Research Article **SRR (2021) 14:124**

Scientific Research and Reviews (ISSN:2638-3500)

Integrative Analysis of Transcriptome and Methylation Data in Human Non-Small Cell Lung Cancer

Xiang AO

Department of Computer Science, City University of Hong Kong, Hong Kong.

ABSTRACT

Human lung cancer is the most prevalent cancer worldwide that consisting of two main subtypes: the non-small cell lung cancer (NSCLC) and the small cell lung cancer (SCLC). NSCLC comprises over 80% of lung cancer and the treatment of NSCLC is mostly guided by tumor stage, although distinctive molecular characteristics between two major subtypes of NSCLC, i.e., lung adenocarcinoma (LUAD) and squamous cell lung carcinoma (LUSC), have been increasingly identified. In this study, we integrated the gene expression data and methylation data to investigate the genetic differences between LUAD and LUSC. We further applied the Boruta package to select key features from LUAD and LUSC tumor samples to build predictive models of tumor stage. We finally obtained 6 key gene expression features and 4 key methylation features that can be reliably used in prediction of LUAD and LUSC stage.

Keywords: Transcriptome; Methylation Data; Lung Cancer

***Correspondence to Author:**
Xiang AO
Department of Computer Science, City University of Hong Kong, Hong Kong.

How to cite this article:
Xiang AO Integrative Analysis of Transcriptome and Methylation Data in Human Non-Small Cell Lung Cancer. Scientific Research and Reviews, 2021; 14:124.

eSciPub LLC, Houston, TX USA.
Website: http://escipub.com/

1. Introduction

Human lung cancer is the most common cancer and the most lethal cancer worldwide. According to statistics from World Health Organization (WHO), it is approximate 2.09 million new cases and 1.76 million deaths each year [1]. Smoking is the major risk factor for lung cancer [2-4]. It was estimated that nearly 90% of lung cancer men and 70% to 80% in women was linked to cigarette smoking [5]. Even passive smoking contributes to lung cancer [6]. Other risk factors include exposure to radon or/and asbestos, and arsenic in drinking water [7].

Lung cancer can be grouped into two subtypes according to its histological difference: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) [8]. NSCLC comprises about 80~85% of lung cancer cases while SCLC accounts for 10~15% of the cases [9]. Although NSCLC takes the major part in lung cancer composition, it's well known that NSCLC is group of distinct diseases with genetic and cellular heterogeneity [10]. In fact, NSCLC has two major subtypes: adenocarcinoma (~50%) and squamous cell carcinoma (~40%). For the convenience of later analysis, we name them as LUAD and LUSC, respectively.

High-throughput genomic profiling of LUAD and LUSC has showed their dissimilar genetic mutations [10]. In LUAD, recurrent mutations have been found *in KRAS,* b-raf proto-oncogene (*BRAF*)*,* epidermal growth factor receptor (*EGFR*)*,* erb-b2 receptor tyrosine kinase 2 (*ERBB2),* tyrosine-protein kinase met (*MET*)*,* fibroblast growth factor receptor 1 (*FGFR1*)*, FGFR2,* anaplastic lymphoma receptor tyrosine kinase (*ALK*)*,* the ROS1 receptor tyrosine kinase, neuregulin 1 (*NRG1*), neurotrophic tyrosine kinase receptor type 1 (*NTRK1*) and RET receptor tyrosine kinase (RET) [11-25]. In contrast, prevalent mutations in LUSC are discoidin domain-containing receptor 2 (*DDR2*), *FGFR1*, *FGFR2*, *FGFR3* and genes in the PI3K pathway [23].

Given the relatively large number of mutations in LUAD and LUSC, far more mutations or even epigenome modifiers could be identified for LUAD and LUSC if we do more comprehensive analyses or explore multi-omics data. The key point is about how do we rank the importance of the results and thus apply it in designing lung cancer treatment. By combining results from multi-omics analyses, a more accurate and precise interpretation of genetic basis and development of NSCLC may be achieved and will ultimately inform the most suitable therapy for individual patients.

Here we integratively analyzed the transcriptome and methylation data from a public available NSCLC cohort, i.e. the Cancer Genome Atlas (TCGA) [26], mainly focused on LUAD and LUSC. We reported a set of differentially methylated CpGs (DMCs) and differentially expressed genes (DEGs) between normal and tumor samples for each of the cancer. Based on the Integration of the DMCs and DEGs, we built predictive models for distinguishing early and late stages of LUAD and LUSC tumors.

