Using Comparative Genomic Hybridization Arrays (aCGH) Techniques to Detect Chronic Obstructive Pulmonary Disease Related Susceptibility Regions

Lin Hua[1]*, Zheng Yang[1], Ping Zhou[1] and Li An[2]*

[1]*Biomedical Engineering Institute, Capital Medical University, Beijing 100069, China*
[2]*Department of Respiratory and Critical Care Medicine, Beijing Chao-Yang Hospital, Capital Medical University, Beijing 100020, China*
yzh31350@163.com, wjzpwyz@163.com

Abstract

In recent years, comparative genomic hybridization arrays (aCGH) techniques have been developed rapidly, and aCGH data analysis can identify chromosomal aberrations that are related to the development of many complex diseases. Currently, Chronic Obstructive Pulmonary Disease (COPD) is predicted to become the third most common cause of death and the fifth most common cause of disability in the world by 2020. Unfortunately, So far the studies to COPD have not been well characterized despite the well-documented role that cigarette smoking plays in the genesis of COPD. Therefore, in this study, we used comparative genomic hybridization arrays (aCGH) techniques to detect COPD related susceptibility regions (potential genomic aberrations) which will provide the support for COPD clinical study. Furthermore, the SW-ARRAY algorithm was used to detect the copy number variable (CNV) regions, and these regions were compared between patients with COPD and patients without COPD. Our results can help understand the disease etiology of COPD.

Keywords: *aCGH; COPD; CNV regions; genomic aberrations*

1. Introduction

Chronic obstructive pulmonary disease (COPD) is an inherently heterogeneous disorder. Within a given individual, there may be varying contributions of emphysema, chronic bronchitis, and long-term smoking. Although smoking is the important environmental risk factor, the existing reports show that only 10% of the chronic heavy smokers develop symptomatic COPD [1-2]. Recently, a series of studies have implicated that COPD represents a complex disease with genetics contributions from multiple genes. It is therefore suggested that there must be some genetic predisposing risk factors contributing to COPD susceptibility.

Recently, it was reported that large, rare deletions within gene regions might be the causal loci for multiple complex phenotypes, and an increasing number of genomic aberrations has been observed in the progression from normal sample to disease sample. A recent study has identified the some chromosomal aberrations in squamous cell carcinoma (SCC) samples by using aCGH data analysis [3]. Previous whole-genome analyses of copy number and gene expression have led to the identification of global cellular processes underlying malignant transformation and progression [4]. Some early aCGH studies on

* To whom correspondence should be addressed. Tel. 010-83911552. E-mail: hualin7750@yahoo.com.cn.
* To whom correspondence should be addressed. Tel. 010-85231893. E-mail: bjzy818@sina.com

breast cancer found that the highly amplified genes were over-expressed and the highly over-expressed genes were amplified [5-7]. Therefore, DNA copy number might influence gene expression across a wide range of DNA copy number alterations. Although few similar studies on COPD were performed, we hypothesize this phenomenon might exist in many complex diseases.

In this paper, to characterize genomic alterations associated with COPD disease, we performed a bioinformatics analysis using aCGH profiles from patients with and without COPD. The most common genomic aberrations in different group were assessed. As a result, we found three common high copy amplifications regions (3q25.2-3q27.1, 5p15.3-p13.1 and 8q24.1-q24.3) and two high copy deletions regions (3p26.3-12.1 and 5q11.1-q35.2) shared by patients with and without COPD. Specially, we found the copy amplification of 2p16.2-p13.22 was only detected for patients with COPD. Similarly, a significantly higher frequency of losses of 8p23 was only detected for patients with COPD. These regions may possibly act as a predictor for a relatively prognosis of COPD patients.

