Authors' contributions:

Wei Ma and Yuanyuan Qiao conceived and designed the study. Dandan Li and Wei Ma performed the computational analysis. Changjian Zhang and Ming Xiong drafted the manuscript. Wei Ma and Yuanyuan Qiao supervised the study.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians 2018;68:394-424.
2. Romaszko AM, Doboszyńska A. Multiple primary lung cancer: A literature review. Advances in clinical and experimental medicine : official organ Wroclaw Medical University 2018;27:725-30.
3. Singh A, Gupta S, Sachan M. Epigenetic Biomarkers in the Management of Ovarian Cancer: Current Prospectives. Frontiers in cell and developmental biology 2019;7:182.
4. Dimitroulis D, Damaskos C, Valsami S, Davakis S, Garmpis N, Spartalis E, Athanasiou A, Moris D, Sakellariou S, Kykalos S, Tsourouflis G, Garmpi A, et al. From diagnosis to treatment of hepatocellular carcinoma: An epidemic problem for both developed and developing world. World journal of gastroenterology 2017;23:5282-94.
5. Guan X. Cancer metastases: challenges and opportunities. Acta pharmaceutica Sinica. B 2015;5:402-18.
6. Jiang WG, Sanders AJ, Katoh M, Ungefroren H, Gieseler F, Prince M, Thompson SK, Zollo M, Spano D, Dhawan P, Sliva D, Subbarayan PR, et al. Tissue invasion and metastasis: Molecular, biological and clinical perspectives. Seminars in cancer biology 2015;35 Suppl:S244-s75.
7. Fidler IJ, Kripke ML. The challenge of targeting metastasis. Cancer metastasis reviews 2015;34:635-41.
8. Stamey TA, Yang N, Hay AR, McNeal JE, Freiha FS, Redwine E. Prostate-specific antigen as a serum marker for adenocarcinoma of the prostate. The New England journal of medicine 1987;317:909-16.
9. Catalona WJ, Partin AW, Slawin KM, Brawer MK, Flanigan RC, Patel A, Richie JP, deKernion JB, Walsh PC, Scardino PT, Lange PH, Subong EN, et al. Use of the percentage of free prostate-specific antigen to enhance differentiation of prostate cancer from benign prostatic disease: a prospective multicenter clinical trial. Jama 1998;279:1542-7.

