The Expression of RN7SL494P (7SL) Predicts Nodal Metastasis and Prognosis in Lung Adenocarcinoma

Xiao Zhu[1,2], Yongmei Huang[1], Hui Luo[1*], Ying Xu[2*]

1 The Marine Biomedical Research Institute, Guangdong Medical University, Zhanjiang, China

2 Computational Systems Biology Lab (CSBL), Institute of Bioinformatics, University of Georgia, Athens, GA USA

[*]Correspondence:

luohui@gdmu.edu.cn (HL)

xyn@uga.edu (YX)

Abstract

The metastasis of lung cancer can spread to the lymph nodes around the lungs. Metastasis, rather than the primary cancer, judges patients survival. Wherefore, a more detailed study on transcriptome of metastatic lung adenocarcinoma (LUAD) including primary carcinoma was carried out. LUAD RNA-seq data and the corresponding clinical information were available from The Cancer Genome Atlas (TCGA), which included 522 cases but only 515 cases have transcriptome data. Differential expression analyses between cases and controls, between primary cancer and metastasis subgroup, or between TNM stages, were respectively carried out using edgeR package. Then, the Kruskal-Wallis tests were used to verify the gradient changes of cancer metastasis or staging with the differential expression genes. The survival analyses were calculated using the Kaplan-Meier algorithm and log-rank test. The functional predictions for the differentially expressed genes were porformed with the Gene Ontology and Kyoto Encyclopedia of Genes and Genomes (GO/KEGG). Single gene set enrichment analysis (single GSEA) was run to explore the biological pathways associated with the expressions of RN7SL494P gene based on the Molecular Signatures Database (MSigDB). 406 and 439 differentially expressed genes were identified respectively in lymph node metastasis or TNM stages. 112/296 intersection genes were associated with nodal metastasis and/or staging, among them only 25 genes were associated with the nodal metastasis, 13 genes were associated with the staging with gradient changes. Only one gene (RN7SL494P) was found to be associated with prognosis. But RN7SL494P was not found joining any biological functions or processes or cellular components with GO/KEGG analyses. Finally, single GSEA enrichment and pathway analyses showed that RN7SL494P might be involving in cancer development process and poor outcome in lung adenocarcinoma. These findings highlight the potential applications of RN7SL494P as a promising molecular predictor not only in nodal metastasis but prognosis evalution in lung adenocarcinoma patients.

Keywords: TCGA; lung adenocarcinoma; RN7SL494P; nodal metastasis; prognosis

1. Introduction

Lung adenocarcinoma (LUAD), a histological subtype of non-small cell lung cancer (NSCLC), rises when healthy cells change and uncontrolledly grow in the outer region of the lung. It is the most common lung cancer, and accounting for about 40 percent of all lung-derived cancers [1].

Lung adenocarcinoma tends to grow in smaller airways, such as bronchioles, which develops more tardily than any other sorts of lung cancer. Once cancerous tissues growing, it may cause cancer cells to fall off. These cells can be taken away in the blood, or float in the lymph fluid which encompasses the lung tissue [2]. The lymph flows through pipes called lymphatic vessels, which inflows into collecting station called lymph nodes [3, 4]. When a cancer cell passes through the bloodstream into a lymph node or a distant body, it is called metastasis.

In this study, we provided a comprehensive screening for nodal metastasis, TNM staging with the transcriptome and clinical data in Lung adenocarcinoma of The Cancer Genome Atlas (TCGA) project. TCGA began in 2006 [5], which is a joint research project between the National Human Genome Research Institute and the National Cancer Institute.

2. Results

2.1. The Differential Expression Genes in Lung Adenocarcinoma

We conducted gene differential expression analysis and found total 13118 differential expression genes, among them, 2800 down-regulated genes and 10318 up-regulated genes. The top 10 significant down- and up-regulated genes were shown in Table 2. We chose all significantly up- and down- regulated mRNA to draw their expression on the heatmap and volcanic map (Figure 2 A and B).

Table 2. The top 10 significant down- and up-regulated genes associated with lung adenocarcinoma.

	Genes	logFC	logCPM	P value	FDR
Down-regulated	RTKN2	-4.068647319	5.46758936	4.80E-226	1.68E-221
	FAM107A	-4.529447985	5.196616095	2.82E-213	4.94E-209
	OTUD1	-2.103762476	4.564398143	1.27E-208	1.47E-204
	EPAS1	-2.695426362	9.192377662	9.16E-203	8.00E-199
	TEK	-3.244335981	4.34234913	5.41E-199	3.78E-195
	S1PR1	-2.841234568	5.220910855	1.01E-197	5.90E-194
	RGCC	-2.871276087	6.177191804	2.42E-197	1.21E-193
	SEMA3G	-3.203024311	3.936371095	8.92E-197	3.90E-193
	SPAAR	-2.710967738	0.421822555	4.05E-195	1.57E-191
	STX11	-2.992538831	3.548342519	6.58E-194	2.30E-190
Up-regulated	PYCR1	3.72498895	6.81287369	2.96E-94	6.06E-92
	TEDC2	3.528502981	2.55920153	2.23E-75	2.87E-73
	IQGAP3	3.632830071	4.869445161	8.84E-70	9.63E-68
	ETV4	3.853849662	5.517196991	8.15E-69	8.66E-67
	FAM83A	6.825890252	7.149252358	1.27E-68	1.34E-66
	TOP2A	3.886140596	6.847342295	3.26E-66	3.14E-64

	GOLM1	2.662542406	8.154528815	5.03E-66	4.83E-64
	SAPCD2	3.78667196	3.975269302	9.14E-66	8.73E-64
	TMEM184A	3.148020595	5.406608823	7.89E-65	7.24E-63
	ALDH18A1	1.633757577	6.730939458	1.98E-64	1.80E-62

