OPEN ACCESS Freely available online

PLOS ONE

Classification of Non-Small Cell Lung Cancer Based on Copy Number Alterations

Bi-Qing Li[2☯], Jin You[3☯], Tao Huang[4]*, Yu-Dong Cai[1]*

1 Institute of Systems Biology, Shanghai University, Shanghai, P.R. China, **2** Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, P. R. China, **3** The Key Laboratory of Stem Cell Biology, Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, P. R. China, **4** Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York City, New York, United States of America

Abstract

Lung cancer is one of the leading causes of cancer mortality worldwide and non–small cell lung cancer (NSCLC) accounts for the most part. NSCLC can be further divided into adenocarcinoma (ACA) and squamous cell carcinoma (SCC). It is of great value to distinguish these two subgroups clinically. In this study, we compared the genome-wide copy number alterations (CNAs) patterns of 208 early stage ACA and 93 early stage SCC tumor samples. As a result, 266 CNA probes stood out for better discrimination of ACA and SCC. It was revealed that the genes corresponding to these 266 probes were enriched in lung cancer related pathways and enriched in the chromosome regions where CNA usually occur in lung cancer. This study sheds lights on the CNA study of NSCLC and provides some insights on the epigenetic of NSCLC.

Citation: Li B-Q, You J, Huang T, Cai Y-D (2014) Classification of Non-Small Cell Lung Cancer Based on Copy Number Alterations. PLoS ONE 9(2): e88300. doi:10.1371/journal.pone.0088300

Editor: Giuseppe Viglietto, UNIVERSITY MAGNA GRAECIA, Italy

Received August 12, 2013; **Accepted** January 6, 2014; **Published** February 5, 2014

Funding: This work was supported by grants from National Basic Research Program of China (2011CB510101, 2011CB510102), Innovation Program of Shanghai Municipal Education Commission (12ZZ087), and the grant of "The First-class Discipline of Universities in Shanghai". The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: tohuangtao@126.com (TH); cai_yud@yahoo.com.cn (Y-DC)

☯ These authors contributed equally to this work.

Introduction

Lung cancer is one of the leading cause of cancer mortality worldwide [1]. Basing on the 2011 International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society (IASLC/ATS/ERS) lung adenocarcinoma classification, it is now classified into 5 different subtypes: Atypical adenomatous hyperplasia (AAH), Adenocarcinoma in situ (AIS) (nonmucinous, mucinous, or mixed nonmucinous/mucinous), Minimally invasive adenocarcinoma (MIA) ($\leq$3 cm lepidic predominant tumor with $\leq$5 mm invasion), Invasive adenocarcinoma, and variants of invasive adenocarcinoma, and each of them has its own histological feature [2]. Non–small cell lung cancer (NSCLC) accounts for 85% of all lung cancers. The most frequent histologic subtypes of NSCC is adenocarcinoma (ACA) and squamous cell carcinoma (SCC), accounting for 50% and 30% of NSCLC cases, respectively [3]. ACA is the most common histologic subtype reported with lung cancer in the never smokers (LCINS) [4], which is a cancer of an epithelium which originates in glandular tissue. SCC is a cancer of squamous epithelial cell, which arises most often in segmental bronchi and related to lobar and main stem bronchus occurs by its extension [5], and its incidence is correlated with smoking period [6] compared with ACA. Historically, well differentiated SCC cells include the morphologic features such as intercellular bridging, squamous pearl formation and individual cell keratinization [5]. Nowadays, medicine development in NSCLC has introduced histologic subtyping, the differentiation of ACA from SCC in biopsy specimens, as an important factor for effective treatment choice and molecular therapy target. For example pemetrexed, antifolate agent, is effective in the treatment of patients with non-squamous NSCLC but should not be recommended for the treatment of squamous cell carcinoma [7]. Bevacizumab, combined with paclitaxel/carboplatin, has excessive toxic effects in squamous-cell carcinoma [8], while it could significantly increase overall survival rate of patients with cancers of non-squamous histology [9,10]. Traditional diagnosis method to distinguish adenocarcinoma from squamous cell carcinoma, is based on the histologic section and patients' smoking habit. However, because of the individual heterogeneity of lung cancer, this method cannot correctly distinguish ACA and SCC in some cases efficiently. Recently, immunohistochemistry is being used in biopsy and cytology material [11] as a complement, and several genes have been discovered as the immunohistochemical marker. Kargi et al. found thyroid transcription factor-1 (TTF-1) is a marker in immunostaining for ACA, while p63 and cytokeratins (CK) 5/6 are marks for SCC [12]. Moreover, molecular targeted therapy has been more and more used in NSCLC as the promising treatment strategy in recent years. It is demonstrated that superior efficacy of tyrosine kinase inhibitors (TKIs) as compared to standard chemotherapy for patients with EGFR-mutant tumors [13]. Kwak et al. also explored the small-molecule inhibitor of the ALK tyrosine kinase could be used as the efficacious therapy in advanced ALK-positive tumors in an early-phase clinical trial [14]. Therefore, it is meaningful to identifying genes which have distinct genetics features in ACA and SCC that could be used as prognostic factor or potential target for medical therapy.