2. Materials and Methods

TCGA is a landmark cancer genomics program, molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types. It contains two different versions of data: the legacy TCGA data and the harmonized TCGA data. Both versions are highly concordant and the major difference between them is the reference genome build they used. The legacy version used GRCh37 Genome Reference Consortium Human build 37 (GRCh37 or hg19) while the harmonized version utilized GRCh38 build (hg38). More details about the differences can be found in [27].

We downloaded the gene expression data and DNA methylation data of two projects TCGA-LUAD and TCGA-LUSC through a R package TCGAbiolinks (version 2.14) [28-30]. Specifically, we used the function GDCquery to inquire desired data. Settings of parameters is showed in Table 1. We downloaded 576 expression samples (311 females and 265 males, age range from 38 to 88) and 507 methylation samples (269 females and 238 males, age range from 33 to 88) from the TCGA-LUAD project. 553 expression samples (145 females

and 408 males, age range from 39 to 90) and 412 methylation samples (109 females and 303 males, age range from 40 to 90) were downloaded from the TCGA-LUSC project.

Table 1 Parameter settings of function GDCquery

Parameter	Expression data	Methylation data
project	“TCGA-LUAD” or “TCGA-LUSC”	
data.category	"Gene expression"	"Raw microarray data"
data.type	"Gene expression quantification"	"Raw intensities"
file.type	"results"	".idat"
experimental.strategy	"RNA-Seq"	"Methylation array"
platform	"Illumina HiSeq"	"Illumina Human Methylation 450"
legacy	TRUE	TRUE

Differential gene expression analysis on two expression datasets was similar to the analysis did in [31]. We directly adopted the R package DESeq2 [42] to obtain the results. Beside the covariates age and gender, we added surrogate variables in DE analysis to reduce the impact from batch effects and other unwanted variations. R package sva (version 3.34.0) [32] was employed to obtain surrogate variables. Significant DE genes were filtered out with threshold $|\log_2(FC) \geq 1|$ and Bonferroni adjusted pvalue ≤ 0.05.

Differential methylated CpGs was analyzed with the dmpFinder function from R package minfi (version 1.32.0) [32-39]. Significant DMC was selected if its beta value difference between tumor and normal samples was larger than 0.2 and Bonferroni adjusted pvalue was less than 0.05.

Gene ontology enrichment analysis was complete with R package clusterProfiler (version 3.14.0) [43].

We applied Boruta [40] to select features for building the predictive models of tumor stages. There were over 430k methylation probes and nearly 20k expression genes. The number of features were too large to deal with. We used the DMCs, DEGs and their normalized M values and expression levels as input to Boruta because those candidate DMCs and DEGs are potentially contribute to tumor development.

We applied Support Vector Machine(SVM) from R package e1071(version 1.7-6) [41] to train models. Here we grouped stageI and stageII tumors as early stage, while grouped stageIII and stageIV tumors as late stage. We hereby selected 277 early stage and 69 late stage LUAD samples and 232 early stage and 39 late stage of LUSC samples. 20% samples of each stage were left out for consequent model testing.

3. Results

Differential gene expression analysis on TCGA-LUAD project identified 4131 DE genes of which 2101 genes were down-regulated and 2030 genes were up-regulated. In contrast, the analysis in TCGA-LUSC project discovered 4878 DE genes of which expression of 2449 genes and 2429 genes were repressed and overexpressed, respectively. Figure 1 shows the volcano plots of the DE results from the two cohorts.

Gene ontology enrichment analysis was done on significant DE genes of the two projects. Top 5 significantly enriched GO terms from the TCGA-LUAD were cell-substrate adhesion (GO:0031589), Ras protein signal transduction (GO:0007265), hemostasis (GO:0007599), DNA replication (GO:0006260), and blood coagulation (GO:0007596). The top 5 from the TCGA-LUSC were DNA-dependent DNA replication (GO:0006261), neutrophil mediated immunity (GO:0002446), DNA replication (GO:0006260), neutrophil activation involved in immune response (GO:0002283), and neutrophil activation (GO:0042119). Figure 2 shows the most 20 significant GO terms in a barplot.

Analysis of differential methylated CpGs was taken on the two projects as well. We found 451 9 DMCs and 4246 DMCs from the TCGA-LUAD

project and the TCGA-LUSC project, DMCs found from both projects. respectively. Notably, there were 1402 common