10. Bussemakers MJ, van Bokhoven A, Verhaegh GW, Smit FP, Karthaus HF, Schalken JA, Debruyne FM, Ru N, Isaacs WB. DD3: a new prostate-specific gene, highly overexpressed in prostate cancer. Cancer research 1999;59:5975-9.
11. Geisler C, Gaisa NT, Pfister D, Fuessel S, Kristiansen G, Braunschweig T, Gostek S, Beine B, Diehl HC, Jackson AM, Borchers CH, Heidenreich A, et al. Identification and validation of potential new biomarkers for prostate cancer diagnosis and prognosis using 2D-DIGE and MS. BioMed research international 2015;2015:454256.
12. Wang S, Qiu Y, Bai B. The Expression, Regulation, and Biomarker Potential of Glypican-1 in Cancer. Frontiers in oncology 2019;9:614.
13. Goossens N, Nakagawa S, Sun X, Hoshida Y. Cancer biomarker discovery and validation. Translational cancer research 2015;4:256-69.
14. Prat A, Pineda E, Adamo B, Galván P, Fernández A, Gaba L, Díez M, Viladot M, Arance A, Muñoz M. Clinical implications of the intrinsic molecular subtypes of breast cancer. Breast (Edinburgh, Scotland) 2015;24 Suppl 2:S26-35.
15. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, Quackenbush JF, Stijleman IJ, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2009;27:1160-7.
16. Yang L, Roberts D, Takhar M, Erho N, Bibby BAS, Thiruthaneeswaran N, Bhandari V, Cheng WC, Haider S, McCorry AMB, McArt D, Jain S, et al. Development and Validation of a 28-gene Hypoxia-related Prognostic Signature for Localized Prostate Cancer. EBioMedicine 2018;31:182-89.
17. Sierzega M, Kaczor M, Kolodziejczyk P, Kulig J, Sanak M, Richter P. Evaluation of serum microRNA biomarkers for gastric cancer based on blood and tissue pools profiling: the importance of miR-21 and miR-331. British journal of cancer 2017;117:266-73.
18. Chandra Gupta S, Nandan Tripathi Y. Potential of long non-coding RNAs in cancer patients: From biomarkers to therapeutic targets. International journal of cancer 2017;140:1955-67.
19. Venza M, Visalli M, Catalano T, Beninati C, Teti D, Venza I. DSS1 promoter hypomethylation and overexpression predict poor prognosis in melanoma and squamous cell carcinoma patients. Human pathology 2017;60:137-46.
20. Yu F, Quan F, Xu J, Zhang Y, Xie Y, Zhang J, Lan Y, Yuan H, Zhang H, Cheng S, Xiao Y, Li X. Breast cancer prognosis signature: linking risk stratification to disease subtypes. Briefings in bioinformatics 2019;20:2130-40.
21. Srivastava A, Creek DJ. Discovery and Validation of Clinical Biomarkers of Cancer: A Review Combining Metabolomics and Proteomics. Proteomics 2019;19:e1700448.
22. Chen X, Wang YW, Zhu WJ, Li Y, Liu L, Yin G, Gao P. A 4-microRNA signature predicts lymph node metastasis and prognosis in breast cancer. Human pathology 2018;76:122-32.
23. Xu G, Zhang M, Zhu H, Xu J. A 15-gene signature for prediction of colon cancer recurrence and prognosis based on SVM. Gene 2017;604:33-40.

24. Wang X, Zhou J, Xu M, Yan Y, Huang L, Kuang Y, Liu Y, Li P, Zheng W, Liu H, Jia B. A 15-lncRNA signature predicts survival and functions as a ceRNA in patients with colorectal cancer. Cancer management and research 2018;10:5799-806.
25. Zhu X, Tian X, Yu C, Shen C, Yan T, Hong J, Wang Z, Fang JY, Chen H. A long non-coding RNA signature to improve prognosis prediction of gastric cancer. Molecular cancer 2016;15:60.

