RESEARCH ARTICLE

Comparison of the Gene Expression Profiles Between Smokers With and Without Lung Cancer Using RNA-Seq

Peng Cheng[1&], You Cheng[2&], Yan Li[1], Zhenguo Zhao[1], Hui Gao[1], Dong Li[1], Hua Li[1], Tao Zhang[1]*

Abstract

Lung cancer seriously threatens human health, so it is important to investigate gene expression changes in affected individuals in comparison with healthy people. Here we compared the gene expression profiles between smokers with and without lung cancer. We found that the majority of the expressed genes (threshold was set as 0.1 RPKM) were the same in the two samples, with a small portion of the remainder being unique to smokers with and without lung cancer. Expression distribution patterns showed that most of the genes in smokers with and without lung cancer are expressed at low or moderate levels. We also found that the expression levels of the genes in smokers with lung cancer were lower than in smokers without lung cancer in general. Then we detected 27 differentially expressed genes in smokers with versus without lung cancer, and these differentially expressed genes were foudn to be involved in diverse processes. Our study provided detail expression profiles and expression changes between smokers with and without lung cancer.

Keywords: Genes - lung cancer - smokers - gene expression profiles - RNA-Seq

Asian Pacific J Cancer Prev, **13**, 3605-3609

Introduction

Lung cancer is the most common cause of cancer-related death in both men and women throughout the world. Lung cancer can be broadly classified into two main types based on the morphological characteristics: non-small cell lung cancer and small cell lung cancer (Xiao et al., 2011). There are many causes of cancer include carcinogens (such as those in tobacco smoke), ionizing radiation, viral infection, etc. Cigarette smoke contains over 60 known carcinogens (Hech, 2003) and tobacco smoke is the main contributor to lung cancer (Biesalski et al., 1998) . Therefore, the tissues of those smokers with and without lung cancer provide great resources to study their gene expression changes and find out those lung cancer related genes to help corresponding treatment.

RNA-Seq technologies are now popularly used in diverse transcriptome studies (such as alternative splicing, gene expression, gene fusions etc.) and exhibit many amazing aspects (Mortazavi et al., 2008; Sultan et al., 2008; Maher et al., 2009; Zhao, 2009; Gan et al., 2010; Guttman et al., 2010; Trapnell et al., 2010; F, 2011; Geng Chen, 2011; Pflueger, 2011). Compared with microarrays, RNA-Seq has many advantages. It needs less RNA samples, products lower background noise, could detect new genes and/or transcripts and so on (Marioni et al., 2008; Wang et al., 2009; Marguerat and Bahler, 2010; Nagalakshmi et al., 2010; Beane, 2011). To better understand the gene expression differences between smokers with and without lung cancer, we analyzed two related datasets from short read archive (Beane, 2011).

We quantified the expression of human genes in these two samples of smokers with and without lung cancer. Then, we compared their expressed genes and studied the gene expression distribution patterns in both two samples. To further investigate the gene expression changes between smokers with and without lung cancer, we carried out differential expression analysis and found out a number of differentially expressed genes. The results show some interesting phenomenon of the gene expression profiles between smokers with and without lung cancer, and highlighting that the RNA-Seq technologies are powerful to study the characteristics of human transcriptome.

Materials and Methods

The RNA-Seq datasets of smokers with and without lung cancer were downloaded from short read archive (SRA) with the accession number: SRX060175 (smokers without lung cancer) and SRX060176 (smokers with lung cancer). The human reference genome hg19 was downloaded from UCSC http://genome.ucsc.edu/. We first extracted the human transcript sequences from hg19. Then the RNA-Seq reads of SRX060175 and SRX060176 were mapped onto those transcripts with two mismatches

[1]*Department of Oncology, PLA general hospital of Chengdu Commond, Clinical Medical College of the Third Medical Military University, Chengdu,* [2]*Department of Otolaryngology–Head and Neck Surgery, PLA General Hospital of Nanjing Commond, Clinical Medical College of Nanjing University, Nanjing, China* [&]*Equal contributors *For correspondence: zhangtaossdd@hotmail.com*

allowed using SeqMap (Hui Jiang, 2008). The gene expression levels were calculated using rSeq (Hui Jiang, 2008) and 0.1 RPKM (reads per kilobase of exon model per million mapped reads) was chosen as the threshold. Differential expression analysis between these two samples of smokers with and without lung cancer was carried out using the software of DESeq and chose its model of without any replicates. Adjusted P-value <0.1 were used as the threshold of differentially expressed genes.