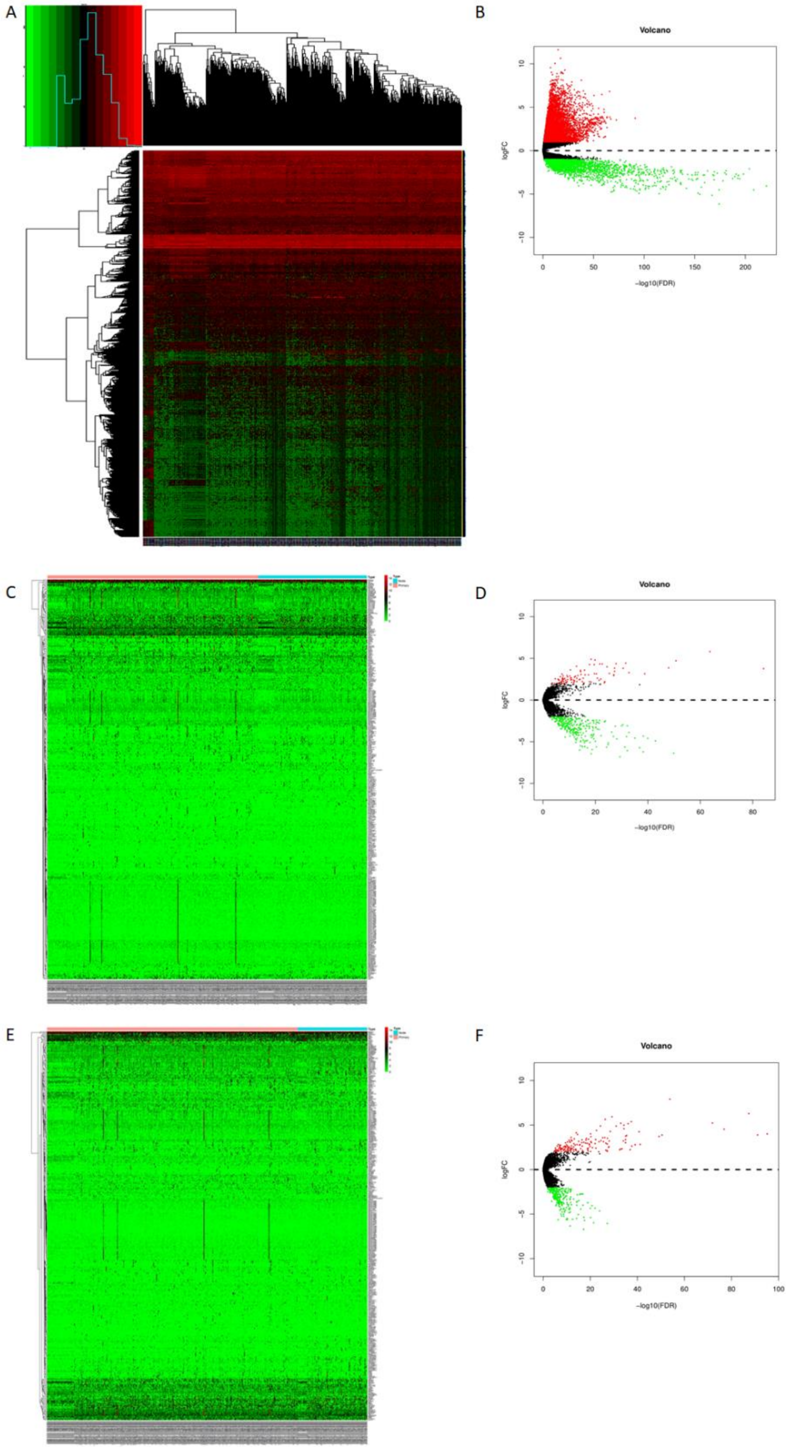

Figure 2. The differentially expressed analyses.

A,B Total differential expression genes in LUAD (A heatmap, B volcano map);

C,D The differential expression genes in nodal metastasis (C heatmap, D volcano map);

E,F The differential expression genes in TNM staging (E heatmap, F volcano map).

2.2. GO and KEGG Analysis of Differentially Expressed Genes

We conducted a GO analysis of all differentially expressed genes in LUAD and found that RN7SL494P was not involved in any biological functions or processes or cellular components in *DAVID* database (Figure 4 A and B). Kobas was used for differential gene functional annotation with KEGG pathway. Indeed, after identifying key KEGG pathways, we also did not find RN7SL494P-related pathways (Supplement Table 1). The functional annotation of the differentially expressed genes with clusterProfiler R package also did not find RN7SL494P-related KEGG pathways (Supplement Table 2). Therefore, a single gene functional enrichment method associated with specific gene would be studied in the following step.

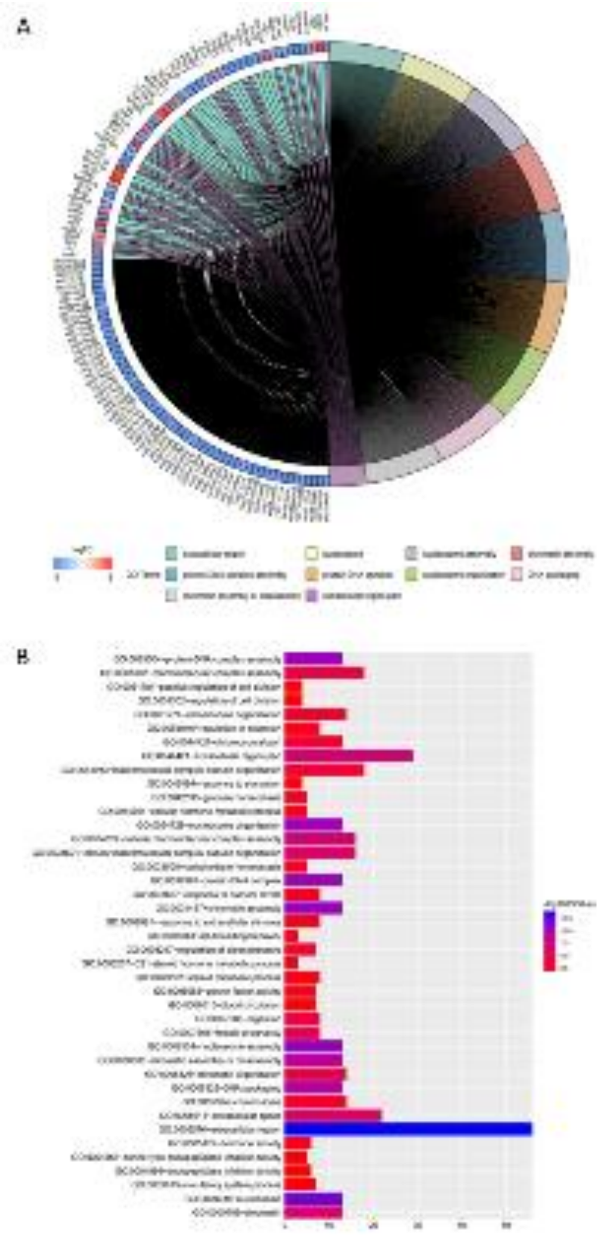

Figure 4. GO analyses of all differentially expressed genes in LUAD.

A The biological functions, biological processes or cellular components in *DAVID* database by GOplot analysis;

B The enrichment of differentially expressed genes.

2.3. The Differentially Expressed Genes Associated With Nodal Metastasis or TNM Staging

Then, based on the lymph node metastasis features of the subjects in Table 1, a total of at least 406 differential genes were obtained, and 312 genes were significantly up-regulated, and 94 genes were significantly down-regulated (Figure 2 C and D). The top 10 significant down- and up-regulated genes associated with cancer metastasis were shown in Table 3. Similarly, the TNM staging-related differentially expressed genes were shown in Figure 2 E and F, and its top 10 significant down- and up-regulated genes were shown in Table 3.