2. Materials and Methods

2.1. Data source

In this study, we used GEO data (GSE12280) to implement our analysis. This gene expression dataset includes 34 patients who presented with centrally located primary squamous cell lung carcinoma (SCC), including 15 patients without lymph node or distant organ metastases within 5 years after surgery; 8 patients with lymph node metastases at the time of surgery, but no distant metastases within 5 years after surgery; 11 patients presenting with distant metastases within 2 years after surgery but without lymph node metastases [3]. Different from previous study, we classified the patients into two groups: patients with COPD (17 patients including 8 no metastases, 4 lymph node metastasis and 5 distant metastasis), and patients without COPD (17 patients including 7 no metastases, 4 lymph node metastasis and 6 distant metastasis). The aim of this classification is to detect the potential chromosomal aberrations difference between patients with and without COPD.

2.2. aCGH analysis according to probes and samples

In this analysis, thresholds for gains and losses were set at log-ratios of 0.3 (gain) and -0.3 (loss), respectively [3]. Thresholds for amplifications were set at log-ratios of 0.8 and thresholds for homozygous deletions were set at -0.8. We analyzed the genomic aberrations according to probes and samples, respectively.

2.3. Detecting CNV regions

In this analysis, we used SW-ARRAY algorithm (Smith–Waterman algorithm adapted for Array CGH) [8] provided by Genovar [9] to detect the copy number variable (CNV) regions. The SW-ARRAY algorithm is a technique originally applied in bioinformatics for the local alignment of DNA and protein sequences, and for the identification of sequence segments with unusual properties. The SW-ARRAY algorithm is described as following:

1) A threshold value t_0 is subtracted from the log ratios, ensuring that the mean of the adjusted scores is negative. The score of a segment of consecutive probes is the sum of the corresponding adjusted log ratios.

2) Highscoring ‘islands’ are identified using the Smith–Waterman algorithm. A locally high-scoring segment or island is defined to be a positive-scoring segment whose score cannot be increased by shrinking or expanding the segment boundaries; let $X(p)$ be the

adjusted score for the *p*th probe ordered along the genome. Let us define the score of the segment from *p* to *q* inclusive as

$$T(p,q)=\sum_{i=p}^{q}X(i) \qquad (1)$$

Define *S(p)* to be the score of the island ending at coordinate *p*, and *B(p)* to be the coordinate of the beginning of the island. Then it can be shown that the following Smith–Waterman recursion will find the islands. Let *S(0)* = 0, and for *p* > 0

$$S(p)=\begin{cases}S(p-1)+X(p), S(p-1)+X(p)>0\\ 0, otherwise\end{cases} \qquad (2)$$

$$B(p)=\begin{cases}B(p-1), S(p)>0\\ p, otherwise\end{cases} \qquad (3)$$

The boundaries $\{B(p_{\max}), p_{\max}\}$ and score $S(p_{\max})$ of the overall maximum-scoring island are output by the algorithm. The segment corresponding to the maximum-scoring island is replaced by a sequence of zeroes and the algorithm repeated until no positive-scoring islands are detected.

3) The statistical significance of an island was estimated by permutation, as the proportion of times that a higher-scoring island was found in 1000 runs in which the adjusted log ratios were permuted between the probes and the highest-scoring island in the shuffled data recorded in each run.

4) The threshold value selection: Values near 1 at any particular position, it means that a copy-number change is indicated. Values near 0 mean that copy-number changes are not indicated. Intermediate values between 0 and 1 mean that the detection of copy-number changes is to some degree sensitive to the choice of threshold value.

3. Results

3.1. Common genomic regions in samples with gain and loss

3.1.1 Frequency according to probes: According to the thresholds for gain and loss defined in the method section, the average gain (%) and loss (%) for patients with COPD (15.81 and 14.13) and patients without COPD (17.87 and 16.25) is very similar (See Figure 1). Green and red colors indicate gain and loss, respectively.