To study the gene expression profiles between smokers with and without lung cancer, we downloaded two related RNA-Seq datasets from short read archive (SRA) with the accession number: SRX060175 (smokers without lung cancer) and SRX060176 (smokers with lung cancer) (Beane, 2011). The corresponding samples were from the human large airway epithelial cells of smokers with and without lung cancer. They were sequenced using the Illumina Genome Analyzer IIx platform with standard Illumina mRNA-Seq protocol. The reads are single-end and 36 bp in length. There are total ~26.94 million and ~27.78 million reads for SRX060175 and SRX060176, respectively.

We used the methods of rSeq (Jiang and Wong, 2009) to quantify the human gene expressions. First, we extract the human transcript sequences from the human reference genome hg19. Then we mapped(Hui Jiang 2008; Marguerat and Bahler, 2010) the RNA-Seq reads of SRX060175 and SRX060176 to those transcripts with two mismatches allowed. We computed the gene expression levels using rSeq and chose 0.1 RPKM (reads per kilobase of exon model per million mapped reads) as the threshold. For the sample of smokers without lung cancer (SRX060175), 16,248 genes expressed higher than 0.1 RPKM; and 16,321 genes for the sample of smokers with lung cancer (SRX060176). Between these two samples, 15433 expressed genes are in common, 815 genes only expressed in SRX060175 and 888 genes only expressed in SRX060176. The results show that most of human genes expressed in the samples of smokers with and without lung cancer are the same.

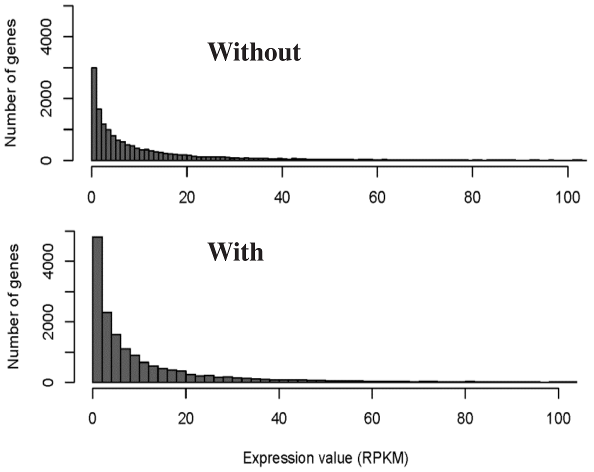

Figure 1. Gene Expression and Distribution of Smokers With and Without Lung Cancer

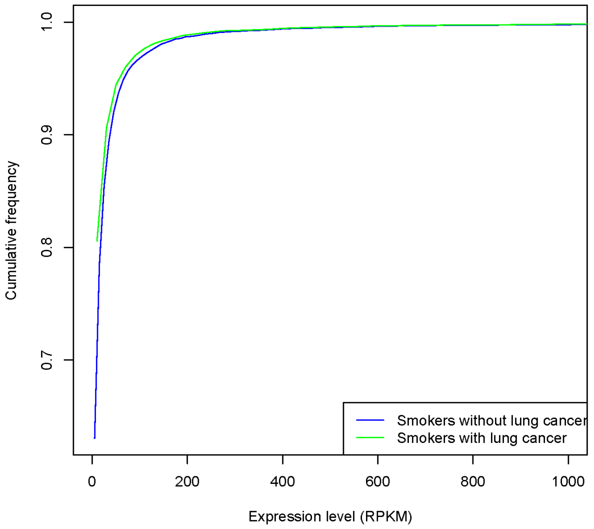

Figure 2. Cumulative Frequency of Gene Expression Levels for Smokers with and without Lung Cancer

