Internship Report

Ding Tong

Duration: 2019.07.19 – 2019.08.29

Summary:

- Learning:
  - I learnt Linux system, Shell programming, Python, Bioconductor, UCSC genome browser, and TCGA database from zero, and now I'm familiar with all of these.
- Work:
  - For practice, processing data from the raw data RNA-Seq (fastq files) to reads count files was done using part of RNA-Seq data from mouse and one relatively small chromosome file from mm10 genome file.
  - RNA-Seq data from Zhang, etl.al (2018) was used to perform the differential expression analysis after processing the raw data to reads count files. A volcano plot was also adopted to do the data visualization. The details of this work are shown below.
  - TCGA data extraction and processing was also done. The details are described below.

Main Work

1. Processing data from RNA-Seq

Methods & Materials

1) Data

a. reads

OE1_R1.fq.gz OE1_R2.fq.gz
OE2_R1.fq.gz OE2_R2.fq.gz
WT1_R1.fq.gz WT1_R2.fq.gz
WT2_R1.fq.gz WT2_R2.fq.gz

b. Reference genome & annotation files

Download from UCSC: hg19_V19

c. Chromosome sizes

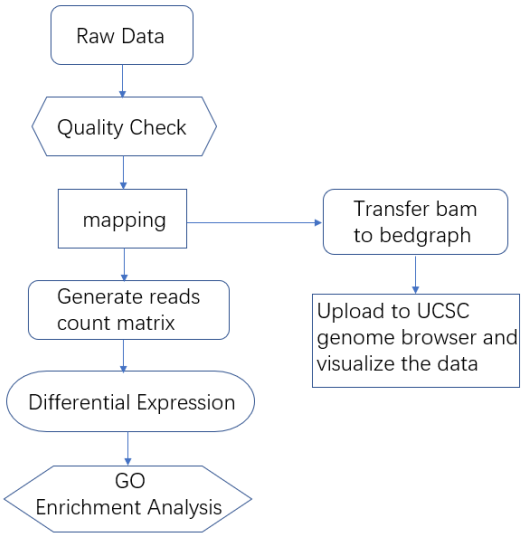
In the genome index folder STAR built

2) Software

a. quality check: fastqc
b. reads mapping: STAR
c. stranded pattern detection: infer_expriment.py from RSeQC
d. reads count matrix generation: featurecounts from RSubread from Bioconductor
e. file manipulation: SAMTools

f. file manipulation: BEDTools
g. data visualization: UCSC genome browser
h. differential expression analysis: DESeq2 from Bioconductor
i. Volcano Plot: package Cairo on R
j. GO enrichment analysis: David
k. Plot of GO results: packages GOplot and ggpubr on R

3) Workflow



4) Computing resource

'RNA3' server in Zhou lab

Results

The results of differential expression analysis from DESeq2 are shown in Table 1. The selection criteria was set as the absolute value of log 2 fold change > 1, p-adjusted < 0.05. To better visualize the results, a volcano plot was shown in Fig 1. Gene Ontology (GO) enrichment analysis was then performed using up and down regulated genes respectively (Fig.2).