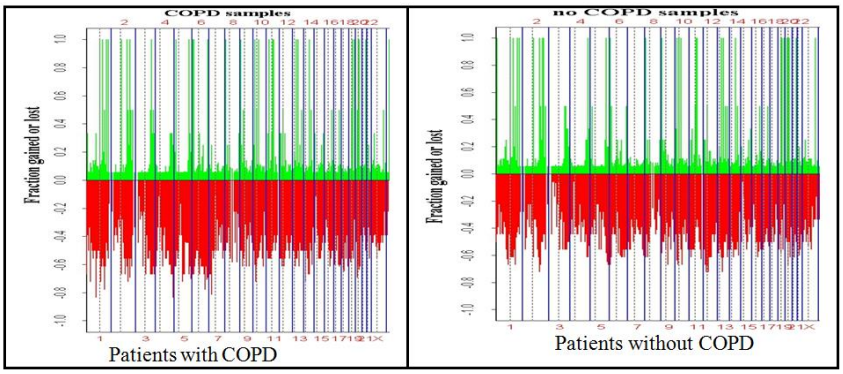


Figure 1. The gain (%) and loss (%) for two groups (patients with COPD and patients without COPD)

In addition, for patients with COPD, we found the highest frequency gain (%) presented in chromosome 14 (90.67) whereas the highest frequency loss (%) presented in chromosome 13 (80.86) (See Figure 2). Green and red colors indicate gain and loss, respectively.

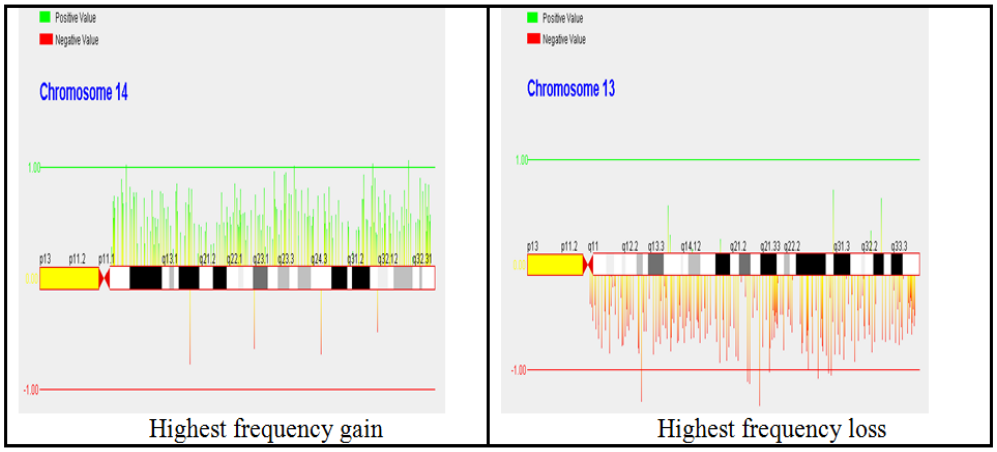


Figure 2. The highest frequency gain (%) and highest frequency loss (%) for patients with COPD

3.1.2. Frequency according to samples: The copy number changes detected in at least 50% of the COPD cases included 7 regions with a gain and 5 regions with a loss (See Table 1). The copy number changes detected in at least 50% of patients without COPD included 4 regions with a gain and 4 regions with a loss (See Table 1). Peak incidences were observed in a smaller sub-region for some of these regions for COPD cases; i.e. gains of 3q26.2-q27.3 (94%), and losses of 3p13-12.1(82%) (See Figure 3). Where, green and red colors indicate gain and loss, respectively.

Indeed, PIK3CA (3q26.32) has been reported previously in SCC squamous cell lung carcinoma sample [10]. For patients without COPD, peak incidences were observed in the same smaller sub-regions; *i.e.*, gains of 3q25.2-3q27.1 (94%), and losses of 3p26.3-12.1 (82%). Three common high copy amplifications (3q25.2-3q27.1, 5p15.3-p13.1 and 8q24.1-q24.3) and two high copy deletions regions (3p26.3-12.1 and 5q11.1-q35.2) were found to be shared by these two groups. Specially, we found the copy amplification of 2p16.2-p13.22 was only detected for COPD cases but not for patients without COPD. In addition, a significantly higher frequency of losses of 8p23 was also detected for COPD cases but not for patients without COPD. These regions may possibly act as a predictor for a relatively prognosis of COPD patients.