Table 3. The top 10 significant down- and up-regulated genes associated with lymph node metastasis or TNM stages.

	Genes	logFC	logCPM	P value	FDR
lymph node metastasis					
	7SK	-6.38529	5.376253	1.92E-54	1.62E-50
Down-regulated	SNORA73B	-4.89863	3.941633	2.30E-47	1.29E-43
	SNORD17	-4.59669	2.395325	2.50E-44	1.20E-40
	SCARNA6	-4.33589	-0.05263	1.57E-42	5.85E-39
	SCARNA5	-6.21022	2.186735	1.80E-42	6.05E-39
	SCARNA10	-5.72043	1.274605	4.91E-41	1.45E-37
	MSTN	-4.5735	1.589033	3.65E-39	9.44E-36
	SCARNA7	-3.88216	-0.07935	7.12E-37	1.68E-33
	SCARNA13	-3.0861	0.701175	3.84E-36	7.73E-33
	RNU4-1	-6.06981	1.417159	3.91E-36	7.73E-33
	NNAT	3.773884	2.325209	2.37E-89	7.97E-85
	LRRC38	5.827189	1.230182	1.32E-68	2.23E-64
Up-regulated	VSX2	4.728637	-1.57565	1.85E-55	2.07E-51

	AC087257.2	3.860068	-2.07862	1.84E-52	1.24E-48
	LINC01433	3.163173	-2.46671	3.82E-43	1.61E-39
	FAM205C	3.293196	-2.91757	7.49E-37	1.68E-33
	AL161668.1	4.428092	-3.68113	1.56E-35	2.77E-32
	RTP1	3.811513	-2.18471	1.65E-34	2.64E-31
	GSG1L2	4.357816	-3.36664	1.11E-31	1.44E-28
	CALB1	3.446571	3.71567	4.30E-31	5.16E-28
TNM stages Down-regulated	7SK	-6.062979794	5.353737093	6.74E-31	6.13E-28
	SNORA73B	-4.647298488	3.923197774	1.76E-27	1.26E-24
	SNORD17	-4.325760083	2.373029662	1.51E-25	9.39E-23
	SCARNA5	-5.981744956	2.167025263	5.95E-25	3.63E-22
	SCARNA6	-4.052245334	-0.070346602	2.75E-24	1.62E-21
	SCARNA10	-5.377554867	1.251887603	1.35E-23	7.42E-21
	MSTN	-4.340706495	1.520356478	1.70E-23	8.96E-21
	SCARNA7	-3.7237236	-0.10084326	2.81E-22	1.26E-19
	RNU4-1	-5.712216563	1.396659528	6.97E-21	2.79E-18
	RNU4-2	-5.357340472	2.502793386	1.26E-20	4.94E-18
Up-regulated	PPIAP46	4.012250624	-0.902056299	1.92E-100	6.46E-96
	HNRNPA1P52	3.896195799	-1.852379148	4.95E-96	8.32E-92
	LRRC38	6.291094962	1.168289657	3.92E-92	4.39E-88
	AC087257.2	4.527651209	-2.097876396	1.53E-81	1.28E-77
	VSX2	5.232030072	-1.594431836	1.90E-76	1.28E-72
	PSG11	7.901940389	-1.563148972	2.93E-58	1.64E-54
	FAM205C	3.883821429	-2.930755861	6.97E-55	3.35E-51
	FXNP2	3.718917178	-3.212051642	1.45E-53	6.09E-50
	MARCH4	2.823672408	0.803655353	1.61E-45	6.02E-42
	RTP1	4.254620305	-2.219012962	5.41E-45	1.82E-41

2.4. The Overlapping Differentially Expressed Genes Associated With Nodal Metastasis and TNM Staging

Venn diagram analysis was carried out to visualize the overlapping differentially expressed genes between lymph node metastasis and TNM stages using VennDiagram R packa ge. 296 overlapping genes were found (Figure 3 A).

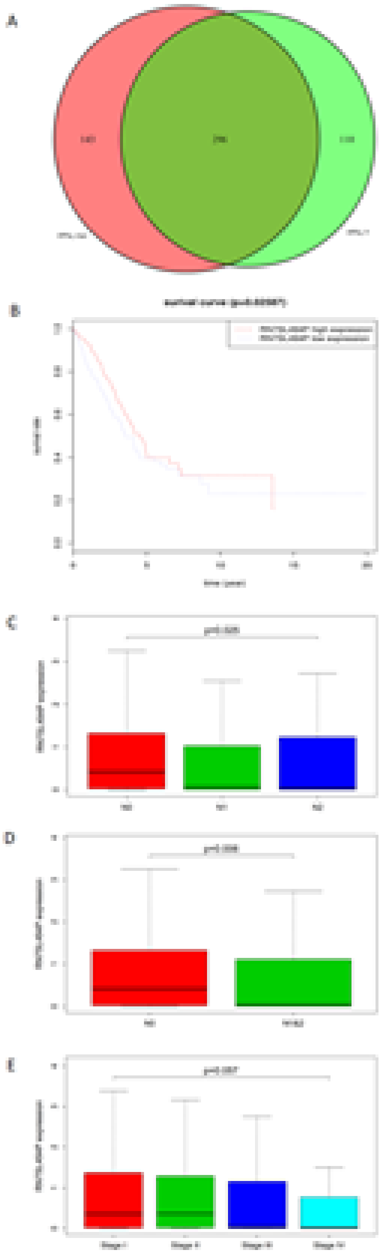

Figure 3. The overlapping differentially expressed genes associated with nodal metastasis and TNM staging.

A The venn diagram of differentially expressed genes between nodal metastasis and TNM staging;

B Survival analysis of differentially expressed RN7SL494P associated with nodal metastasis;

C Kruskal-Wallis test for differentially expressed RN7SL494P associated with the gradient changes on lymph node metastasis (N0 *vs.* N1 *vs.* N2);

D Kruskal-Wallis test for differentially expressed RN7SL494P associated with the gradient changes on lymph node metastasis (N0 *vs.* N1&N2);

E Kruskal-Wallis test for differentially expressed RN7SL494P associated with the gradient changes on TNM staging.

2.5. The Gradient Changes of Differentially Expressed Genes Associated With Nodal Metastasis and TNM Staging

We analyzed the gradient changes of differentially expressed genes in lymph node metastasis (from N0 to N2) and TNM stage (from I to IV) with Kruskal-Wallis test. Because the number of N3samples was only two, this subgroup would not be considered in this section. 112 differentially expressed genes were associated with the gradient changes of lymph node metastasis, or TNM stages, or metastasis and TNM stages (Table 4). Among them, 25 differentially expressed genes were associated with the lymph node metastasis; 13 differentially expressed genes were associated with the TNM stages; and only 7 genes (SCARNA7, AC105999.2, RANBP20P, RN7SL151P, SYNPR, AL512638.1, and TMIGD1) were simultaneously associated with lymph node metastasis and TNM stages.