a. up-regulation

Row.names	baseMean	log2FoldC	lfcSE	stat	pvalue	padj	OE_1	OE_2	WT_1	WT_2
ENST00000487273.2	1054.44	2.15142	0.30991	6.94201	3.87E-12	2.15E-08	435.929	339.743	2028.43	1413.65
ENST00000264832.3	870.572	1.62385	0.30629	5.30171	1.15E-07	0.00032	501.941	352.234	1432.15	1195.97
ENST00000342505.4	12443.4	1.36911	0.27862	4.91386	8.93E-07	0.00151	7987.47	5904.71	19409.7	16471.6
ENST00000409652.4	749.307	1.40552	0.29315	4.79462	1.63E-06	0.00151	447.139	374.717	1180.74	994.63
ENST00000358399.3	869.92	1.48819	0.31996	4.65118	3.30E-06	0.0023	438.42	475.474	1585.78	980.003
ENST00000315436.3	71.9516	2.43257	0.5279	4.60798	4.07E-06	0.00251	22.4192	22.483	152.562	90.3427
ENST00000227507.2	2841.2	1.33623	0.29392	4.54628	5.46E-06	0.00304	1925.56	1299.85	4354.45	3784.93
ENST00000237289.4	805.809	1.31469	0.3001	4.38077	1.18E-05	0.00548	538.061	387.208	1191.48	1106.48
ENST00000371933.3	514.69	1.41983	0.34431	4.12366	3.73E-05	0.01383	269.031	290.614	951.899	547.219
ENST00000245907.6	164.193	1.61617	0.39785	4.06226	4.86E-05	0.01674	100.886	61.6201	239.586	254.68
ENST00000373970.3	881.347	1.19254	0.29622	4.0258	5.68E-05	0.01674	622.756	451.326	1218.34	1232.96
ENST00000380615.3	247.969	1.48351	0.36661	4.04655	5.20E-05	0.01674	107.114	153.218	424.379	307.165
ENST00000397128.2	297.532	1.51741	0.38203	3.97196	7.13E-05	0.01889	133.27	174.035	570.495	312.328
ENST00000371873.5	274.143	1.50675	0.38976	3.86583	0.00011	0.02567	169.39	116.579	518.924	291.678
ENST00000216117.8	932.641	1.14356	0.29719	3.84787	0.00012	0.02652	676.313	487.132	1341.9	1225.22
ENST00000316059.6	531.743	1.19634	0.321	3.72696	0.00019	0.03595	331.306	314.762	886.362	594.541

b. down-regulation

Row.nam	baseMean	log2FoldC	lfcSE	stat	pvalue	padj	OE_1	OE_2	WT_1	WT_2
ENST0000	1764.81	-1.6065	0.33397	-4.8103	1.51E-06	0.00151	3486.19	1828.62	848.758	895.683
ENST0000	7199.2	-1.459	0.30185	-4.8335	1.34E-06	0.00151	12990.7	8125.53	4146.02	3534.55
ENST0000	625.394	-1.5574	0.33102	-4.705	2.54E-06	0.00202	1152.1	715.293	345.95	288.236
ENST0000	2574.34	-1.2008	0.2715	-4.4228	9.74E-06	0.00493	3986.88	3189.26	1532.06	1589.17
ENST0000	11266.9	-1.2112	0.28376	-4.2682	1.97E-05	0.00843	18041.2	13433.2	5985.36	7607.72
ENST0000	20026.1	-1.4225	0.34258	-4.1524	3.29E-05	0.01308	39587.4	18753.3	11473.3	10290.5
ENST0000	5172.82	-1.3206	0.32815	-4.0243	5.71E-05	0.01674	9712.51	5063.68	3044.79	2870.32
ENST0000	2843.08	-1.425	0.35526	-4.0113	6.04E-05	0.0168	5715.66	2571.39	1571.81	1513.46
ENST0000	3252.84	-1.425	0.36058	-3.9519	7.75E-05	0.01961	6602.46	2878.66	1849	1681.23
ENST0000	205.424	-1.5638	0.39968	-3.9127	9.13E-05	0.02208	396.073	218.169	112.81	94.6447
ENST0000	3407.86	-1.1836	0.30942	-3.8253	0.00013	0.02795	5939.85	3525.67	2008.01	2157.9
ENST0000	11224.7	-1.2173	0.32105	-3.7916	0.00015	0.03085	20411.5	10985	6942.63	6559.74
ENST0000	1895.56	-1.2925	0.34328	-3.7652	0.00017	0.03308	3560.92	1823.62	1228.01	969.678
ENST0000	380.153	-1.1877	0.31805	-3.7343	0.00019	0.03595	594.109	462.984	227.768	235.751
ENST0000	279.819	-1.3664	0.3703	-3.69	0.00022	0.04025	510.66	296.443	146.115	166.058
ENST0000	4091.91	-1.2755	0.34666	-3.6793	0.00023	0.04067	7857.94	3725.52	2559.17	2225.01
ENST0000	7945.98	-1.2948	0.35377	-3.66	0.00025	0.04254	15429.4	7151.27	5198.91	4004.33
ENST0000	1079.45	-1.1677	0.3203	-3.6457	0.00027	0.04364	1864.53	1124.15	620.99	708.115
ENST0000	3482.13	-1.0675	0.29351	-3.6369	0.00028	0.04387	5687.01	3743.01	2187.43	2311.05