Results

Gene expression distribution

We investigated the gene expression level distributions in the two samples of smokers without and with lung cancer. For these two samples of SRX060175 and SRX060176, there are 2,987 and 3,037 genes expressed at the range of 0.1-1 RPKM; 7,255 and 7,675 genes in the range of 1-10 RPKM; 4,706 and 4,445 genes at the range of 10-50 RPKM; 752 and 679 genes at the range of 50-100 RPKM; 548 and 485 genes are equal or greater than 100 RPKM (Table 1). As we can see that the majority of human genes in both smokers with and without lung cancer samples are expressed lower than 50 RPKM (92% for SRX060175 and 92.87% for SRX060176), and remain a small portion of human genes expressed at higher levels (Figure 1).

Table 1. Statistics of Gene Expression Level

Sample	Minimum	Lower quartile	Median	Mean	Upper quartile	Maximum
SRX060175	0.1	1.6	5.75	23.4	16.92	7152.85
SRX060176	0.1	1.56	5.22	22.41	15.38	14326.9

*Those genes with expression level lower than 0.1 RPKM were regarded as unexpressed

We also calculated the minimum, lower quartile, median, mean, upper quartile and maximum gene expression values in these two samples (Table 1). The two highest expression level genes in SRX060175 sample are TPT1 (7152.85 RPKM) and MALAT1 (6681.13 RPKM); and MALAT1 (14326.9 RPKM) and TPT1 (7331.23 RPKM) for sample SRX060176. TPT1 is involved in calcium binding and microtubule stabilization and MALAT1 is associated with metastasis, and positively regulates cell motility via the transcriptional and/or post-transcriptional regulation of motility-related genes. We plotted the cumulative frequency of human gene expression levels in samples of smokers with and without lung cancer (Figure 2). We found that the curve of sample SRX060175 is almost above the one of sample SRX060176, suggesting that the most of human genes in smokers with lung cancer are expressed lower than smokers without lung cancer.

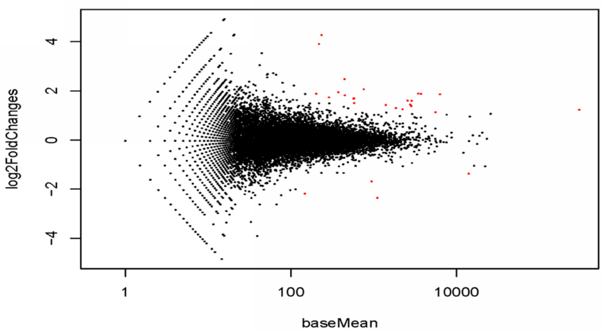

Figure 3. Differential Expression Between Smokers with and Without Lung Cancer. Plot is shown in the log2 fold changes against the base means, the dots that colouring in red are represent those genes that significant (adjusted P-value <0.1) at 10% FDR

Table 2. Differentilly Expressed Genes Between Smokers with and Without Lung Cancer