Table 4. The gradient changes of differentially expressed genes associated with lymph node metastasis or TNM stages with the Kruskal-Wallis test, and the survival analysis of patients with the differentially expressed genes.

Genes	**Lymph node metastasis (N0-N1-N2)**		**TNM stages (I-II-III-IV)**		**Log-rank test**
	Gradient change	***P***	**Gradient change**	***P***	*P*
NNAT	NA	0.019	NA	0.025	/
VSX2	yes,downtrend	0.008	/	0.586	0.08025
SCARNA7	yes,downtrend	0.011	yes,downtrend	0.018	0.34227
AL161668.1	NA	0	NA	0.002	/
SNORA12	NA	0.013	NA	0.003	/
GSG1L2	yes,upward	0	/	0.604	0.36278
CYP2B6	/	0.287	NA	0.032	/
ALB	/	0.197	NA	0.008	/
VN1R35P	yes,upward	0.003	/	0.157	0.08025
SNORA71A	NA	0.04	/	0.842	/
AL451054.3	NA	0	NA	0.012	/
AC105999.2	yes,upward	0.042	yes,upward	0.012	0.13752
RN7SL3	yes,upward	0.048	/	0.266	0.09487
LINC01819	yes,downtrend	0.016	NA	0.021	/
RANBP20P	yes,downtrend	0.019	yes,downtrend	0.016	0.07001
RNU5A-1	/	0.066	yes,downtrend	0.015	0.75953
RN7SKP255	/	0.101	NA	0.005	/
AL513304.1	yes,upward	0.019	/	0.073	0.37489
HIST1H4F	/	0.191	NA	0	/

RN7SKP203	NA	0.006	/	0.334	/
HIST1H4L	/	0.342	NA	0.048	/
RN7SL769P	NA	0.01	/	0.116	/
RN7SL151P	yes,downtrend	0.006	yes,downtrend	0.009	0.28316
GKN1	NA	0.039	/	0.272	/
FXNP2	NA	0.006	/	0.508	/
RNY3	NA	0.003	/	0.067	/
AC112495.1	yes,downtrend	0.012	NA	0.002	0.88522
SYNPR	yes,downtrend	0.034	yes,downtrend	0.002	0.14163
RN7SL480P	yes,downtrend	0.03	/	0.169	0.97413
RN7SL116P	yes,downtrend	0.019	/	0.057	0.71102
AC036111.1	NA	0.004	/	0.195	
RNA5-8SP2	NA	0	/	0.088	
RN7SL300P	NA	0.026	/	0.079	
HIST1H2AH	yes,upward	0.014	NA	0.012	0.89036
PSG11	/	0.126	NA	0.002	
GLRA4	yes,downtrend	0.003	/	0.322	0.08082
RN7SL359P	NA	0	/	0.052	/
AL135929.2	NA	0.006	/	0.14	/
CYP11B1	NA	0.029	/	0.123	/
RN7SL342P	NA	0.02	/	0.062	/
SPAG11B	yes,upward	0.028	/	0.064	0.54783
RN7SL732P	NA	0.005	/	0.082	/
CYP1D1P	NA	0	NA	0.002	/
RN7SL791P	NA	0	NA	0.002	/
RN7SKP189	NA	0.002	/	0.696	/
RN7SKP71	yes,downtrend	0.011	NA	0.025	0.24259
RN7SL217P	NA	0.029	NA	0.041	/
RN7SL272P	NA	0	NA	0.016	/
RHOXF2B	NA	0	/	0.093	/

RN7SL464P	NA	0.003	/	0.214	/
CRISP1	NA	0.007	/	0.074	/
FGF4	/	0.379	NA	0.019	/
CRP	NA	0.026	/	0.066	/
PSG2	/	0.347	NA	0.03	/
RN7SL197P	NA	0.017	/	0.644	/
RN7SL646P	NA	0.003	/	0.111	/
RN7SL554P	NA	0.001	/	0.317	/
PPP1R3A	NA	0.009	/	0.226	/
RN7SL597P	/	0.056	NA	0.017	/
RN7SL308P	NA	0.001	NA	0.003	/
AC106872.1	NA	0	NA	0.003	/
AL135929.1	NA	0.007	/	0.086	/
AL512638.1	yes,upward	0.002	yes,upward	0	0.80925
RN7SL711P	/	0.104	yes,downtrend	0.022	0.6968
HMGB3P18	NA	0.018	NA	0.022	/
RN7SL126P	NA	0.021	/	0.106	/
RN7SL630P	NA	0.002	/	0.066	/
RN7SL494P	yes,downtrend	0.025	/	0.057	0.02587
RN7SL7P	NA	0.024	/	0.23	/
RN7SL786P	NA	0.021	/	0.118	/
AC108515.1	NA	0	NA	0.005	/
RN7SKP185	NA	0.023	yes,downtrend	0.02	0.66366
RN7SKP90	NA	0	yes,downtrend	0.017	0.91288
AC008808.2	/	0.814	NA	0.024	/
RN7SL390P	NA	0.012	/	0.445	/
SCARNA3	NA	0	NA	0.007	/
MIR124-2HG	NA	0.002	NA	0.012	/
RN7SL297P	NA	0.001	NA	0.002	/
RNU1-88P	NA	0.004	/	0.35	/

RN7SL314P	NA	0.078	NA	0.038	/
RN7SL575P	NA	0.049	/	0.272	/
RN7SL302P	NA	0.04	/	0.099	/
AL513475.2	NA	0.046	/	0.401	/
KRT38	/	0.148	yes,upward	0.031	0.30421
OR4A16	NA	0.004	NA	0.003	/
FRG2	NA	0.003	/	0.699	/
LINC02557	NA	0.001	/	0.462	/
LINC01221	NA	0.002	/	0.076	/
AC012065.1	yes,upward	0	NA	0	0.25382
LINC01040	NA	0.014	NA	0.024	/
IGLV3-26	NA	0.003	NA	0.011	/
CRCT1	yes,upward	0.019	NA	0.013	0.51194
GAGE12J	NA	0.017	NA	0.007	/
CELA3A	yes,downtrend	0.035	NA	0.003	0.60893
RN7SL260P	NA	0.005	/	0.102	/
AC245291.3	/	0.105	NA	0.018	/
AC105031.2	yes,upward	0.001	NA	0.013	0.88735
AC245128.1	NA	0.008	NA	0.043	/
AC008517.1	NA	0.002	/	0.357	/
DRAXINP1	/	0.111	NA	0	/
RN7SL14P	NA	0.032	/	0.214	/
DDX11L16	NA	0.002	NA	0.02	/
ANHX	NA	0.043	NA	0.007	/
FAM9A	NA	0.018	NA	0	/
TMIGD1	yes,upward	0.001	yes,upward	0.027	0.42473
PSG7	/	0.251	yes,upward	0.001	0.74669
AC105460.1	NA	0.01	NA	0.001	/
AC080128.1	/	0.215	NA	0.036	/
BX510359.3	/	0.064	NA	0.002	/

AL139002.1	NA	0.022	/	0.747	/
MIR3976HG	/	0.195	NA	0.002	/
SPAG11A	NA	0.003	NA	0.008	/

d, the deleted base. $P_{\text{corrected}}$, multiple testing by the Bonferroni correction.