Table 1. Results from DESeq2

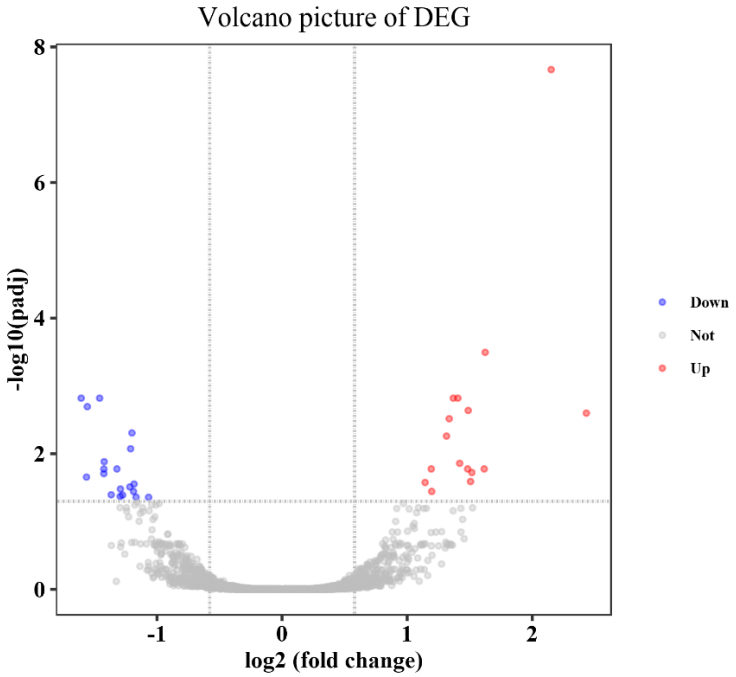
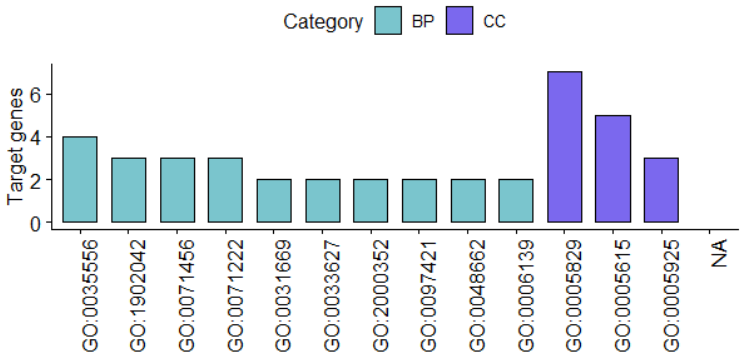
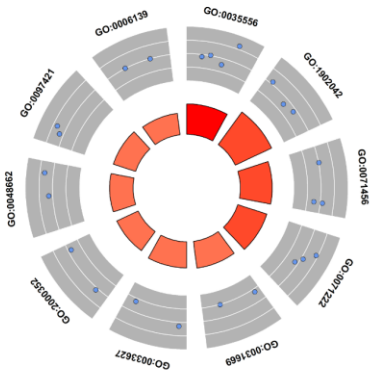