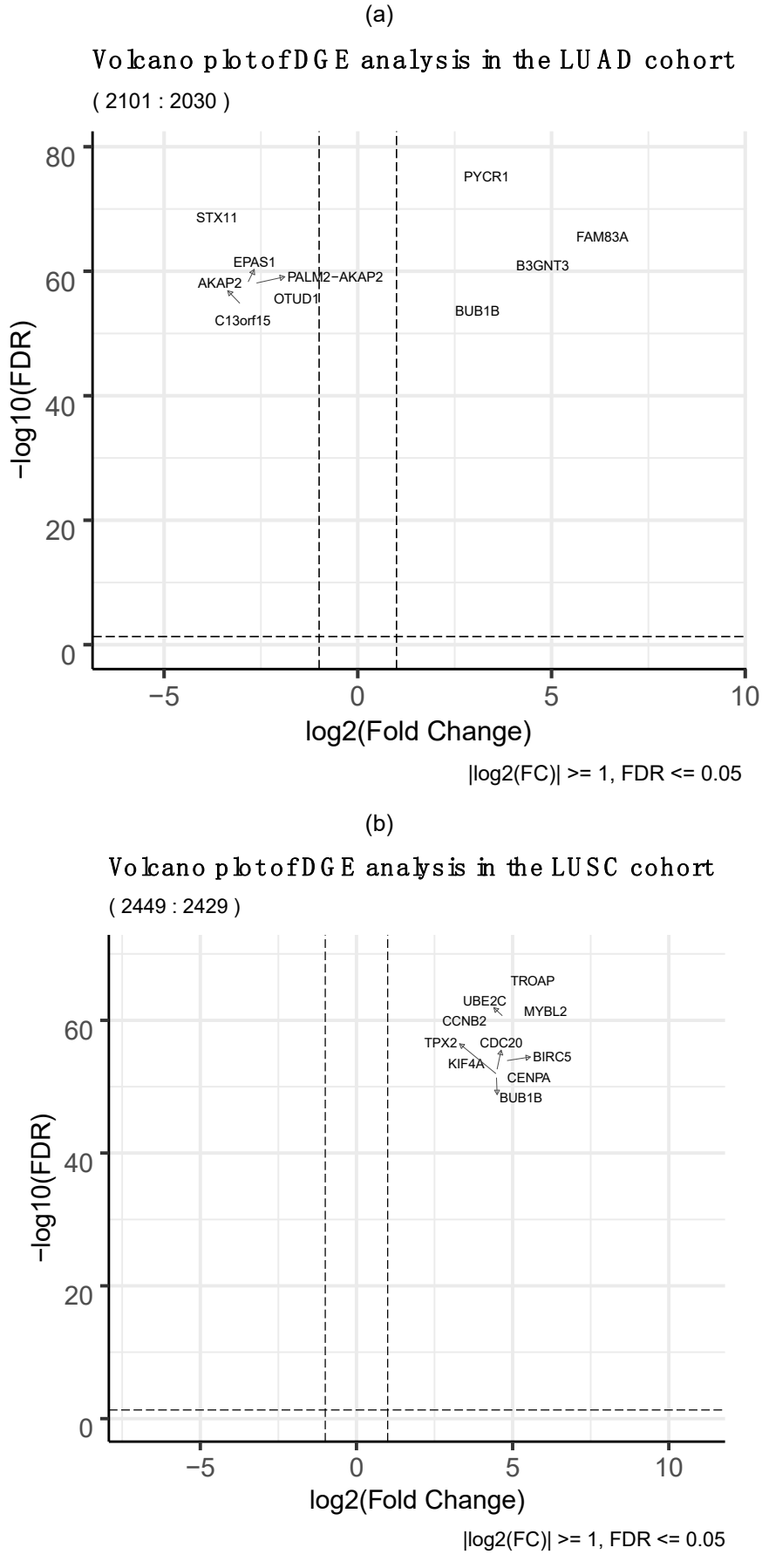


Figure 1 Volcano plots of DGE analysis on the two cohorts.