Table 1. Gains and losses of two groups according to samples (over 50%)

Sample	Cytoband	event	Frequency (%)	Sample	Cytoband	event	Frequency (%)
COPD	1q21.2-q25.1	gain	52.9-76.5	without COPD	22q13.1-13.3	gain	52.9-70.6
COPD	12p13.3-p11.2	gain	52.9-82.3	without COPD	3q25.2-3q27.1	gain	52.9-94.1
COPD	2p16.2-p13.2	gain	52.9-88.2	without COPD	5p15.3-p13.1	gain	52.9-76.5
COPD	20p13-p11.2	gain	52.9-82.3	without COPD	8q24.1-q24.3	gain	52.9-82.3
COPD	3q25.2-3q27.1	gain	52.9-100.0	without COPD	3p26.3-12.1	loss	52.9-82.3
COPD	5p15.3-p13.1	gain	52.9-64.7	without COPD	4q32.3-34.2	loss	52.9-70.6
COPD	8q24.1-q24.3	gain	52.9-94.1	without COPD	5q11.1-q35.2	loss	52.9-88.2
COPD	13q12.2-q14.1	loss	52.9-76.5	without COPD	9p24.3-p21.1	loss	52.9-76.5
COPD	3p26.3-12.1	loss	52.9-82.3				
COPD	4p16.3-p11	loss	52.9-76.5				
COPD	5q11.1-q35.2	loss	52.9-82.3				
COPD	8p23.3-p12	loss	52.9-70.6				

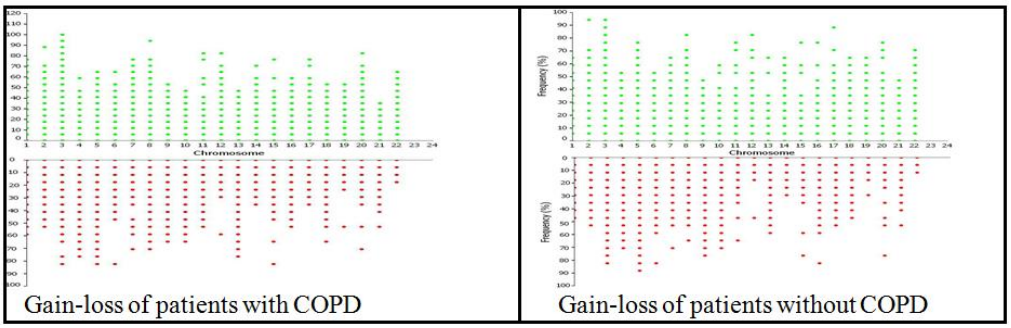


Figure 3. Gains and losses of two groups (patients with COPD and patients without COPD) according to samples

3.2. Detecting CNV regions according to samples

We input parameters include median absolute deviation (MAD) and island block length to start the SW-ARRAY algorithm. Setting higher MAD value and island block length will result in stricter CNV region detection. In order to detect more CNVs, we selected MAD=0.6 and island block length=6. As a result, 439 CNV regions were found. These regions include 292 gain regions and 147 loss regions (See Table 2). From Table 2, we can see that except the gains at 18 chromosome regions (See Figure 4, green and red colors indicate gain and loss, respectively) were restricted to patients without COPD whereas the losses were restricted to COPD cases (Fisher: P=0.024), there were no significant differences in the prevalence of gains and losses between two groups at other chromosome regions.