2.6. Survival Analysis of Differentially Expressed Genes Associated With Nodal Metastasis and TNM Staging

We analyzed survival time with all 30 differential expression genes which associated with the gradient changes on lymph node metastasis and/or TNM stages, just one gene (RN7SL494P) was found to be associated with patient survival time (Table 4 and Figure 3B), which was simultaneously associated with the gradient changes on lymph node metastasis (P = 0.02587 for N0 *vs.* N1 *vs.* N2; Figure 3C), and 0.006 for N0 *vs.* N1&N2; Figure 3D). But this gene did not be associated with the gradient changes on TNM stages (P = 0.057; Figure 3E).

2.7. Single GSEA Enrichment and Pathway Analysis

The associations between RN7SL494P co-expressions and cancer-related pathways were carried out, and there was only one enriched pathway KEGG_RENIN_ANGIOTENSIN_SYSTEM which associated with higher expressions of RN7SL494P gene (Figure 5A and B). But there were 45 KEGG functional pathways associated with lower expressions of this gene in LUAD (Figure 5D). Figure 5A and 5B are examples showing that RN7SL494P expression levels were inversely associated with different pathways. The co-expression genes with the low-expressions of RN7SL494P were abounded in some biological or pathological pathways like

NUCLEOTIDE_EXCISION_REPAIR, MISMATCH_REPAIR, CELL_CYCLE, and OXIDATIVE_PHOSPHORYLATION *et al.* (Figure 5D). These findings suggest that low-expression of RN7SL494P might be associated with cancer development process and poor outcome in patients with lung adenocarcinoma.

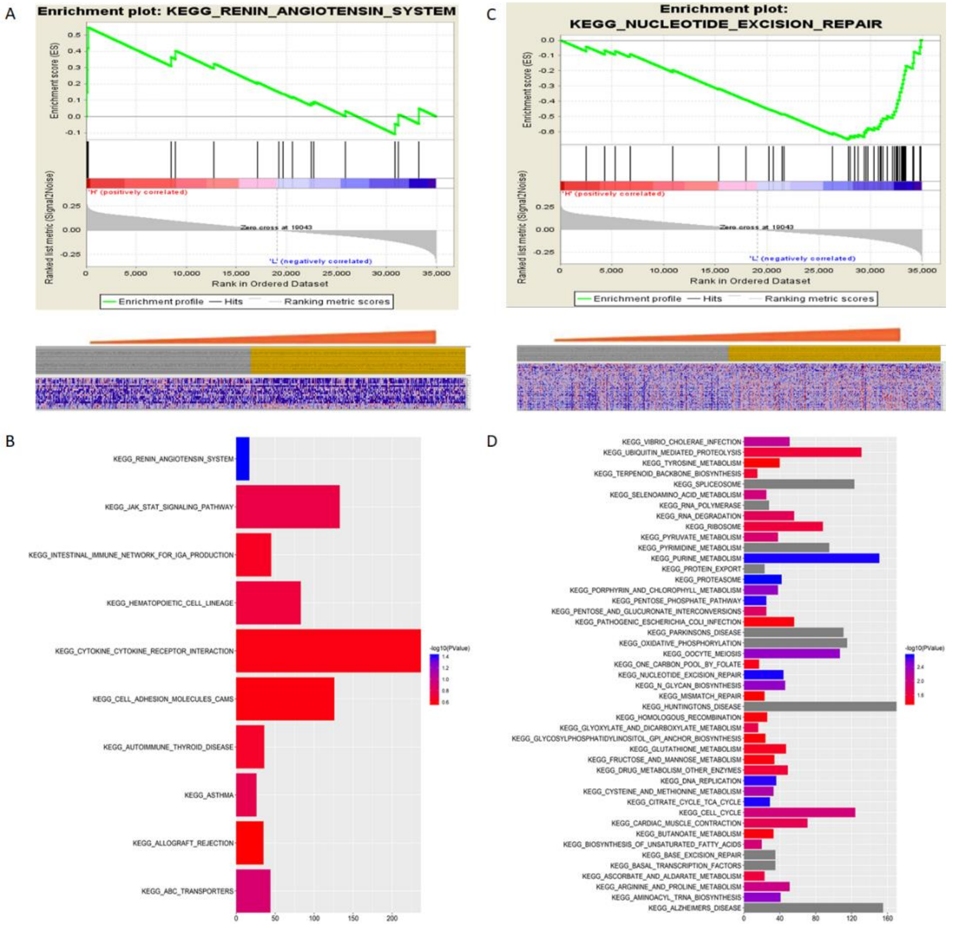

Figure 5. Single GSEA analyses.

A KEGG_RENIN_ANGIOTENSIN_SYSTEM which associated with higher expressions of RN7SL494P;

B The genes co-expressed with higher expressions of RN7SL494P were enriched in biological pathways associated with KEGG_RENIN_ANGIOTENSIN_SYSTEM;

C An examples showing that the genes co-expressed with lower expression of RN7SL494P were associated with KEGG_NUCLEOTIDE_EXCISION_REPAIR;

D The genes co-expressed with lower expressions of RN7SL494P were enriched in 45 biological pathways.

3. Discussion

Many patients were diagnosed as cancer metastasis, which makes treatment very difficult. The 5-year survival rate for metastatic lung cancer was about 1 percent [6]. When tumors spread outside the lungs, it may be difficult to cure successfully. Because none of these patients have a single best treatment, the choice of treatment strategies relies on the location, size and stages, subtypes and the lymph nodes involved.

Scientists have exploited methods for cancer patients who can screen for metastasis. The main target of screening is to reduce the number of people who die from cancer, especially from cancer metastasis. To study the “drive genes” in metastatic lung adenocarcinoma, we examined the differentially expressed genes with the repository data of RNA-seq from TCGA. We comprehensively analyzed the gene expression in lung adenocarcinoma, especially in the course of tumor metastasis.

We identified the differential expression genes which associated with lymph node metastasis and TNM stages in lung adenocarcinoma. We found that RN7SL494P gene not only possessed the above characteristics, but also prognostic significance in metastatic cancer. Subsequently, RN7SL494P single GSEA enrichment analysis further demonstrated the roles and functions of RN7SL494P.