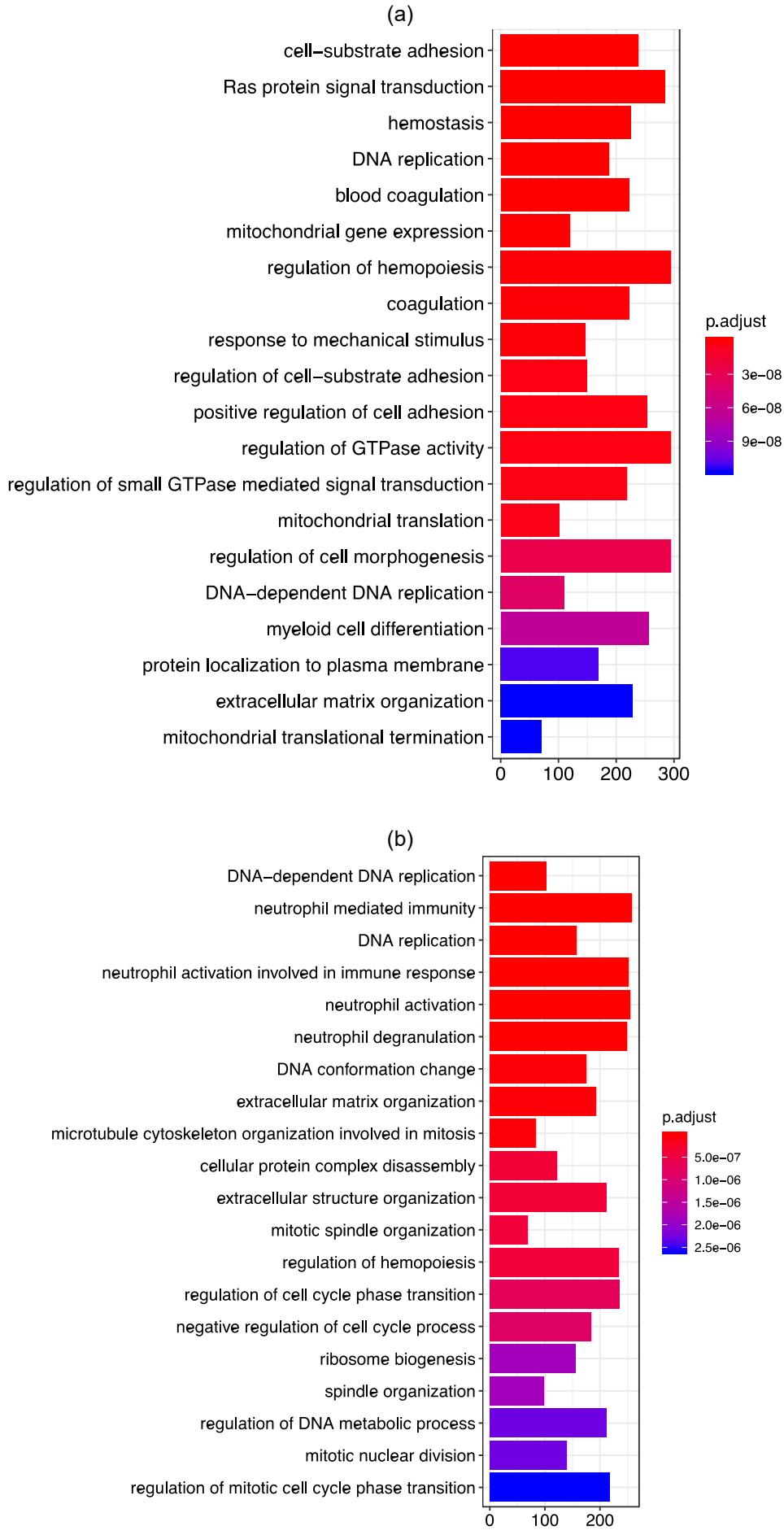