Gene names	log2FoldChange	P-value	Adjusted P-value
ZNF319	4.266149763	1.57E-15	2.71E-11
IL1B	3.901762687	2.52E-13	2.18E-09
HBB	-2.356397849	7.15E-12	4.12E-08
C1orf161	1.858806437	9.01E-10	2.92E-06
FCGBP	2.472613288	7.64E-10	2.92E-06
MUC5B	1.88916265	1.09E-09	2.92E-06
SCGB3A1	1.878350811	1.18E-09	2.92E-06
XKR9	2.063198021	7.77E-09	1.68E-05
LOC643406	1.600265714	2.80E-07	0.000484672
REXO1L1	1.593950663	2.63E-07	0.000484672
HBA2	-1.681075078	8.58E-07	0.001348971
CCL3	1.944180128	1.83E-06	0.002632677
LMOD3	-1.371161928	3.35E-06	0.00361775
LOC442293	1.442990071	2.98E-06	0.00361775
SPC25	1.814019793	3.20E-06	0.00361775
LPAL2	1.697444741	4.15E-06	0.004226582
CAMK2N1	1.681708887	5.28E-06	0.005073578
SDC4	1.379587271	7.45E-06	0.006780726
HERC2P4	1.426942678	1.07E-05	0.009254681
MALAT1	1.231037835	2.30E-05	0.018983029
TMEM212	1.295619039	4.00E-05	0.030723029
TNIP3	1.50087345	4.09E-05	0.030723029
GPR109B	-2.182816074	5.86E-05	0.039853493
LRRIQ1	1.732550304	5.99E-05	0.039853493
MUC2	1.2498093	5.90E-05	0.039853493
WDR3	1.880201783	9.87E-05	0.063216619
SCGB1A1	1.13039707	0.000150812	0.093163991

Differential expression analysis

To further study the gene expression differences between smokers with and without lung cancers, we carry out differential expression analysis to find out those differentially expressed human genes between these two samples. For calculating the differentially expressed genes, we used the software of DESeq (Anders and Huber, 2010) and chose its model of without any replicates. Using the threshold of adjusted P-value < 0.1, we found that 27 genes differentially expressed in smokers with lung cancer versus smokers without lung cancer, with 4 down-regulated and 23 up-regulated (Figure 3 and Table 2).

About those differentially expressed genes, they have diverse functions and involved in different pathways. Several of those differentially expressed genes have important functions with lung, such as HBB and HBA2 genes are involved in oxygen transport from the lung to the various peripheral tissues (Wajcman et al., 1992; Sanna et al., 1994); MALAT1 (metastasis associated lung adenocarcinoma transcript 1) is a large and infrequently spliced non-coding RNA, it is associated with metastasis and positively regulates cell motility while the transcriptional and/or post-transcriptional regulation of motility-related genes (Huang da, 2009; Tseng, 2009; Guo, 2010). Other differentially expressed genes are related with various functions, IL1B are involved in the inflammatory response, being identified as endogenous pyrogens; SPC25 Acts as a component of the essential kinetochore-associated NDC80 complex, which is required for chromosome segregation and spindle checkpoint activity; SDC4 is cell surface proteoglycan that bears heparan sulfate; MUC2 coats the epithelia of the intestines, airways, and other mucus membrane-containing organs and so on. We also carried out functional annotation clustering using DAVID (Huang da et al., 2009; Huang da, 2009; Guo, 2010; Geng Chen, 2011), but only three genes (MUC2, FCGBP, MUC5B) could clustered together and met the criterion that adjusted P-value <0.1.

Discussion

In this study, we investigated the gene expression differences between smokers with and without lung cancer with two transcriptome sequencing datasets downloaded from short read archive. We first estimated the gene expression levels between these two samples and found that the majority expressed genes of them are the same, indicating that the expression profile differences between smokers with and without lung cancer might be not the unique expressed genes but the subtle expression changes of the genes. Analyzed results also show that most of human genes are expressed at a low or moderate level in both two samples of smokers with and without lung cancers, remain a small portion of human genes expressed at extremely high levels. It suggests that lung cancer disease seems does not disturb the whole trends of gene expression distribution. To know more about the expression divergence between these two samples, we then inferred the differentially expressed genes between the smokers with and without lung cancer. Because there are no replicates of these two samples, we used stringent criteria to call the differential expression. Finally, 27 genes were found differentially expressed between smokers with and without lung cancer. Further analyses suggested that some of those differentially expressed genes have crucial functions in lung tissues but other differentially expressed genes are involved in diverse functional pathways.