Figures

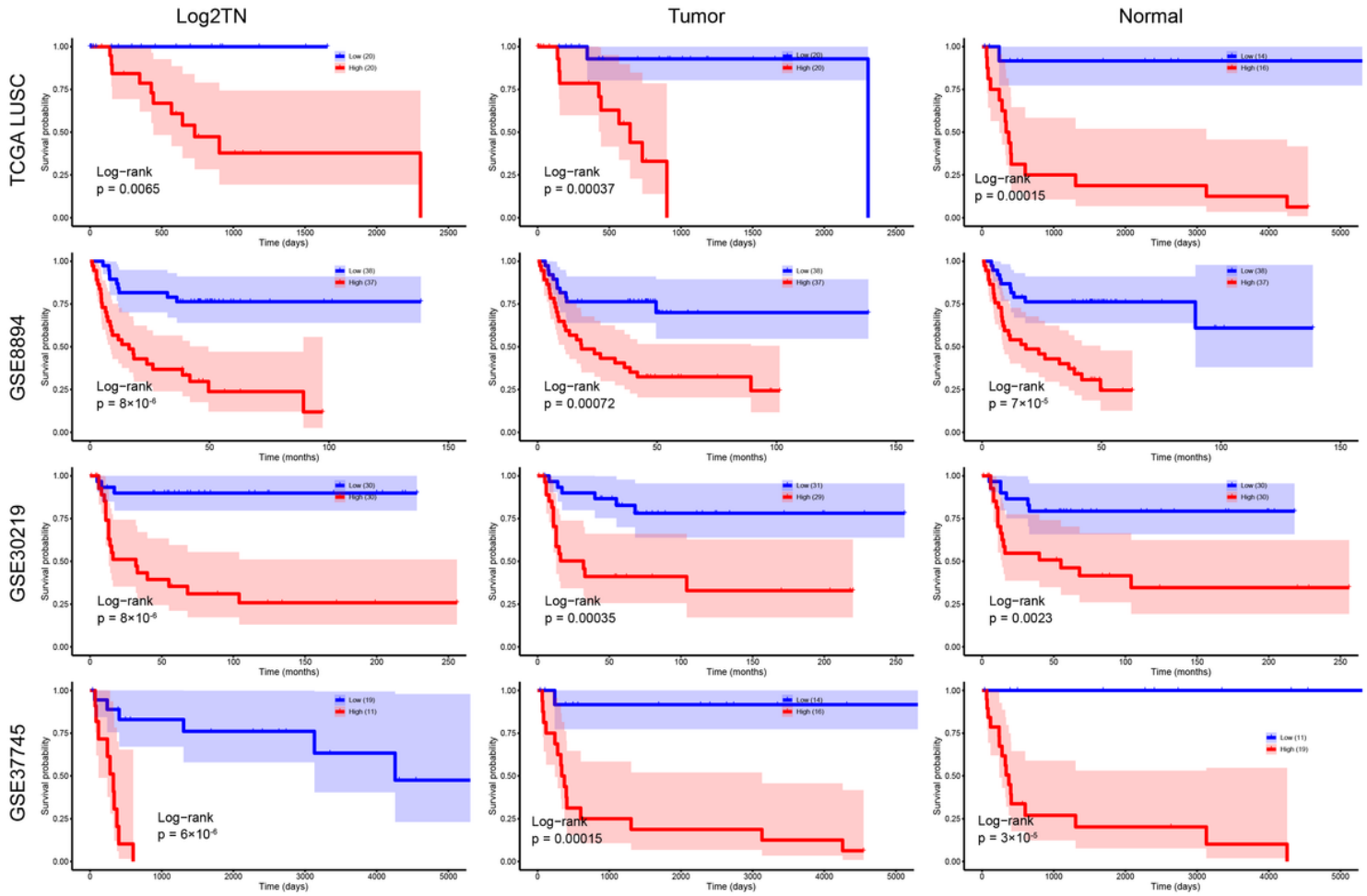

Figure 1

Performance of gene signatures from Log2TN, tumor and normal on TCGA LUSC and three GEO datasets.

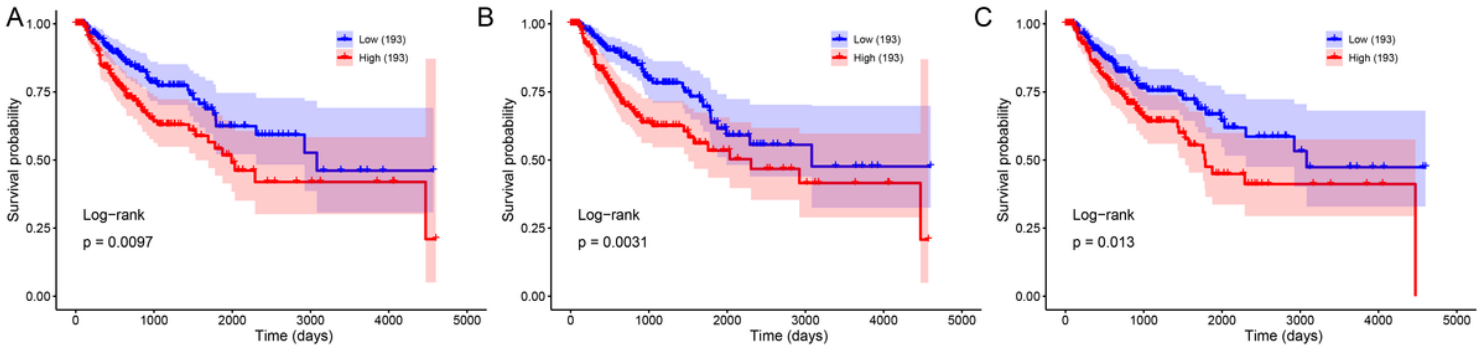

Figure 2

Performance of gene signatures from Log2TN, tumor and normal on the full TCGA LUSC dataset. A: signature of Log2TN; B: signature of tumor; C: signature of normal.