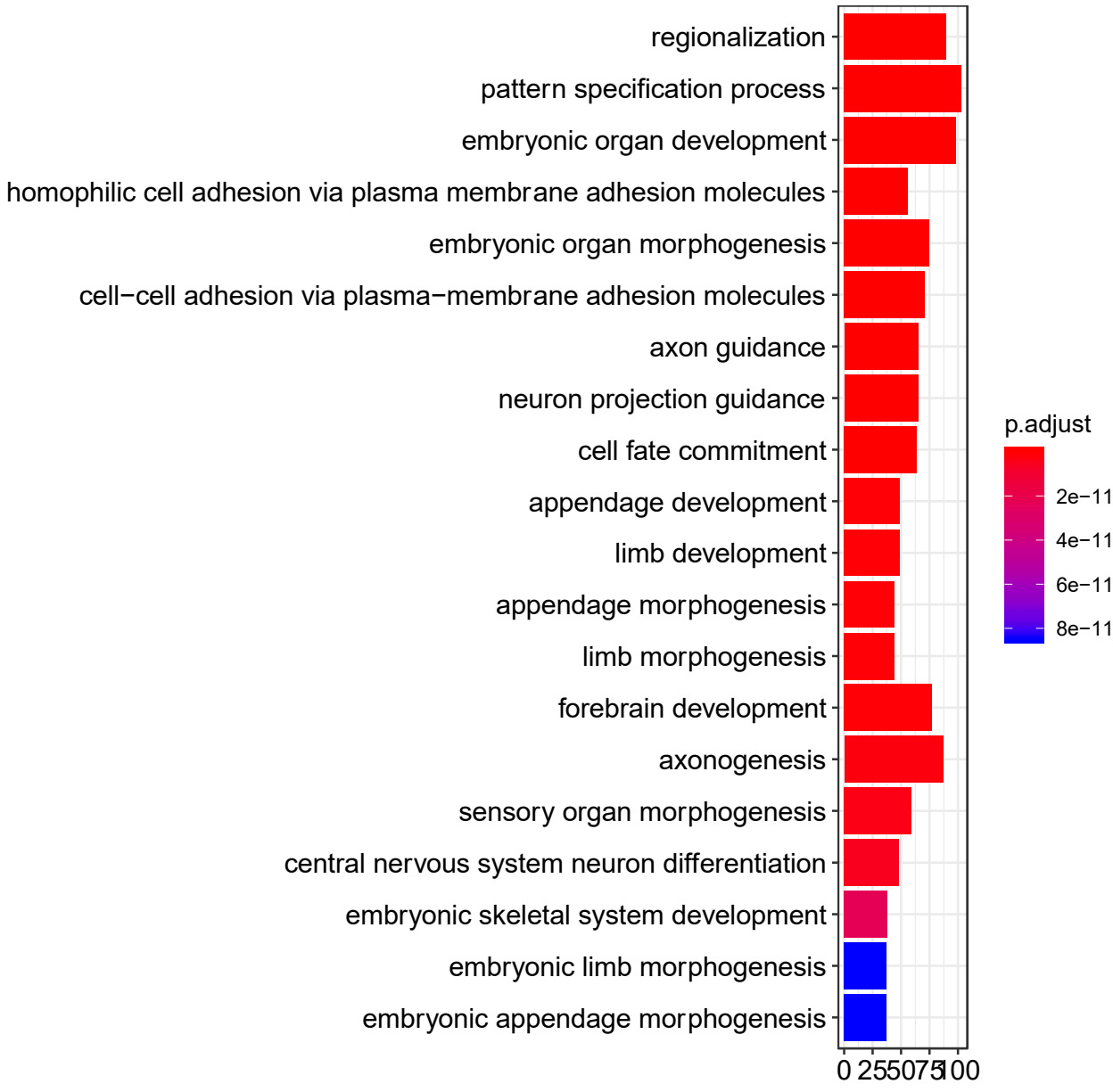