Genes coding for the secreted intestinal mucins MUC2 has been mapped to chromosomes 11 (p15). Inactivation of Muc2 causes lung tumor formation with spontaneous progression to invasive carcinoma, and this occurs in the absence of the overt inflammatory response. The reduced representation of goblet cells is characteristic of many aberrant crypt foci (ACF) of both humans and rodents, which are considered early preneoplastic lesions (Velcich et al., 2002). MUC2 gene expression data support the hypothesis that the reduction in these cells and, thus, reduction of the mucus they produce, plays a role in tumor formation.

Tumors with increased expression of mucin genes tended to be associated with post-operative relapse, especially when MUC5B genes were overexpressed ($p = 0.015$). Tumors from smokers tended to have higher MUC5B and MUC5AC gene expression ratios than those of non-smokers (MUC5B: 1.71 vs. 0.76, $p = 0.023$ and MUC5AC: 1.46 vs. 0.81, $p = 0.040$), and were more likely to overexpress much genes (52.9% of tumors from smokers vs. 23.1% of tumors from non-smokers had overexpression of mucin genes $p = 0.039$) (Yu et al., 1996). It is noteworthy that a high percentage of squamous-cell carcinomas also expressed mucin genes and proteins. This finding seems to validate "Yesner's diagram": lung cancers derived from the same pluripotent cells, and squamous-cell carcinoma may preserve their mucin-secretory potential.

IgG Fc binding protein (FcγBP) that binds the Fc portion of IgG molecules has been reported in mucin secreting cells in colon, small intestine, gall bladder, cystic duct, bronchus, sub mandibullar gland, cervix uteri, and in fluids secreted by these cells in human (O'Donovan et al., 2002). The FcγBP gene investigated in the study has potential as a genetic marker in lung cancer. In each of the malignant lung tumors tested the ratio of FcγBP mRNA expression (relative to normal tissue) was less than one, whereas in all the lung tumor and in three out of four of the hyperplastic nodules the ratio of FcγBP expression was greater than one. Measurement of FcγBP mRNA expression in lung tumors and surrounding normal tissues would thus have enabled us to predict the benign or malignant nature of these lung nodules.

Lung cancer is widely affecting the health of human and can lead to cancer-related death, it is vital to study the molecular mechanisms that causing lung cancer. RNA-Seq is now a more flexible and more accurate technology than gene microarrays to investigate the gene expression changes among those healthy and unhealthy lung tissues. It provides us great abilities to study the properties of human gene expressions and generate an unprecedented view of the human transcriptome. Our study shows that the transcriptome sequencing data from smokers with and without lung cancer provide us great opportunities to compare the gene expression profiles between these two samples. Therefore, the RNA-Seq technologies are very powerful to reveal the characteristics of the gene expressions and enable us to study the gene activities more comprehensively. Our results uncover some interesting phenomenon of the gene expression profiles between smokers with and without lung cancer. We believe that more and more intriguing findings will be reported with the progress in sequencing technologies and bioinformatics algorithms. These advances will definitely bring many benefits to the human cancer treatments.

Acknowledgements

Author Contributions: All authors had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

References

Anders S, Huber W (2010). Differential expression analysis for sequence count data. *Genome Biol*, **11**, R106.

Beane J VJ, Schembri F, Anderlind C, et al (2011). Characterizing the impact of smoking and lung cancer on the airway transcriptome using RNA-Seq. *Cancer Prev Res (Phila)*, **4**, 803-17.

Biesalski HK, Bueno de Mesquita B, Chesson A, et al (1998). European consensus statement on lung cancer: risk factors and prevention. lung cancer panel. *CA Cancer J Clin*, **48**, 167-76; discussion 4-6.

Chepelev I, Wei G, Tang Q, Zhao K (2009). Detection of single nucleotide varisations in expressed exons of the human genome using RNA-Seq. *Nucleic Acids Res*, **37**, e106.

F O (2011). Milos P M RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet*, **12**, 87-98.