RN7SL494P (7SL) located on 15q21.2, belongs to a long noncoding RNA (lncRNA) class pseudogene. As an eukaryotic small cytoplasmic RNAs, 7SL RNA is essential for protein translocation that binds to the ribosome and targets the newborn protein in the endoplasmic reticulum to secrete or insert the membrane during the assembly of human signal recognition particle (SRP) [7, 8]. A study with RNA sequencing from 11 human tissues showed that 7SL was is the highest expression of ncRNAs and could be an order of magnitude higher than any mRNA [9]. 7SL stimulates the GTPase activities of the SRP and its signal receptor (SR) complex [10, 11].

Defines a set of genes based on previous biological experiments, for example, knowledges about co-expression or biochemical pathways. A recent study showed the S-structure domain of 7SL RNA is related to the cellular activity in mitochondria [12]. Furthermore, except the function of NUCLEOTIDE_EXCISION_REPAIR, the results of single-GSEA demonstrated that RN7SL494P was also associated with CELL_CYCLE, RIBOSOME, DNA_REPLICATION, and UBIQUITIN_MEDIATED_PROTEOLYSIS. Thus, RN7SL494P (7SL) may play a role in the process of translation and assembly of peptides, and its dysfunction may cause pathological occurrence.

We found the high expression of RN7SL494P improved tumor survival rates in lung adenocarcinoma (high expression 41.80% vs. low expression 39.70%; Figure 3B). Yang et al. [13] found that the over-expression of FOXP3 could inhibit the transcription of 7SL RNA through binding to its promoter and subsequently strengthens the translation of p53 and conduced to repressing the growth of multiple tumors (but not include lung cancer). This study suggested that 7SL (RN7SL494P) RNA may be a direct target of FOXP3 and may be enmeshed in the configuration of FOXP3/P53 feedback loop. This indicated that there were so many complex regulatory networks in the process of tumor formation. We speculated that RN7SL494P gene may display "inconsistent functions" in different tumor microenvironments.

4. Conclusions

In the current study, we used the TCGA database to analyze expressions of genes in lung adenocarcinoma. We found that the expression of RN7SL494P (7SL) was obviously associated with nodal metastasis along with gradient changes, and its prognostic value was also better than any other genes with differential expressions.

5. Methods

5.1. The LUAD Data and Pipeline

The LUAD data from the National Cancer Institute's Genomic Data Commons data portal (https://portal.gdc.cancer.gov/repository) were downloaded on August 5, 2017 using gdc-client.exe software. This gave us 594 level-3 RNA-seq (515 cases) and 522 clinical XML datasets. The clinical data are showed in Table 1. The pipeline and its details of this study are showed in Figure 1.

Table 1. Clinical and laboratory features of the subjects included in the study.

Characteristics	Alive (n=355)	Dead with tumor (n=125)	Dead tumor free (n=42)	Total (n=522)
Age (ys)				
Mean (SD)	65.1 (9.8)	64 (10.8)	69.8 (9.8)	65 (10.3)

Median [MIN, MAX]	66 [33,88]	66.5 [40,84]	72 [53,85]	66 [33,88]
Gender				
FEMALE	193 (54.37%)	71 (56.80%)	16 (38.10%)	280 (53.64%)
MALE	162 (45.63%)	54 (43.20%)	26 (61.90%)	242 (46.36%)
metastasis				
N0	258 (72.68%)	54 (43.20%)	23 (54.76%)	335 (64.18%)
N1	51 (14.37%)	33 (26.40%)	14 (33.33%)	98 (19.77%)
N2	36 (10.14%)	34 (27.20%)	5 (11.90%)	75 (14.37%)
N3	2 (0.56%)	0 (0.00%)	0 (0.00%)	2 (0.38%)
unknown	8 (2.25%)	4 (3.20%)	0 (0.00%)	12 (2.30%)
TNM stage				
I	4 (1.13%)	0 (0.00%)	1 (2.38%)	5 (0.96%)
IA	108 (30.42%)	22 (17.60%)	4 (9.52%)	134 (25.67%)
IB	109 (30.70%)	19 (15.20%)	12 (28.57%)	140 (26.82%)
II	0 (0.00%)	0 (0.00%)	1 (2.38%)	1 (0.19%)
IIA	32 (9.01%)	15 (12.00%)	3 (7.14%)	50 (9.58%)
IIB	45 (12.68%)	25 (20.00%)	3 (7.14%)	73 (13.98%)
IIIA	36 (10.14%)	25 (20.00%)	13 (30.95%)	74 (14.18%)
IIIB	4 (1.13%)	4 (3.20%)	3 (7.14%)	11 (2.11%)
IV	11 (3.10%)	13 (10.40%)	2 (4.76%)	26 (4.98%)
unknown	6 (1.69%)	2 (1.60%)	0 (0.00%)	8 (1.53%)

TNM, tumor, nodes, metastasis-classification.

Clinical XML: cases=522 files=522

RNA-seq: cases=515 files=594

7 cases no RNA-seq data.

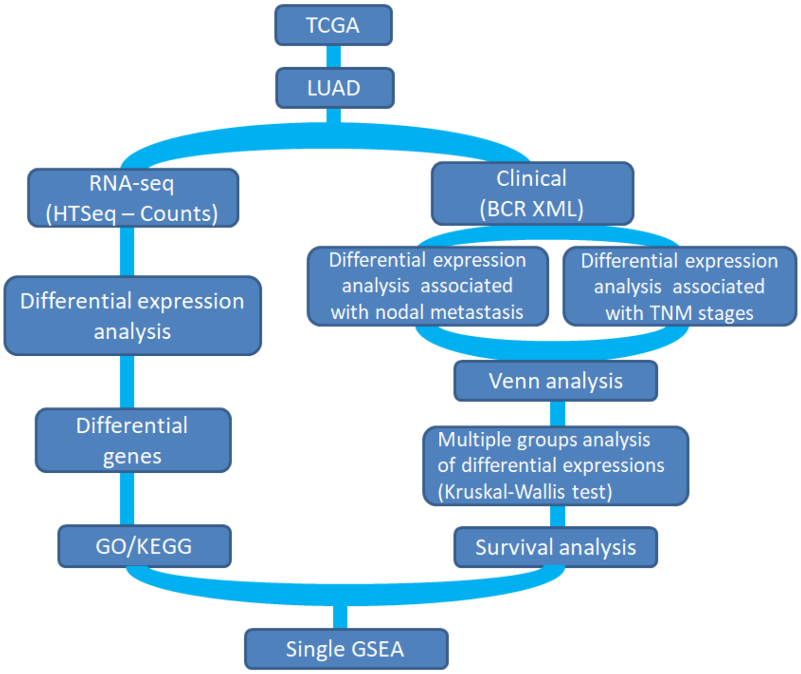
TCGA
LUAD
RNA-seq (HTSeq – Counts)
Clinical (BCR XML)
Differential expression analysis
Differential expression analysis associated with nodal metastasis
Differential expression analysis associated with TNM stages
Venn analysis
Differential genes
Multiple groups analysis of differential expressions (Kruskal-Wallis test)
GO/KEGG
Survival analysis
Single GSEA

Figure 1. The pipeline of this study.