Figure 2 A barplot shows the most 20 enriched GO terms from DE analysis on the TCGA-LUAD project (a) and the TCGA-LUSC project (b).

To do the enrichment analysis on the results of DMC analysis, we mapped the significant DMCs to its nearest gene and then did the GO analysis. The top 5 enriched GO terms in TCGA-LUAD were: regionalization (GO:0003002), pattern specification process (GO:0007389), embryonic organ development (GO:0048568), homophilic cell adhesion via plasma membrane adhesion molecules(GO:0007156), and embryonic organ morphogenesis (GO:0048562). In contrast, the 5 most enriched GO terms in TCGA-LUSC were: embryonic organ development (GO:0048568), muscle tissue development (GO:0060537), cell fate commitment (GO:0045165), embryonic organ morphogenesis (GO:0048562), and muscle organ development (GO:0007517). Figure 3 presents the top 20 enriched GO terms in barplots.

Boruta found 5 key features both for LUAD and LUSC samples from their methylation data and RNA-seq expression data. Table 2 listed the 10 key features.

(a)



(b)

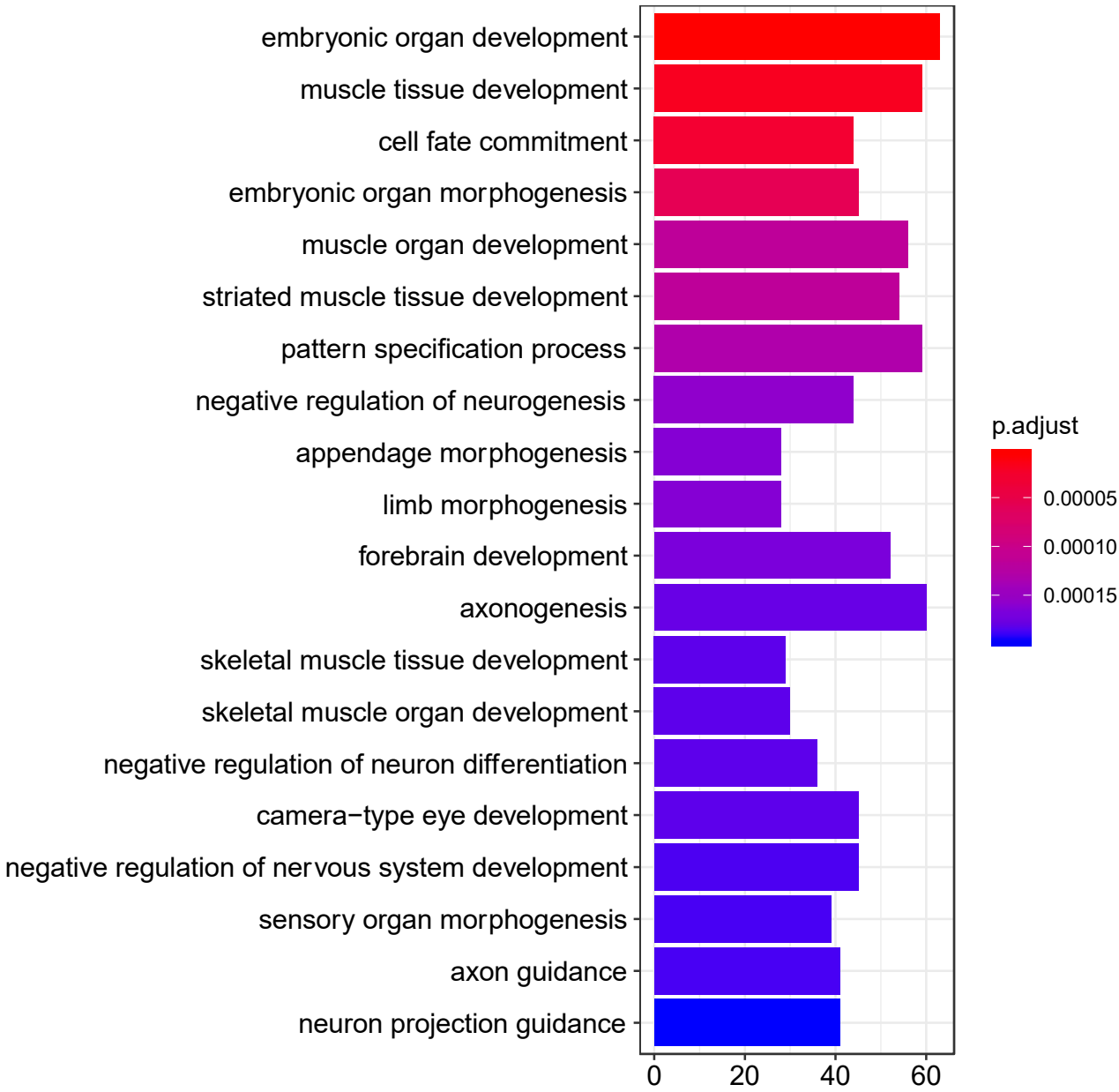


Figure 3 The most 20 enriched GO terms from DMC analysis on the TCGA-LUAD project (a) and the TCGA-LUSC project (b).

Table 2. Key features found from LUAD and LUSC

Cohort	Name	Chr	Start*	End	Strand	Ref Gene#
LUAD	cg18595065	5	72597126		+	TMEM174
LUAD	cg26335602	6	28129616		-	ZNF389
LUAD	CHRNB4	15	78916636	78952874	-	
LUAD	C1QL1	17	43037062	43045644	-	
LUAD	FAM136A	2	70523109	70529220	-	
LUSC	cg00242951	3	9596482		+	LHFPL4
LUSC	cg02551745	3	156809115		+	LEKR1
LUSC	HOMER3	19	19017769	19052041	-	
LUSC	MAPK8IP2	22	51039131	51049978	+	
LUSC	KLHDC7B	22	50986462	50989451	+	

* CpG position was showed in this column

The nearest genes were showed for CpGs

We further used SVM to train models for prediction of the tumor stage with the 5 key features and 20% samples to test the models. We achieved an accuracy of 0.8 and 0.845 in LUAD and LUSC tumor samples.

4. Discussion

Lung cancer is the most common cancer and the most lethal cancer worldwide. Every year, it's estimated that there are more than 2 million new cases and 1.5 million mortality. Non-small cell lung cancer predominates the cancer and contains two common subtypes: adenocarcinoma and squamous cell carcinoma. Here we investigated the transcriptome and DNA methylation of two public LUAD and LUSC projects. After applying differential gene expression analysis and differential methylation CpG analysis on both projects' data, we reported genes and pathways that potentially contribute to NSCLC pathogenesis.

Differential gene expression analysis results showed the expression patterns of LUAD and LUSC. In LUAD, DGE analysis found 4131 DE genes totally while 4878 DE genes were discovered in LUSC. Figure 4.4 shows a Venn diagram between DE genes of the LUAD and the LUSC. DE genes were further divided into upregulated genes and downregulated genes. As we can see, the two projects shared a large part of common DE genes (42%, 2666/6319), 99.3% of which altered at the same direction. This clue shows that LUAD and LUSC are two heterogonous cancers, yet they shared some common traits or some uniform genetic factors underlaid the progress of LUAD and LUSC.