Gan Q, Chepelev I, Wei G, et al (2010). Dynamic regulation of alternative splicing and chromatin structure in Drosophila gonads revealed by RNA-seq. *Cell Res*, **20**, 763-83.

Geng Chen KY, Shi L, Fang Z, et al (2011). Comparative analysis of human protein-coding and noncoding RNAs between brain and 10 mixed cell lines by RNA-Seq. *PLoS One*, **21**, e28318.

Guo F LY, Liu Y, Wang J, Li Y, Li G (2010). Inhibition of metastasis-associated lung adenocarcinoma transcript 1 in CaSki human cervical cancer cells suppresses cell proliferation and invasion. *Acta Biochim Biophys Sin (Shanghai)*, **42**, 224-9.

Guttman M, Garber M, Levin JZ, et al (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol*, **28**, 503-10.

Hech SS (2003). Tobacco carcinogens, their biomarkers and tobacco-induced cancer. *Nat Rev Cancer*, **3**, 733-44.

Huang da W SBT, Lempicki R A (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, **4**, 44-57.

Huang da W, Sherman BT, Lempicki RA (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*, **37**, 1-13.

Hui Jiang WHW (2008). SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics*, **25**, 2395-6.

Jiang H, Wong WH (2009). Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, **25**, 1026-32.

Maher CA, Kumar-Sinha C, Cao X, et al (2009). Transcriptome sequencing to detect gene fusions in cancer. *Nature*, **458**, 97-101.

Marguerat S, Bahler J (2010). RNA-seq: from technology to biology. *Cell Mol Life Sci*, **67**, 569-79.

Marioni JC, Mason CE, Mane SM, et al (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*, **18**, 1509-17.

Mortazavi A, Williams BA, McCue K, et al (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, **5**, 621-8.

Nagalakshmi U, Waern K, Snyder M (2010). RNA-Seq: a method for comprehensive transcriptome analysis. *Curr Protoc Mol Biol*, *Chapter 4*, Unit 4 11 1-3.

O'Donovan N, Fischer A, Abdo E, et al (2002). Differential expression of IgG Fc binding protein (FcgammaBP) in human normal thyroid tissue, thyroid adenomas and thyroid carcinomas. *J Endocrinol*, **174**, 517-24.

Pflueger D TS, Sboner A, Habegger L, et al (2011). Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. *Genome Res*, **21**, 56-67.

Sanna MT, Giardina B, Scatena R, et al (1994). Functional alterations in adult and fetal hemoglobin Sassari Asp-alpha 126(H9)-->His. The role of alpha 1 alpha 2 contact. *J Biol Chem*, **269**, 18338-42.

Sultan M, Schulz MH, Richard H, et al (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**, 956-60.

Trapnell C, Williams BA, Pertea G, et al (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, **28**, 511-5.

Tseng JJ, HY, Hsu SL, Chou MM (2009). Metastasis associated lung adenocarcinoma transcript 1 is up-regulated in placenta previa increta/percreta and strongly associated with trophoblast-like cell invasion in vitro. *Mol Hum Reprod*, **15**, 725-31.

Velcich A, Yang WC, Heyer J, et al (2002). Colorectal cancer in mice genetically deficient in the mucin Muc2. *Science*, **295**, 1726-9.

Wajcman H, Kister J, Vasseur C, et al (1992). Structure of the EF corner favors deamidation of asparaginyl residues in hemoglobin: the example of Hb La Roche-sur-Yon [beta 81 (EF5) Leu----His]. *Biochim Biophys Acta*, **1138**, 127-32.

Wang Z, Gerstein M, Snyder M (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, **10**, 57-63.

Xiao H, Ding J, Gao S, et al (2011). Never smokers with lung cancer: analysis of genetic variants. *Asian Pac J Cancer Prev*, **12**, 2807-9.

Yu CJ, Yang PC, Shun CT, et al (1996). Overexpression of MUC5 genes is associated with early post-operative metastasis in non-small-cell lung cancer. *Int J Cancer*, **69**, 457-65.