5.2. Differential Gene Expression Analysis

The differential expressions of RNA-seq were analyzed using edgeR package [14]. It used empirical Bayesian estimation and accurate tests based on the negative binomial distributions. As edgeR suggested, genes with very low reads were often not interested in differential expression analysis; therefore, the average count-per-million (CPM) was an important criterion which could define whether a gene is reasonably expressed. Then, the package reported log2 (fold change), log2 (counts per million), and corresponding statistical significant and their corresponding error discovery rates. The differential expression genes with upregulation or downregulation were selected based on these parameters.

5.3. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway Analysis

The GO provides a platform for assorting genes or their products hierarchically into terms. These terms fall into three categories: molecular functions (the molecular activity), cellular component (the functional gene products), and biological processes (the cellular or physiological effects) [15, 16]. The DAVID 6.7 was used to perform the functional annotation analysis [17], the ggplot2 and the GOplot R packages were used to view the results.

Then we used two methods including the Kobas algorithm [18] and clusterProfiler R package to analyze the KEGG pathway [19] of differential expression genes respectively. The significant upward and downward differential expression genes from LUAD RNA-seq were analyzed, and *P* value less than 0.05 was considered as the screening criterion.

5.4. Gene Set Variation Analysis (GSVA) of KEGG Pathways

A comprehensive human gene annotations document (c5.all.v5.2.symbols.gmt) for the GO function category was downloaded from the Molecular Signatures Database (MSigDB) [20]. To reduce mRNA-SEQ data from transcriptional abundance of gene level to transcriptional activity index of gene function level, Gene Set Variation Analysis (GSVA) algorithm [21] was carried out according to enrichment scores.

5.5. Kruskal-Wallis Test

In the differential expression analysis associated with cancer metastasis or TNM stages, the clinical data like lymph node metastasis and TNM stages were selected. The Kruskal-Wallis tests were used to perform the differential expressions in the among multiple cancer groups (N0, N1, N2, and maybe N3; or stage I, II, III, and IV). The Kruskal-Wallis test by grade is a nonparametric substitution method for one-way ANOVA, and this method expands the double-sample Wilcoxon test in the case of more than two groups [22] (see below).

$$P = \frac{1}{s^2}\left[\sum_{i=1}^{k} \frac{R_i}{n_i} - N\frac{(N+1)^2}{4}\right] \text{…….…………………..(1)}$$

*s*²: the sample variance; *k*: number of groups; *R*$_i$: the total for the ith row; *n*$_i$: the size of the ith group; *N*: the total number of observations.

5.6. Survival Analyses

Two risk groups were established according to the cut-off value derived from the median of the corresponding gene expressions in the analysis of the associations of patient prognosis with gene expressions. The Kaplan-Meier algorithm and log-rank test were carried out to evaluate the survival differences between the two risk groups, and a P value less than 0.05 was considered to be statistically significant.

5.7. Gene Set Enrichment Analysis (GSEA) and Single Gene Set Enrichment Analysis (Single-GSEA)

GSEA assesses genomic level expression data. According to the median of the hub gene expression (high and low expressions), 515 lung cancer samples from the RNA-seq were divided into two groups. These two groups of GSEA were used to identify the potential function of the hub gene and the annotated c5.all.v6.2.symbols.gmt was selected as the reference gene sets. The difference at the nominal $P < 0.05$, FDR < 0.05 and the enrichment score (ES) > 0.6 were defined as the cutoff standard.

The single gene “RN7SL494P” (found it related to metastasis and prognosis in this study) related gene sets from Molecular Signatures Database (MSigDB) [23] was used to decide whether the sets show statistical difference comparing the low and the high expression categories with java-dependent GSEA 3.0 software package [24].

Authors’ contributions

Conceptualization: XZ; Formal analysis, XZ; Methodology, XZ; Project administration, YX and HL; Software, XZ and YH; Supervision, YX and HL; Validation, XZ; Writing—original draft, XZ; Writing—review and editing, XZ and YH. All authors read and approved the final manuscript.

Author details

1 The Marine Biomedical Research Institute, Guangdong Medical University, Zhanjiang, China; 2 Computational Systems Biology Lab (CSBL), Institute of Bioinformatics, University of Georgia, Athens, GA USA

Acknowledgements

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Available of data and materials

The datasets used and/or analyzed during the current study are available from the first author on reasonable request.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Funding

This work was supported partly by National Natural Science Foundation of China (81541153); Guangdong Provincial Science and Technology Programs (2016A050503046 and 2015A050502048); The Public Service Platform of South China Sea for R&D Marine Biomedicine Resources (GDMUK201808).

References

1. Casali, C.; Rossi, G.; Marchioni, A.; Sartori, G.; Maselli, F.; Longo, L.; Tallarico, E.; Morandi, U., A single institution-based retrospective study of surgically treated bronchioloalveolar adenocarcinoma of the lung: clinicopathologic analysis, molecular features, and possible pitfalls in routine practice. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer* **2010,** 5, (6), 830-6.
2. Devarakonda, S.; Morgensztern, D.; Govindan, R., Genomic alterations in lung adenocarcinoma. *The Lancet. Oncology* **2015,** 16, (7), e342-51.
3. Watanabe, N.; Ishii, T.; Takahama, T.; Tadokoro, A.; Kanaji, N.; Dobashi, H.; Bandoh, S., Anaplastic lymphoma kinase gene analysis as a useful tool for identifying primary unknown metastatic lung adenocarcinoma. *Internal medicine* **2014,** 53, (23), 2711-5.
4. Yu, Y.; Jian, H.; Shen, L.; Zhu, L.; Lu, S., Lymph node involvement influenced by lung adenocarcinoma subtypes in tumor size </=3 cm disease: A study of 2268 cases. *European journal of surgical oncology : the journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology* **2016,** 42, (11), 1714-1719.
5. McCain, J., The cancer genome atlas: new weapon in old war? *Biotechnology healthcare* **2006,** 3, (2), 46-51B.
6. Planchard, D.; Smit, E. F.; Groen, H. J. M.; Mazieres, J.; Besse, B.; Helland, A.; Giannone, V.; D'Amelio, A. M., Jr.; Zhang, P.; Mookerjee, B.; Johnson, B. E., Dabrafenib plus trametinib in patients with previously untreated BRAF(V600E)-mutant metastatic non-small-cell lung cancer: an open-label, phase 2 trial. *The Lancet. Oncology* **2017,** 18, (10), 1307-1316.
7. Walter, P.; Blobel, G., Signal recognition particle contains a 7S RNA essential for protein translocation across the endoplasmic reticulum. *Nature* **1982,** 299, (5885), 691-8.