The most 20 enriched GO terms from DE genes in the LUAD project were largely different from those in the LUSC project, with only one single term shared. It indicated that the biological causes of LUAD and LUSC were greatly distinct. Treatment to LUAD and LUSC should be designed specifically.

Differential methylation CpG analysis on two projects revealed the abnormal methylation regions in tumor samples. 4519 DMCs and 4246 DMCs in total were found from the TCGA-LUAD project than the TCGA-LUSC project, respectively. Almost a third of the DMCs were found from both projects. It denotes, on one hand, LUAD and LUSC tumors are both originated from lung tissue so that they share some common methylation patterns, on the other hand, distinct methylation patterns contribute to the development of LUAD and LUSC.

By integrative analysis of the methylation and expression data of LUAD and LUSC tumors, we found 6 key genes and 4 key methylation site that potentially contribute to the development of NSCLC tumors.

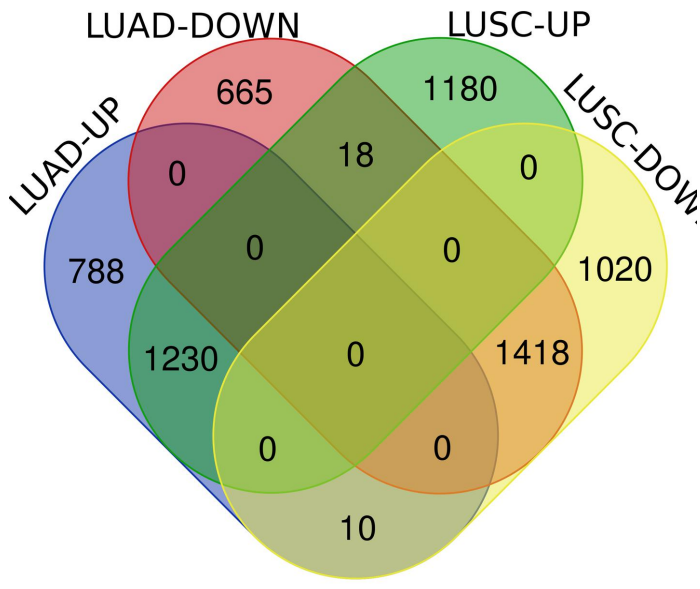


Figure 4 A Venn diagram of DE genes.
The postfix "up" represents up-regulated genes and the postfix "down" means down-regulated genes.

5. Acknowledgement

The results shown here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Reference

[1] Cancer. Retrieved from https://www.who.int/news-room/fact-sheets/detail/cancer

[2] Cersosimo, Robert J. "Lung cancer: a review." American journal of health-system pharmacy 59.7 (2002): 611-642.

[3] Rahal, Zahraa, et al. "Smoking and lung cancer: a geo-regional perspective." Frontiers in oncology 7 (2017): 194.

[4] Jeon, Jihyoun, et al. "Smoking and lung cancer mortality in the United States from 2015 to 2065: a comparative modeling approach." Annals of internal medicine 169.10 (2018): 684-693.

[5] Walser, Tonya, et al. "Smoking and lung cancer: the role of inflammation." Proceedings of the American Thoracic Society 5.8 (2008): 811-815.

[6] Du, Yihui, et al. "Lung cancer occurrence attributable to passive smoking among never smokers in China: a systematic review and meta-analysis." Translational Lung Cancer Research 9.2 (2020): 204.

[7] Lung Cancer Risk Factors. Retrieved from https://www.cancer.org/cancer/lung-cancer/causes-risks-prevention/risk-factors.html

[8] Travis, William D., et al. "The 2015 World Health Organization classification of lung tumors: impact of genetic, clinical and radiologic advances since the 2004 classification." Journal of thoracic oncology 10.9 (2015): 1243-1260.

[9] What is lung cancer. Retrieved from https://www.cancer.org/cancer/lung-cancer/about/what-is.html

[10] Chen, Zhao, et al. "Non-small-cell lung cancers: a heterogeneous set of diseases." Nature Reviews Cancer 14.8 (2014): 535-546.

[11] Davies, Helen, et al. "Mutations of the BRAF gene in human cancer." Nature 417.6892 (2002): 949-954.

[12] Santos, Eugenio, et al. "Malignant activation of a K-ras oncogene in lung carcinoma but not in normal tissue of the same patient." Science 223.4637 (1984): 661-664.

[13] Lynch, Thomas J., et al. "Activating mutations in the epidermal growth factor receptor underlying responsiveness of non–small-cell lung cancer to gefitinib." New England Journal of Medicine 350.21 (2004): 2129-2139.

[14] Paez, J. Guillermo, et al. "EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy." Science 304.5676 (2004): 1497-1500.

