A Bioinformatics Analysis of aCGH data for Chronic Obstructive Pulmonary Disease

Lin Hua1,*, Zheng Yang1, Ping Zhou1, Li An2,*

1Biomedical Engineering Institute, Capital Medical University, Beijing 100069, China
2Department of Respiratory and Critical Care Medicine, Beijing Chao-Yang Hospital, Capital Medical University, Beijing 100020, China

Abstract. Chronic Obstructive Pulmonary Disease (COPD) is a complex human disease which has higher prevalence and higher mortality. 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. 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. In this study, we used a bioinformatics analysis to identify the potential genomic aberrations related to COPD disease using aCGH data provided by a previous study. In addition, we used the SW-ARRAY algorithm to detect the copy number variable (CNV) regions, and compared these regions between patients with COPD and patients without COPD. Our results can help understand the disease etiology of COPD.

Keywords: aCGH; COPD; gain; loss

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. Recently, a series of studies have implicated that COPD represents a complex disease with genetics contributions from multiple genes. It is therefore suggesting that there must be some genetic predisposing risk factors contributing to COPD susceptibility. Currently, 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. Previous whole-genome analyses of

* To whom correspondence should be addressed. E-mail:hualin7750@yahoo.com and bjzy818@sina.com

copy number and gene expression have led to the identification of global cellular processes underlying malignant transformation and progression. Some early aCGH studies on breast cancer found that the highly amplified genes were over-expressed and the highly over-expressed genes were amplified. 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. Therefore, in this paper, to characterize genomic alterations associated with COPD disease, we performed a bioinformatic 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 and two high copy deletions regions 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

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) [1]. Different from previous study, we classified the patients into two groups: patients with COPD (17 patients), and patients without COPD (17 patients). The aim of this classification is to detect the potential chromosomal aberrations difference between patients with and without COPD. Thresholds for gains and losses were set at log-ratios of 0.3 (gain) and -0.3 (loss), respectively. 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. In addition, we used SW-ARRAY algorithm (Smith–Waterman algorithm adapted for Array CGH) [2] provided by Genovar [3] to detect the copy number variable (CNV) regions.

3 Results

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. 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). 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. 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. 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%). Indeed,
PIK3CA (3q26.32) has been reported previously in SCC squamous cell lung carcinoma sample. For patients without COPD, peak incidences were observed in the same smaller sub-region; 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. Similarly, 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.

Furthermore, we input parameters include median absolute deviation (MAD) and island block length to start the SW-ARRAY algorithm to detect CNV regions. 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. We found that except the gains at 18 chromosome regions 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. Also, we used SNPnexus tool [4] 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 with subregions with highest frequency gains or losses for COPD samples. These regions included 8 genes, PEX5L, TNIK, PYDC2, NLGN1, KCNMB3, CGNL1, GABRB2 and KCNK16. 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. In addition, GABRB2 was also approved specifically to asthma.

4 Discussions

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 and two high copy deletions regions 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 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 summary, joint data 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 [5-7], which will help 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 (11J133).

Main References