Previous analysis has showed CNAs are common in almost all human cancers [15,16]. In NSCLC, CNAs increase with disease progression and CNAs are both positionally and functionally clustered [17]. Furthermore, Giovanni Tonon el at. found despite their distinct histopathological phenotypes, ACA and SCC genomic profiles showed a nearly complete overlap, with only one clear SCC-specific amplicon on 3q26–29 [18].

In this study, to figure out the key genes distinguishing ACA and SCC from each other, we compare the genome-wide copy number alterations (CNAs) patterns of 208 early stage ACA and 93 early stage SCC tumor samples. By means of the feature selection and analysis methods, including the Maximum Relevance Minimum Redundancy method (mRMR) and the Incremental Feature Selection (IFS) method, 266 optimal CNA probes were selected for the discrimination of ACA and SCC. The classification model was built with Nearest Neighbor Algorithm (NNA). As a result, the classifier achieved a overall MCC of 0.6616. Further analysis on the 266 CNA related genes showed that they were closely associated with lung cancer.

Materials and Methods

Dataset

We used the copy number alterations data from the non-small cell lung cancer study of Huang et al. [19]. In their study, a series of 301 snap-frozen tumor samples from NSCLC patients was collected during surgery or biopsy from the Massachusetts General Hospital (MGH), Boston, MA and the National Institute of Occupational Health, Oslo, Norway. The clinical information of these 301 samples was given in File S1. The copy number profiling of 208 early stage adenocarcinoma tumors (ACA) samples and 93 early stage squamous cell carcinoma tumors (SCC) were retrieved from NCBI Gene Expression Omnibus (GEO) with the accession number of GSE34140. The copy number profile was obtained using the using Affymetrix 250 K Nsp GeneChip. Only 256,554 probes on somatic chromosomes were analyzed. The SNP probes were mapped to the RefSeq genes with 2 kb extension both upstream and downstream using the UCSC Genome Browser. Among the 256,554 probes on somatic chromosomes, 104,256 probes were mapped to 11,700 genes [19].

mRMR method

We used Maximum Relevance Minimum Redundancy (mRMR) method to rank the importance of the probes [20]. mRMR method could rank probes based on both their relevance to the class of samples and the redundancy among probes. A smaller index of a probe denotes that it has a better trade-off between maximum relevance to class of samples and minimum redundancy.

Both relevance and redundancy were quantified by mutual information (MI), which estimates how much one vector is related to another. The MI equation was defined as below:

$$I(x,y)=\iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}\,dxdy \tag{1}$$

In equation (1), x, y are vectors, $p(x,y)$ is their joint probabilistic density, and $p(x)$ and $p(y)$ are the marginal probabilistic densities.

Let Ω denote the whole probe set, Ω_s denote the already-selected probe set containing m probes and Ω_t denote the to-be-selected probe set containing n probes. The relevance D between a probe f in Ω_t and the class of sample c can be calculated by:

$$D=I(f,c) \tag{2}$$

The redundancy R between a probe f in Ω_t and all the probes in Ω_s can be calculated by:

$$R=\frac{1}{m}\sum_{f_i\in\Omega_s} I(f,f_i) \tag{3}$$

To get the probe f_j in Ω_t with maximum relevance and minimum redundancy, the mRMR function combines equation (2) and equation (3) and is defined as below:

$$\max_{f_j\in\Omega_t}\left[I(f_j,c)-\frac{1}{m}\sum_{f_i\in\Omega_s} I(f_j,f_i)\right](j=1,2,...,n) \tag{4}$$

The mRMR probe rating would be executed N rounds when given a probe set with N (N = m+n) probes. After N rounds of execution, a probe set S is produced:

$$S=\left\{f_1',f_2',...,f_h',...,f_N'\right\} \tag{5}$$

In S, index h indicates at which round that the probe is selected. The smaller the index h is, the earlier the probe satisfies equation (4) and the better the probe is.