Table 2. The number of CNV regions for patients with COPD and patients without COPD

Chromosome	Gain (COPD)	Gain (without COPD)	Loss (COPD)	Loss (without COPD)	total	*p*-value
Chr3	3	3	2	0	8	0.464 (Fisher)
Chr4	2	5	5	6	18	0.637 (Fisher)
Chr5	29	28	0	2	59	0.096
Chr6	20	19	11	23	73	0.103
Chr7	3	6	3	1	13	0.266
Chr8	9	9	5	5	28	1.000(Fisher)
Chr9	3	0	3	2	8	0.464 (Fisher)
Chr10	1	1	1	2	5	1.000(Fisher)
Chr11	9	11	3	5	28	1.000(Fisher)
Chr12	22	18	3	2	45	0.831
Chr13	6	5	7	4	22	1.000(Fisher)
Chr14	1	0	3	1	5	1.000(Fisher)
Chr15	0	3	5	2	10	0.167(Fisher)
Chr16	22	16	7	4	49	1.000
Chr17	1	1	1	1	4	1.000
Chr18	6	9	11	2	28	0.024 (Fisher)
Chr19	3	1	3	0	7	0.167(Fisher)
Chr20	9	8	7	5	29	1.000
total	149	143	80	67	439	0.502

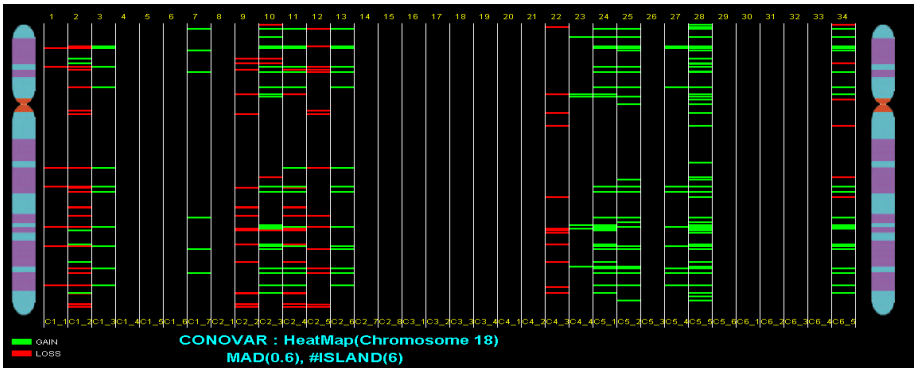


Figure 4. CNV regions (gain and loss) at 18 chromosome

3.3. Detecting SNPs

We used SNPnexus tool [11] which provides a comprehensive set of annotations for genomic variation data by characterizing related functional consequences at different levels of several major annotation systems to detect SNPs within sub-regions with highest frequency gains or losses for COPD samples.

These regions included 8 genes, PEX5L, TNIK, PYDC2, NLGN1, KCNMB3, CGNL1, GABRB2 and KCNK16 (See Table 3). Previous evidences have approved some of these genes are lung disease related. For example, it has been reported the possible target gene KCNMB3 (3q26.32) was significantly targeted in squamous cell carcinoma of the lung [12]. In addition, GABRB2 was also approved specifically to asthma [13].

Table 3. The number of detected SNPs in CNV regions

Probe	gene	Chromosome	Start	End	The number of SNPs	Status	Frequency (%)
RP11-89j17	-	3	175774663	175941716	1736	gain	100.0
RP11-21C11	PEX5L	3	179511186	179653383	1330	gain	100.0
RP11-362K14	TNIK	3	170975588	171101033	1431	gain	100.0
RP11-484I19	-	3	103202967	103378300	1894	gain	100.0
RP11-53d15	PYDC2	3	191158112	191252237	1109	gain	100.0
RP11-44A1	NLGN1	3	173687985	173855782	1655	gain	100.0
RP11-91k9	KCNMB3	3	178971663	179129082	1778	gain	100.0
RP11-196f13	NLGN1	3	173680233	173721234	344	gain	100.0
RP11-163h6	NLGN1	3	173181496	173348022	1732	gain	100.0
RP11-120o11	CGNL1	15	57586567	57757359	2284	loss	82.3
RP11-30G4	-	3	28813262	28966852	1644	loss	82.3
RP11-62g13	GABRB2	5	160654723	160806037	2074	loss	82.3
RP11-133j6	KCNK16	6	39247684	39433103	2356	loss	82.3