8. Zwieb, C.; van Nues, R. W.; Rosenblad, M. A.; Brown, J. D.; Samuelsson, T., A nomenclature for all signal recognition particle RNAs. *Rna* **2005,** 11, (1), 7-13.

9. Castle, J. C.; Armour, C. D.; Lower, M.; Haynor, D.; Biery, M.; Bouzek, H.; Chen, R.; Jackson, S.; Johnson, J. M.; Rohl, C. A.; Raymond, C. K., Digital genome-wide ncRNA expression, including SnoRNAs, across 11 human tissues using polyA-neutral amplification. *PloS one* **2010,** 5, (7), e11779.

10. Peluso, P.; Herschlag, D.; Nock, S.; Freymann, D. M.; Johnson, A. E.; Walter, P., Role of 4.5S RNA in assembly of the bacterial signal recognition particle with its receptor. *Science* **2000,** 288, (5471), 1640-3.

11. Zhang, X.; Kung, S.; Shan, S. O., Demonstration of a multistep mechanism for assembly of the SRP x SRP receptor complex: implications for the catalytic role of SRP RNA. *Journal of molecular biology* **2008,** 381, (3), 581-93.

12. Chen, K.; Wang, Y.; Sun, J., A statistical analysis on transcriptome sequences: The enrichment of Alu-element is associated with subcellular location. *Biochemical and biophysical research communications* **2018,** 499, (3), 397-402.

13. Yang, Y.; Cheng, J.; Ren, H.; Zhao, H.; Gong, W.; Shan, C., Tumor FOXP3 represses the expression of long noncoding RNA 7SL. *Biochemical and biophysical research communications* **2016,** 472, (3), 432-6.

14. Robinson, M. D.; McCarthy, D. J.; Smyth, G. K., edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **2010,** 26, (1), 139-40.

15. Gene Ontology, C.; Blake, J. A.; Dolan, M.; Drabkin, H.; Hill, D. P.; Li, N.; Sitnikov, D.; Bridges, S.; Burgess, S.; Buza, T.; McCarthy, F.; Peddinti, D.; Pillai, L.; Carbon, S.; Dietze, H.; Ireland, A.; Lewis, S. E.; Mungall, C. J.; Gaudet, P.; Chrisholm, R. L.; Fey, P.; Kibbe, W.

A.; Basu, S.; Siegele, D. A.; McIntosh, B. K.; Renfro, D. P.; Zweifel, A. E.; Hu, J. C.; Brown, N. H.; Tweedie, S.; Alam-Faruque, Y.; Apweiler, R.; Auchinchloss, A.; Axelsen, K.; Bely, B.; Blatter, M.; Bonilla, C.; Bouguerleret, L.; Boutet, E.; Breuza, L.; Bridge, A.; Chan, W. M.; Chavali, G.; Coudert, E.; Dimmer, E.; Estreicher, A.; Famiglietti, L.; Feuermann, M.; Gos, A.; Gruaz-Gumowski, N.; Hieta, R.; Hinz, C.; Hulo, C.; Huntley, R.; James, J.; Jungo, F.; Keller, G.; Laiho, K.; Legge, D.; Lemercier, P.; Lieberherr, D.; Magrane, M.; Martin, M. J.; Masson, P.; Mutowo-Muellenet, P.; O'Donovan, C.; Pedruzzi, I.; Pichler, K.; Poggioli, D.; Porras Millan, P.; Poux, S.; Rivoire, C.; Roechert, B.; Sawford, T.; Schneider, M.; Stutz, A.; Sundaram, S.; Tognolli, M.; Xenarios, I.; Foulgar, R.; Lomax, J.; Roncaglia, P.; Khodiyar, V. K.; Lovering, R. C.; Talmud, P. J.; Chibucos, M.; Giglio, M. G.; Chang, H.; Hunter, S.; McAnulla, C.; Mitchell, A.; Sangrador, A.; Stephan, R.; Harris, M. A.; Oliver, S. G.; Rutherford, K.; Wood, V.; Bahler, J.; Lock, A.; Kersey, P. J.; McDowall, D. M.; Staines, D. M.; Dwinell, M.; Shimoyama, M.; Laulederkind, S.; Hayman, T.; Wang, S.; Petri, V.; Lowry, T.; D'Eustachio, P.; Matthews, L.; Balakrishnan, R.; Binkley, G.; Cherry, J. M.; Costanzo, M. C.; Dwight, S. S.; Engel, S. R.; Fisk, D. G.; Hitz, B. C.; Hong, E. L.; Karra, K.; Miyasato, S. R.; Nash, R. S.; Park, J.; Skrzypek, M. S.; Weng, S.; Wong, E. D.; Berardini, T. Z.; Huala, E.; Mi, H.; Thomas, P. D.; Chan, J.; Kishore, R.; Sternberg, P.; Van Auken, K.; Howe, D.; Westerfield, M., Gene Ontology annotations and resources. *Nucleic acids research* **2013,** 41, (Database issue), D530-5.

16. Xu, Y.; Guo, M.; Shi, W.; Liu, X.; Wang, C., A novel insight into Gene Ontology semantic similarity. *Genomics* **2013,** 101, (6), 368-75.

17. Huang da, W.; Sherman, B. T.; Lempicki, R. A., Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research* **2009,** 37, (1), 1-13.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2019 doi:10.20944/preprints201903.0064.v1

18. Xie, C.; Mao, X.; Huang, J.; Ding, Y.; Wu, J.; Dong, S.; Kong, L.; Gao, G.; Li, C. Y.; Wei, L., KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic acids research* **2011,** 39, (Web Server issue), W316-22.

19. Ogata, H.; Goto, S.; Sato, K.; Fujibuchi, W.; Bono, H.; Kanehisa, M., KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic acids research* **1999,** 27, (1), 29-34.

20. Liberzon, A.; Birger, C.; Thorvaldsdottir, H.; Ghandi, M.; Mesirov, J. P.; Tamayo, P., The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell systems* **2015,** 1, (6), 417-425.

21. Hanzelmann, S.; Castelo, R.; Guinney, J., GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC bioinformatics* **2013,** 14, 7.

22. Katz, B. M.; McSweeney, M., A Multivariate Kruskal-Wallis Test With Post Hoc Procedures. *Multivariate behavioral research* **1980,** 15, (3), 281-97.

23. Liberzon, A.; Subramanian, A.; Pinchback, R.; Thorvaldsdottir, H.; Tamayo, P.; Mesirov, J. P., Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **2011,** 27, (12), 1739-40.

24. Subramanian, A.; Tamayo, P.; Mootha, V. K.; Mukherjee, S.; Ebert, B. L.; Gillette, M. A.; Paulovich, A.; Pomeroy, S. L.; Golub, T. R.; Lander, E. S.; Mesirov, J. P., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **2005,** 102, (43), 15545-50.