Nearest neighbor algorithm (NNA)

Nearest Neighbor Algorithm (NNA) [21,22], which has been widely used in bioinformatics and computational biology [23,24,25,26,27], was adopted to predict the class of samples. The "nearness" was calculated according to the following equation

$$D(\mathbb{P}_1,\mathbb{P}_2)=1-\frac{\mathbb{P}_1,\mathbb{P}_2}{\|\mathbb{P}_1\|\cdot\|\mathbb{P}_2\|} \tag{6}$$

where $\mathbb{P}_1$ and $\mathbb{P}_2$are two vectors representing two samples, $\mathbb{P}_1\cdot\mathbb{P}_2$ is their dot product, $\|\mathbb{P}_1\|$ and $\|\mathbb{P}_2\|$ are their moduluses. The smaller the $D(\mathbb{P}_1,\mathbb{P}_2)$, the more similar the two samples are.

For an intuitive illustration of how NNA works, see Fig.5 of [28].

Jackknife Cross-Validation Method

Jackknife Cross-Validation Method [23,24,29,30] (also called the Leave-one-out cross-validation, LOOCV) was used to evaluate the performance of a classifier. In Jackknife Cross-Validation Method, every sample is tested by the predictor that is trained with all the other samples. Let TP denotes true positive. TN denotes true negative. FP denotes false positive and FN denotes false negative. To evaluate the performance of our predictor, the prediction accuracy, specificity, sensitivity and MCC (Matthews's correlation coefficient) were calculated as below:

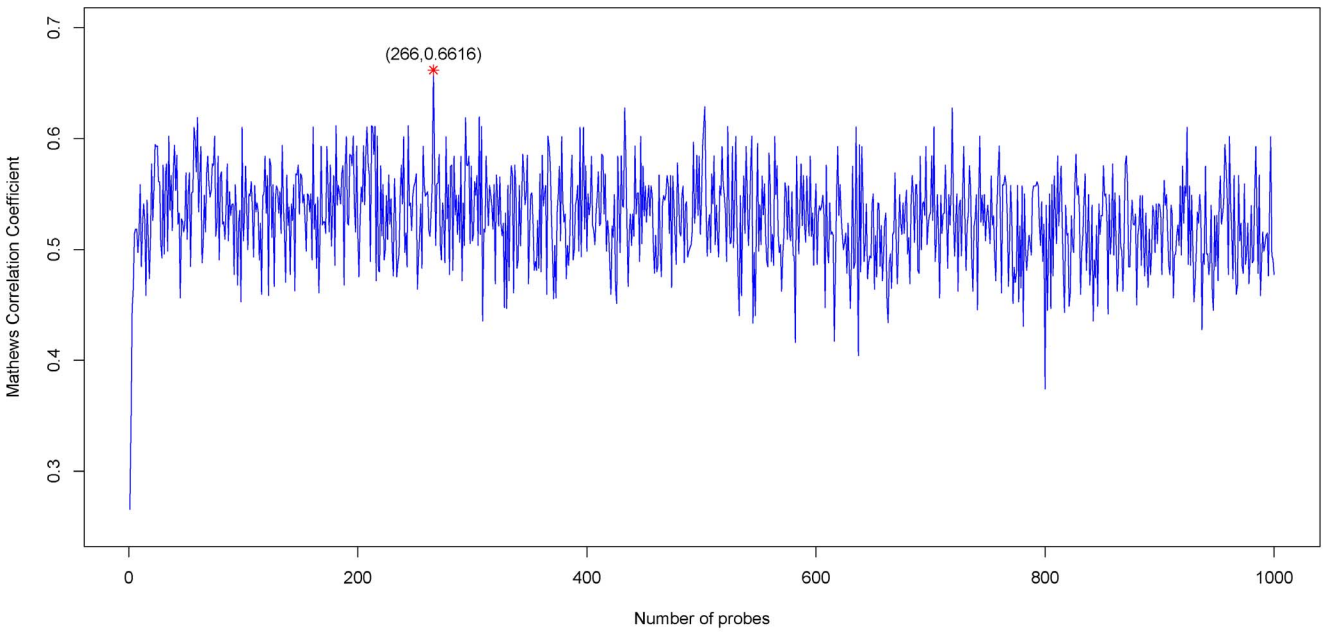

Figure 1. IFS curve for the adenocarcinoma (ACA) and squamous cell carcinoma (SCC) samples classification. The IFS curves were drawn based on the data in File S3. The MCC reached the peak when the number of probes was 266. The 266 probes thus obtained were used to compose the optimal probe set for discrimination of adenocarcinoma (ACA) and squamous cell carcinoma (SCC).
doi:10.1371/journal.pone.0088300.g001

$$\begin{cases} accuracy = \dfrac{TP+TN}{TP+TN+FP+FN} \\ sensitivity = \dfrac{TP}{TP+FN} \\ specificity = \dfrac{TN}{TN+FP} \\ MCC = \dfrac{TP\times TN - FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \end{cases} \quad (7)$$

Incremental Feature Selection (IFS)

Based on the ranked probes rated by mRMR evaluation, we used Incremental Feature Selection (IFS) [31,32,33] to determine the optimal number of probes. During IFS procedure, probes in the ranked probe set are added one by one from higher to lower rank. A new probe set is composed when one probe is added. Thus N probe sets would be composed given N ranked probes. The i-th probe set is:

$$S_i = \{f_1, f_2, ..., f_i\} (1 \le i \le N) \quad (8)$$

For each of the N probe sets, an NNA predictor was constructed and tested using LOOCV. With N prediction accuracies, sensitivities, specificities and MCCs calculated, we obtain an IFS table with one column being the index i and the other columns to be the prediction accuracy, sensitivity, specificity and MCC. The optimal probe set (S_{optimal}) is the one, using which the predictor achieves the best prediction performance.

Functional enrichment analysis of CNAs genes

Functional annotation tool of GATHER [34] was used for KEGG pathway, GO and chromosome region enrichment analysis. All the genes in the human genome were selected as background during the enrichment analysis.

Results and Discussion

The mRMR Result

Listed in the File S2 are two kinds of outcomes obtained by running the mRMR software: one is called the "MaxRel feature list" that ranked all the probes according to their relevance to the class of samples; the other one is the "mRMR feature list" that ranked the probes according to the criteria of maximum relevance and minimum redundancy. In the mRMR probe list, the smaller the index of a probe was, the more important the probe would be for the discrimination of two kinds of NSCLC. Accordingly, the mRMR feature list could be used to establish the optimal feature set in the IFS procedure.

IFS and Final Optimal Feature Set

Based on these two tables, 1000 feature subsets were constructed according to Eq.8. An NNA predictor was modeled for each subset and was evaluated by LOOCV. Shown in Fig.1 is the IFS curve plotted based on the data in File S3. The x-axis is the number of probes used for the classification, and the y-axis is the MCC values of classifiers evaluated by LOOCV. The maximum MCC was 0.6616 when 266 probes were utilized. With such a classifier, the prediction sensitivity, specificity and accuracy were 0.9567, 0.6452 and 0.8605, respectively. These 266 probes were regarded as the optimal biomarkers for the discrimination of two kinds of NSCLC. The information of these 266 probes were given in File S4. Shown in Fig.2 is the heatmap based on these 266 probes. It can be seen that most of the 208 ACA samples and 93 SCC samples can be distinguished.