4. Discussions

Chronic obstructive pulmonary disease (COPD) has been predicted to become the third most common cause of death and it remains under-recognized and under-diagnosed. In this study, we provided a bioinformatics analysis of the chromosomal regions with copy number changes in COPD cases compared to cases without COPD by using aCGH data. Application of aCGH allows a direct coupling to the copy number changes with the potential target genes. As a result, we found three common high copy amplifications regions (3q25.2-3q27.1, 5p15.3-p13.1 and 8q24.1-q24.3) and two high copy deletions regions (3p26.3-12.1 and 5q11.1-q35.2) shared by patients with COPD and patients without COPD. Specially, the copy amplification of 2p16.2-p13.22 was only detected for COPD cases but not for cases without COPD. These loci can be further explored for their potential use as predictive markers in COPD patients. In addition, candidate genes acquired by detecting SNPs in CNV regions, such as KCNMB3 and GABRB2, may contribute to the pathology of COPD.

However, except the gains at 18 chromosome regions were restricted to cases without COPD whereas losses were restricted to COPD samples (Fisher: P=0.024), there were no significant differences in the prevalence of gains and losses between two groups at other chromosome regions. The most likely explanation for this result is that the samples used in this analysis were all patients who presented with centrally located primary squamous cell lung carcinoma, and COPD did not perform a major role in disease. Therefore, more aCGH

data about COPD case-control samples will be needed to perform further analysis. Furthermore, amplifications and homozygous deletions are relatively small regions, which may be missed by CGH techniques. The latest new technique-laser microdissection [3] applied for the vast majority of cases will get a much higher percentage of cells allowing a more reliable detection of copy number changes.

5. Conclusions

In conclusion, joint analysis of array comparative genomic hybridization (aCGH) copy number data and microarray gene expression data will uncover biological relationships relevant to our understanding of COPD [14-16]. Therefore, our future study is combining large-scale data from a variety of analyses at the SNP, gene and protein levels, which will help direct toward better understanding of COPD pathology.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant Nos. 31100905) and the Science Technology Development Project of Beijing Municipal Commission of Education (SQKM201210025008). This study is also funded by the excellent talent cultivation project of Beijing (2012D005018000002) and the young backbone teacher's cultivation project of Beijing Municipal Commission of Education, and supported by the foundation-clinical cooperation project of capital medical university (11JL30, 11JL33 and 12JL75).