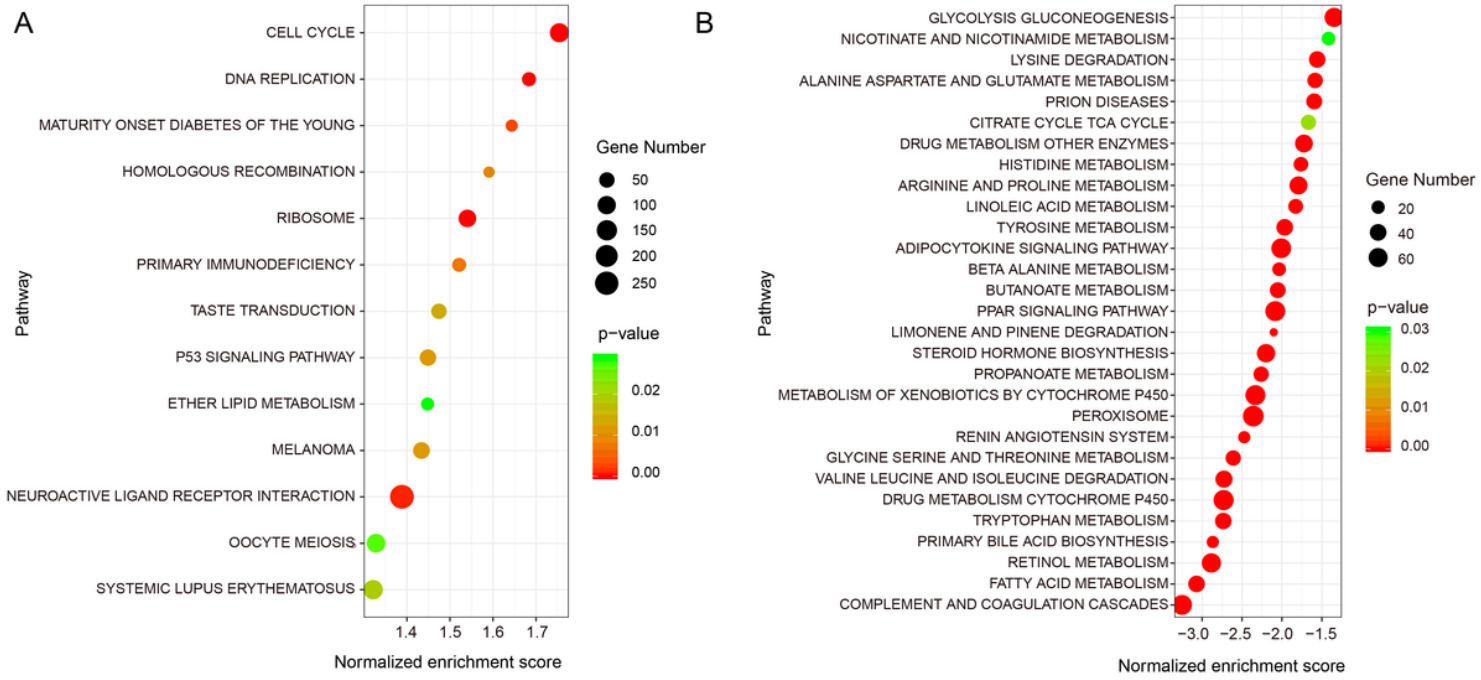

Figure 3

Enriched KEGG pathways. A: positive enrichment pathways; B: negative enriched pathways.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarytable1.xlsx
- OnlineFigures1.png