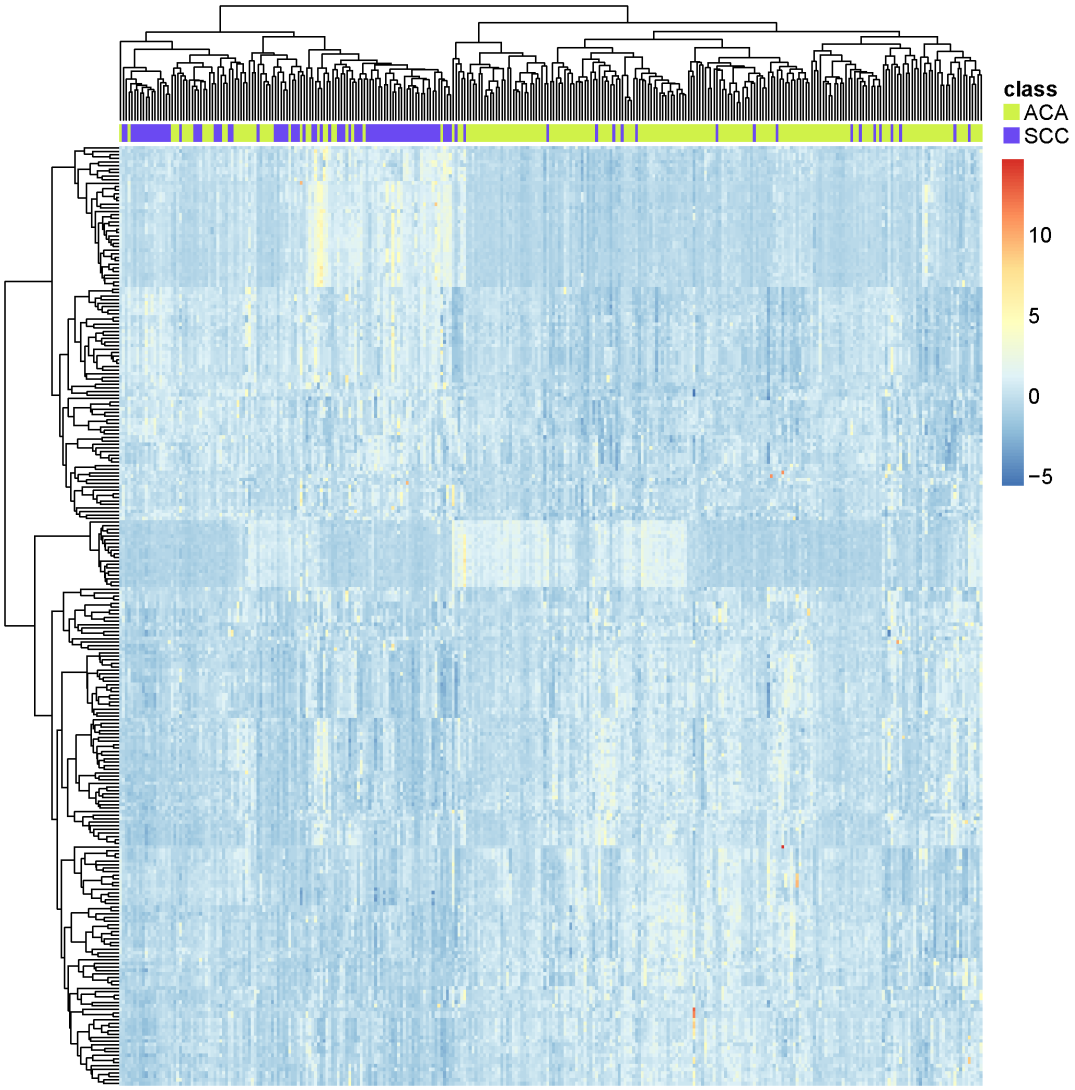

Figure 2. Heatmapof 208 adenocarcinoma (ACA) samples and 93 squamous cell carcinoma (SCC) samples with 266 selected probes. Samples are arranged along the X axis and probes along the Y axis. Each square represents the copy number of a given probe in an individual sample. Red is increased copy number and blue is decreased copy number relative to the mean- and sample-centered scaled copy number across the samples. Adenocarcinoma (ACA) and squamous cell carcinoma (SCC) samples were presented with green and blue, respectively.
doi:10.1371/journal.pone.0088300.g002

KEGG and GO enrichment results of CNAs genes

The KEGG pathway enrichment analysis of CNAs genes indicated that they were enriched in Wnt signaling pathway, Focal adhesion, ECM-receptor interaction and so on (Table 1). It is reported Wnt signaling pathway is activated during the carcinogenesis of NSCLC [35], and inhibition of Wnt-2-mediated signaling could induce non-small-cell lung cancer cells apoptosis [36]. Focal adhesion and ECM-receptor interaction are pathways in the biological processes interactions of cells with extracellular matrix (ECM), which play crucial roles in cell motility, cell proliferation, cell differentiation, regulation of gene expression and cell survival [37,38]. The proteins of these pathways are up-regulated in NSCLCs [39], and take part in the activation of local invasion and distant metastasis of cancer cells [40]. As the KEGG pathway enrichment result, the GO enrichment result of these CNAs genes also shows enrichment in the terms of cell adhesion and intracellular signaling cascade. The GO enrichment result of these CNAs genes were listed in File S5.