[15] Pao, William, et al. "EGF receptor gene mutations are common in lung cancers from “never smokers” and are associated with sensitivity of tumors to gefitinib and erlotinib." Proceedings of the National Academy of Sciences 101.36 (2004): 13306-13311.

[16] Shepherd, Frances A., et al. "Erlotinib in previously treated non–small-cell lung cancer." New England Journal of Medicine 353.2 (2005): 123-132.

[17] Engelman, Jeffrey A., et al. "MET amplification leads to gefitinib resistance in lung cancer by activating ERBB3 signaling." science 316.5827 (2007): 1039-1043.

[18] Fernandez-Cuesta, Lynnette, et al. "CD74–NRG1 fusions in lung adenocarcinoma." Cancer discovery 4.4 (2014): 415-422.

[19] Kohno, Takashi, et al. "KIF5B-RET fusions in lung adenocarcinoma." Nature medicine 18.3 (2012): 375-377.

[20] Rikova, Klarisa, et al. "Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer." Cell 131.6 (2007): 1190-1203.

[21] Soda, Manabu, et al. "Identification of the transforming EML4–ALK fusion gene in non-small-cell lung cancer." Nature 448.7153 (2007): 561-566.

[22] Stephens, Philip, et al. "Intragenic ERBB2 kinase mutations in tumours." Nature 431.7008 (2004): 525-526.

[23] Cancer Genome Atlas Research Network. "Comprehensive genomic characterization of squamous cell lung cancers." Nature 489.7417 (2012): 519.

[24] Vaishnavi, Aria, et al. "Oncogenic and drug-sensitive NTRK1 rearrangements in lung cancer." Nature medicine 19.11 (2013): 1469-1472.

[25] Weiss, Jonathan, et al. "Frequent and focal FGFR1 amplification associates with therapeutically tractable FGFR1 dependency in squamous cell lung cancer." Science translational medicine 2.62 (2010): 62ra93-62ra93.

[26] Tomczak, Katarzyna, Patrycja Czerwińska, and Maciej Wiznerowicz. "The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge." Contemporary oncology 19.1A (2015): A68.

[27] Gao, Galen F., et al. "Before and after: comparison of legacy and harmonized TCGA genomic data commons’ data." Cell systems 9.1 (2019): 24-34.

[28] Colaprico, Antonio, et al. "TCGAbiolinks: an R /

Bioconductor package for integrative analysis of TCGA data." Nucleic acids research 44.8 (2016): e71-e71.

[29] Silva, Tiago C., et al. "TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages." F1000Research 5 (2016).

[30] Mounir, Mohamed, et al. "New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx." PLoS computational biology 15.3 (2019): e1006701.

[31] Ao, Xiang, and Shuaicheng Li. "TRANSCRIPT LEVEL ANALYSIS IMPROVES THE UNDERSTANDING OF BLADDER CANCER."

[32] Leek JT, Johnson WE, Parker HS, Fertig EJ, Jaffe AE, Zhang Y, Storey JD, Torres LC (2020). sva: Surrogate Variable Analysis. R package version 3.36.0.

[33] Aryee, Martin J., et al. "Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays." Bioinformatics 30.10 (2014): 1363-1369.

[34] Maksimovic, Jovana, Lavinia Gordon, and Alicia Oshlack. "SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips." Genome biology 13.6 (2012): R44.

[35] Fortin, Jean-Philippe, et al. "Functional normalization of 450k methylation array data improves replication in large cancer studies." Genome biology 15.11 (2014): 503.

[36] Triche Jr, Timothy J., et al. "Low-level processing of Illumina Infinium DNA methylation beadarrays." Nucleic acids research 41.7 (2013): e90-e90.

[37] Fortin, Jean-Philippe, and Kasper D. Hansen. "Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data." Genome biology 16.1 (2015): 180.

[38] [38] Andrews, Shan V., et al. ""Gap hunting" to characterize clustered probe signals in Illumina methylation array data." Epigenetics & chromatin 9.1 (2016): 1-21.

[39] [39] Fortin, Jean-Philippe, Timothy J. Triche Jr, and Kasper D. Hansen. "Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi." Bioinformatics 33.4 (2017): 558-560.

[40] Kursa, Miron B., and Witold R. Rudnicki. "Feature selection with the Boruta package." J Stat Softw 36.11 (2010): 1-13.

[41] Meyer, David, and FH Technikum Wien. "Support vector machines." The Interface to libsvm in package e1071 28 (2015).

[42] Love, Michael I., Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome biology 15.12 (2014): 1-21.

[43] Yu, Guangchuang, et al. "clusterProfiler: an R package for comparing biological themes among gene clusters." Omics: a journal of integrative biology 16.5 (2012): 284-287.