Figure 1. Volcano plot

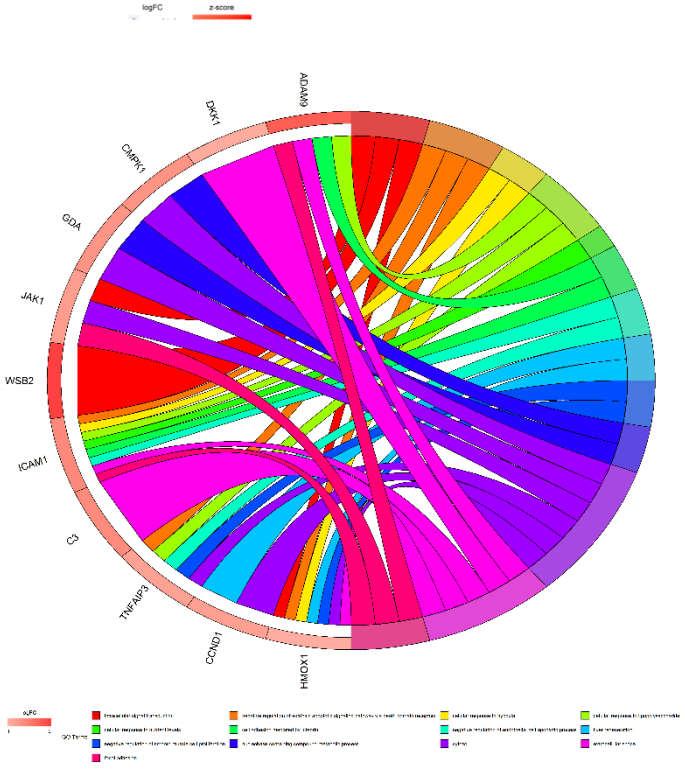


a.

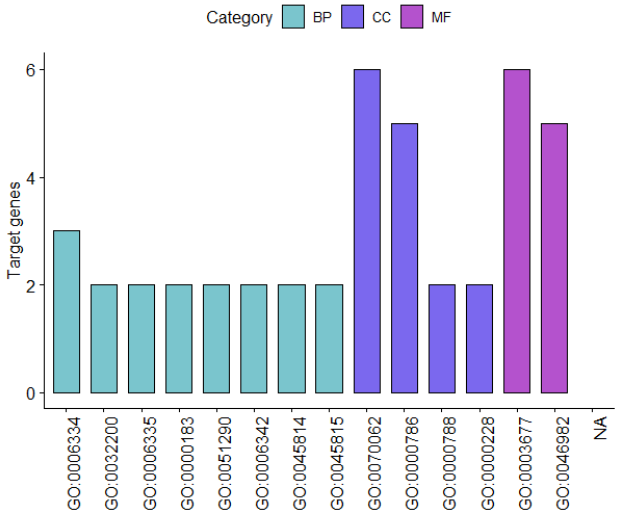


ID	Description
GO:0035556	intracellular signal transduction
GO:1902042	negative regulation of extrinsic apoptotic signaling pathway via death domain receptors
GO:0071456	cellular response to hypoxia
GO:0071222	cellular response to lipopolysaccharide
GO:0031669	cellular response to nutrient levels
GO:0033627	cell adhesion mediated by integrin
GO:2000352	negative regulation of endothelial cell apoptotic process
GO:0048662	negative regulation of smooth muscle cell proliferation
GO:0097421	liver regeneration
GO:0006139	nucleobase-containing compound metabolic process

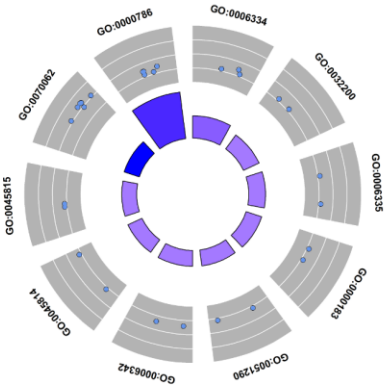
b.



c.