Chromosome region enrichment result of CNAs genes

It is reported copy number gain in region 3q26 [18,41] and in region 8p12 [42] seem to be more common in squamous histology compared with adenocarcinoma. The analysis of our result shows that including these two regions, copy number alterations of 2q34, 10p15, 18q11, 8p23, 3p21, 3q27, 22q12, Xq13, 2q36, 10p11, 10p12 also have the significance in discrimination between SCC and ACA, and deserved further researches on them (Table 2).

CNAs genes identified in this study

In this study, we identified several candidate genes corresponding to 266 CNAs probes that can be used to distinguish two kinds of NSCLC. 50 of them also has a significant correlation to the Smoking Pack-year including TP63, SOX2 and PPP2R2B (see File S4). With literature retrieval of gene function and significance comparison by p-value, we focused on 8 genes which are most probably related to distinguish ACA and SCC from each other. Among them, TP63 has been reported as a biomarker to

Table 1. KEGG enrichment result of CNAs genes.

Pathway	KEGG ID	Your Genes (With Ann)	Your Genes (No Ann)	Genome (With Ann)	Genome (No Ann)	P-value
Wnt signaling pathway	hsa04310	6	32	141	2951	0.0077
Focal adhesion	hsa04510	7	31	227	2865	0.0204
ECM-receptor interaction	hsa04512	4	34	82	3010	0.0193

Your Genes (With Ann): The number of genes from your list with the annotation.
Your Genes (No Ann): The number of genes from your list without the annotation.
Genome (With Ann): The number of genes in the genome (excluding those in your list) with the annotation.
Genome (No Ann): The number of genes in the genome (excluding those in your list) without the annotation.
P-value: The negative logarithm of the *p* value calculated using a Fisher's exact test.
doi:10.1371/journal.pone.0088300.t001

discriminate between SCC and ACA, and it is listed top in our result. Some of other genes are reported to have different gene expression level in ACA and SCC or in patients with distinct smoking habits. In accord with the KEGG and GO enrichment result, PPP2R2B is a gene in wnt signaling pathway, while ITGA9 takes a part in focal adhesion and ECM-receptor interaction. All above illustrates that our result is biologically significant and the 8 genes may be candidate biomarkers for distinguishing ACA and SCC from each other and deserved further studies on them. Below, we will briefly discuss their relationships with NSCLC.

TP63 (Tumor protein 63) is listed top one in the optimal probe set with a CNA fold change of 0.7827 comparing ACA with SCC. It is a tumor suppressor p53 homologue and essential for p53 dependent apoptosis in response to DNA damage [43]. Mi Jin Kim et al. found P63 is a useful immunohistochemical panel in differentiating ACA from SCC of the lung with the positive rate 91% of SCC and 9% of ACA in their studies [44]. The chromosome location of TP63 is 3q27–29. Therefore, our result is coincide with former researches and TP63 may play a key role in distinguish ACA and SCC from each other.

EPHA4 (Ephrin type-A receptor 4) is related to the fourth probe in our optimal probe set with a CNA fold change of 1.0846 comparing ACA with SCC, and is a member of the Eph receptor family, the largest receptor tyrosine kinase family of transmembrane proteins with their ligands, the ephrins, affecting the growth, migration and invasion of cancer cells in culture as well as tumor growth, invasiveness, angiogenesis and metastasis in vivo [45]. Junya Fukai et al. found EphA4 promotes cell proliferation and migration through a novel EphA4-FGFR1 signaling pathway in the human glioma U251 cell line [46]. One of the Eph receptors EphA2 is reported over expression in smokers and predicts poor survival in non-small cell lung cancer [47]. A mutation in EphA2 (G391R) was identified in two of 28 squamous cell lung cancers (7%), but not in any adenocarcinomas or large-cell lung carcinomas [48]. These all indicate that EphA4 may be a candidate biomarker for distinguishing ACA and SCC from each other and deserved further studies on it.

