CITY UNIVERSITY OF HONG KONG
香港城市大學

# Gene Expression Profiling of Non-small Cell Lung Cancer using cDNA Microarrays
# 利用 cDNA 微陣列晶片技術研究非小細胞肺癌的基因表達譜

Submitted to
Department of Biology and Chemistry
生物及化學系
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
哲學博士學位

by

Au Siu Kie
區兆基

September 2009
二零零九年九月

# Abstract

In the Chinese communities, there is a high incidence of non-smoking-related lung cancer, almost invariably of adenocarcinoma (ADC) histology. Recent evidence suggests that it is a distinct biological entity different from lung cancer of other aetiologies. On the other hand, squamous cell carcinoma (SCC) of lung is almost always associated with smoking.

In the current project, the mRNA expression profiles of ADC and SCC samples taken from 84 patients undergoing surgery were studied in relation to the histology, demographics, stage (lymph node metastases) and EGFR mutation status. The 10K human clone set from Incyte Genomics, Inc. (Palo Alto, California, USA) was used as targets in the spotted cDNA microarray.

In phase I of the study, RNA were pooled from patients stratified according to (1) smoking history (ever-smoker versus never-smoker) (2) lymph node metastases (present or absent) (3) histology (ADC versus SCC). Transcript profiling was done by cDNA microarray using the full set of 10K targets. There were only 6 pools of RNA because SCC occurred in smokers exclusively. The experiments were repeated in triplicate for each pool of RNA. In Phase II of the study, based on the intensities of gene expression and the differential expression between the strata, a subset of 1385 genes were selected for the fabrication of a lower density cDNA microarray for profiling of all 84 lung cancer individual samples using normal lung parenchymal tissue from the same individual as control. The data were filtered by pre-set quality criteria and normalized by localized weighted regression (LOWESS).

There were 40 ADC (all never-smokers) and 44 SCC (22 current and 22 ex-smokers) cases. EGFR mutations occurred in 24 ADC (60%)

but none in SCC. The gene expression profiles were significantly different with respect to parameters including histology, EGFR mutation, gender, lymph node metastases and age. According to pre-set criteria, 53 genes and 17 genes were found to be differentially expressed between tumour and normal lung parenchyma in SCC and ADC respectively. Sixteen genes were differentially expressed between SCC and ADC. Unsupervised clustering clearly segregated ADC from SCC and mutated *EGFR* from wild-type *EGFR* tumours. By support vector machine (SVM) or K-nearest neighbor (KNN) algorithms based on the 16 top significant genes with respect to each corresponding parameter, the accuracies of correct class prediction of histology, EGFR mutation status and gender were 82 - 87%, 54.5 - 70.1% and 67.8 - 75% respectively. The KEGG "cell communication" pathway was the most significant pathway overlapping with those differentially expressed genes in SCC. No significantly overlapping pathway was identified for ADC, probably because of the smaller number of genes with differential expression identified.

For validation of the microarray results, 11 genes (*A2M*, *ADH1B*, *CAV1, CCT5,* CD74, *CEACAM6*, *HIST1H2AE*, *HIST1H2BD*, *SFTPC*, *SPTBN1 and TACSTD1*) were selected (based on literature review of their biological relevance to the pathogenesis of malignancies) for real-time reverse-transcription PCR (qRT-PCR). The results between microarray and qRT-PCR were consistent although statistical significance of correlation was reached only for 6 out of the 11 genes.

DNA sequence analysis of exons 18-21 of the *EGFR* gene showed that the most common mutations were deletion in exon 19 (16.7%) and substitution L858R in exon 21 (66.7%). Most of the cases had a single mutation (91.7%) and the incidence of double mutations was 8.3%. According to pre-set criteria, 7 genes were found to be differentially expressed between mutated and wild-type *EGFR* ADC cases.

Unsupervised clustering segregated distinct transcript expression profiles between the 2 groups.

Because of its great potential of being a therapeutic target, *CD74* and its ligand *MIF* were further studied by qRT-PCR and immunohistochemisty (IHC) in 41 primary ADC, 46 primary SCC and 11 metastatic CA of the lung patients. By qRT-PCR, the transcript expression level of *CD74* was near normal in primary ADC but markedly suppressed in both primary SCC and metastatic CA. On the other hand, mRNA expression level of *MIF* in all 3 groups were markedly elevated. By IHC, CD74 protein expression was strong in primary ADC but absent in the other 2 groups. *MIF* protein expression was strong for all 3 groups. Studies of blockade by siRNA in cell lines are in progress to elucidate the functions of these two genes on tumour cell survival.

We have confirmed that there are distinct mRNA expression profiles between ADC and SCC of lung, and between those ADC of wild-type or mutated *EGFR* gene. *CD74* and *MIF* are potentially useful biomarkers for ADC of lung at both transcript and protein levels. The differential expression of biomarkers in different subgroups of lung cancer in cDNA microarray has been confirmed by the qRT-PCR results. Further functional or larger clinico-pathological studies on the biomarkers discovered are warranted.

# Table of Content