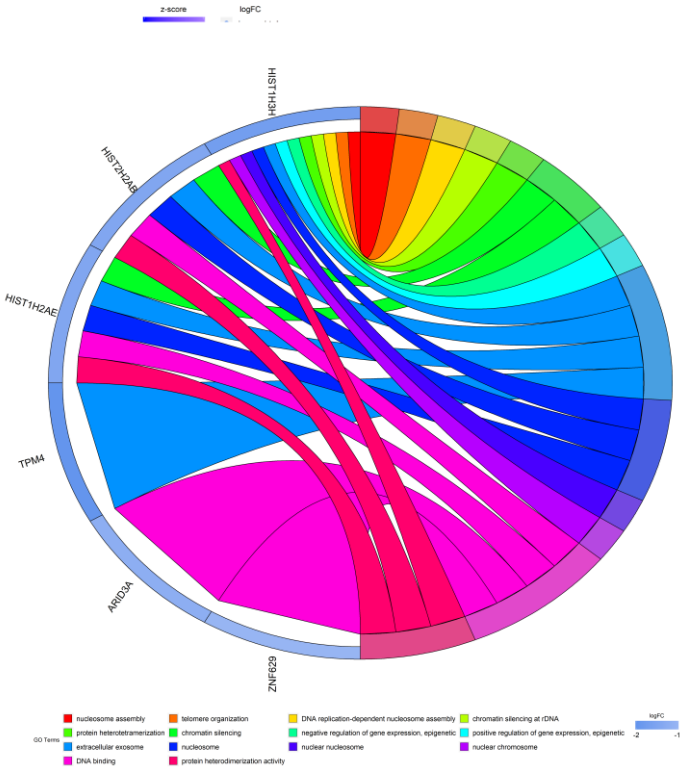


d.



ID	Description
GO:0006334	nucleosome assembly
GO:0032200	telomere organization
GO:0006335	DNA replication-dependent nucleosome assembly
GO:0000183	chromatin silencing at rDNA
GO:0051290	protein heterotetramerization
GO:0006342	chromatin silencing
GO:0045814	negative regulation of gene expression, epigenetic
GO:0045815	positive regulation of gene expression, epigenetic
GO:0070062	extracellular exosome
GO:0000786	nucleosome

e.



f.

Figure 2. Results of GO enrichment analysis. (a) – (c) Bar, circle, and chord plot of the results of GO of up-regulated genes; (d) – (f) Bar, circle, and chord plot of the results of GO of down-regulated genes

2. TCGA data extraction and processing

Methods & Materials

1) Data

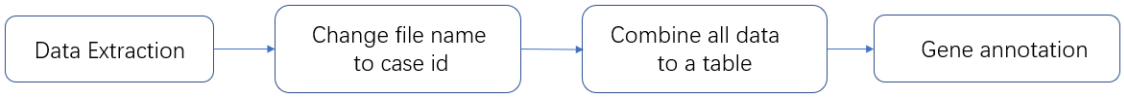
a. manifest file and metadata file
Download from TCGA database

2) Software

a. TCGA data extraction: gdc-client

b. files renaming and repeat group processing: python
c. data combination: R
d. gene annotation: biomaRt and curl packages from Bioconductor

3) Workflow



Results

The results of every step's processing are shown in Fig 2. The final result is a csv file that contains all the expression data of files downloaded from TCGA with patient id in the first row, gene symbol in the third column.

Figure 2. Results of processing step by step

Note. All the data and results are available upon request

Reference

Zhang, K., Zhang, X., Cai, Z., Zhou, J., Cao, R., Zhao, Y., ... & Liu, G. (2018). A novel class of microRNA-recognition elements that function only within open reading frames. *Nature structural & molecular biology*, *25*(11), 1019.