References

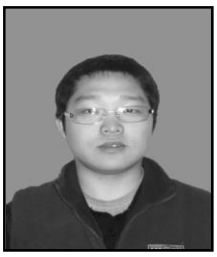
[1] Y. Guo, Y. Gong, C. Pan, Y. Qian, G. Shi, Q. Cheng, Q. Li, L. Ren, Q. Weng, Y. Chen, T. Cheng, L. Fan, Z. Jiang and H. Wan, "BMC Medical Genomics", vol. 5, no. 64, **(2012)**.
[2] S. G. Pillai, D. Ge, G. Zhu, X. Kong, K. V. Shianna, A. C. Need, S. Feng, C. P. Hersh, P. Bakke, A. Gulsvik, A. Ruppert, K. C. L. Carlsen, A. Roses, W. Anderson, I. Investigators, S. I. Rennard, D. A. Lomas, E. K. Silverman and D. B. Goldstein, "PLoS Genetics", vol. 5, no. 3, **(2009)**, pp. e1000421.
[3] M. C. Boelens, K. Kok, P. v. d. Vlies, G. v. d. Vries, H. Sietsma, W. Timens, D. S. Postma, H. J. M. Groen and A. v. d. Berg, "Lung Cancer" vol. 66, no. 3, **(2009)**, pp. 372.
[4] S. F. Chin, A. E. Teschendorff, J. C. Marioni, Y. Wang, N. L. Barbosa-Morais, N. P. Thorne, J. L. Costa, S. E. Pinder, M .A. v. d. Wiel, A. R. Green, I. O. Ellis, P. L. Porter, S. Tavaré, J. D. Brenton, B. Ylstra and C. Caldas, "Genome Biol", vol. 8, no. 10, **(2007)**, pp. R215.
[5] E. Hyman, P. Kauraniemi, S. Hautaniemi, M. Wolf, S. Mousses, E. Rozenblum, M. Ringnér, G. Sauter, O. Monni, A. Elkahloun, O. P. Kallioniemi and A. Kallioniemi, "Cancer Res", vol. 62, no. 21, **(2002)**, pp. 6240.
[6] J. R. Pollack, T. Sørlie, C. M. Perou, C. A. Rees, S. S. Jeffrey, P. E. Lonning, R. Tibshirani, D. Botstein, A. L. Bø rresen-Dale and P. O. Brown, "Proc Natl Acad Sci U S A", vol. 99, no. 20, **(2002)**, pp. 12963.
[7] H. K. Solvang, O. C. Lingjærde, A. Frigessi, A. L. Børresen-Dale and V. N. Kristensen, "BMC Bioinformatics", vol. 12, **(2011)**, pp. 197.
[8] T. S. Price, R. Regan, R. Mott, Å. Hedman, B. Honey, R. J. Daniels, L. Smith, A. Greenfield, A. Tiganescu, V. Buckle, N. Ventress, H. Ayyub, A. Salhan, S. Pedraza-Diaz, J. Broxholme, J. Ragoussis, D. R. Higgs, J. Flint and S. J. L. Knight, "Nucleic Acids Res", vol. 33, no. 11, **(2005)**, pp. 3455.
[9] K. S. Jung, S. Moon, Y. J. Kim, B. J. Kim and K. Park, BMC Bioinformatics. 13 (Suppl 7), S12 **(2011)**
[10] O. Kawano, H. Sasaki, K. Okuda, H. Yukiue, T. Yokoyama, M. Yano and Y. Fujii, "Lung Cancer", vol. 58, no. 1, **(2007)**, pp. 159.
[11] A. Z. D. Ullah, N. R. Lemoine and C. Chelala, "Nucleic Acids Res", 40 (Web Server issue), W65, **(2012)**.
[12] J. U. Kang, S. H. Koo, K. C. Kwon, J. W. Park and J. M. Kim, "BMC Cancer", vol. 9, pp. 237, **(2009)**.
[13] J. A. Hirota, A. Budelsky, D. Smith, B. Lipsky, R. Ellis, Y. Y. Xiang, W. Y. Lu and M. D. Inman, "Clinical & Experimental Allergy", vol. 40, no. 5, **(2010)**, pp. 820.
[14] F. Amir, and J. Shahram, IJAST, vol. 34, **(2011)**, pp. 65.

[15] T. Raja and S.Yahya, IJAST, vol. 39, **(2012)**, pp. 29.
[16] S. Y. Hadi and S. M. Nima, IJAST, vol. 45, **(2012)**, pp. 91.

Authors

Lin Hua*

Biomedical Engineering Institute,
Capital Medical University, Beijing 100069, China
E-mail:hualin7750@yahoo.com.cn



Zheng Yang

Biomedical Engineering Institute,
Capital Medical University, Beijing 100069, China
E-mail: yzh31350@163.com

Ping Zhou

Biomedical Engineering Institute,
Capital Medical University, Beijing 100069, China
E-mail:wjzpwyz@163.com

Li An*

Department of Respiratory and Critical Care Medicine,
Beijing Chao-Yang Hospital, Capital Medical University, Beijing 100020, China
E-mail:bjzy818@sina.com