PPP2R2B (Serine/threonine-protein phosphatase 2A 55 kDa regulatory subunit B beta isoform) is related to the fifth probe in our optimal probe set with a CNA fold change of 1.0781 comparing ACA with SCC. It is the regulatory subunit B beta isoform of PP2A, and is implicated in the negative control of cell growth and division [49]. Recently genome-wide association study (GWAS) of lung cancer in the Chinese population revealed that

Table 2. Chromosome region enrichment result of CNAs genes.

Chromosome region	Your Genes (With Ann)	Your Genes (No Ann)	Genome (With Ann)	Genome (No Ann)	P-value
2q34	5	162	24	30139	5.09E-07
10p15	5	162	55	30108	2.04E-05
18q11	4	163	46	30117	0.0002
3q26	5	162	105	30058	0.0004
8p23	6	161	174	29989	0.0005
3p21	7	160	251	29912	0.0006
3q27	4	163	72	30091	0.0008
22q12	5	162	142	30021	0.0014
Xq13	4	163	100	30063	0.0027
2q36	3	164	51	30112	0.0033
10p11	3	164	62	30101	0.0056
10p12	3	164	63	30100	0.0058

Your Genes (With Ann): The number of genes from your list with the annotation.
Your Genes (No Ann): The number of genes from your list without the annotation.
Genome (With Ann): The number of genes in the genome (excluding those in your list) with the annotation.
Genome (No Ann): The number of genes in the genome (excluding those in your list) without the annotation.
P-value: The negative logarithm of the *p* value calculated using a Fisher's exact test.
doi:10.1371/journal.pone.0088300.t002

chromosome 5q32 (rs2895680 in PPP2R2B-STK32A-DPYSL3, $P = 6.60 \times 10-9$) was lung cancer susceptibility loci and interacted with smoking dose [50]. As well as PPP2R2B is on the top of our result, the contribution of it in the NSCLC is worthy to be further elucidated.

ITGA9 (Integrin alpha-9) is related to the twelfth probe in our optimal probe set with a CNA fold change of 1.1034 comparing ACA with SCC, which belongs to the integrin family and is expressed on a wide range of cell types. It interacts with many ligands for example fibronectin, tenascin-C and ADAM12, and takes part in several processes such as cell adhesion, migration, lung development, lymphatic and venous valve development, and in wound healing [51]. ITGA9 has been found down expression in NSCLC [52], and exhibiting strong cell growth inhibition activity [53]. Statistical analysis of Alexey a. Dmitriev et al. suggested that the methylation/deletion level of ITGA9 has significant changes in ACA and SCC [53]. Our analysis presented the gene copy number of ITGA9 is dissimilar in NSCLC subtypes, implying ITGA9 as a candidate molecular to discriminate between SCC and ACA.

SOX2 (Sex-determining region Y-Box 2) is related to the nineteenth probe in our optimal probe set with a CNA fold change of 0.7790 comparing ACA with SCC, and has been reported to be differentially expressed between ACA and SCC. It is located at chromosome 3q26 and high-level amplification of SOX2 have been reported in approximately 20% of lung squamous cell carcinomas [54,55]. SOX2 is a transcription factor controlling the expression of a number of genes involved in embryonic development and keeps neural cells undifferentiated [56]. Suppression of SOX2 in amplified SOX2 cells has greater antiproliferative effects compared with other genes on 3q26.33 including PIK3CA and TP63.

FHIT (fragile histidine triad) is related to the thirty-third probe in our optimal probe set with a CNA fold change of 1.1110 comparing ACA with SCC, and behaves in vitro as a typical diadenosine triphosphate hydrolase cleaving A-5′-PPP-5′A to yield AMP and ADP [57], but little is known about its physiological function. It is considered as a tumor suppressor in many human cancers and its restoration in Fhit-negative cancer cell lines suppresses tumorigenicity and induces apoptosis [58]. Jennifer E. Tseng el at. found that the frequency of loss of FHIT expression is related with smoking habit in Stage I Non-Small Cell Lung Cancer [59]. In the studies of Gemma Toledo et al. FHIT expression was related to tumor histology: 52 of 54 (96.3%) SCC and 20 of 44 (45.5%) ACA were negative for FHIT ($P<0.0001$) [60]. As SCC is closely correlated with a history of tobacco smoking [6], and our results show the copy number of FHIT is significantly lower in SCC, FHIT may be a possible biomarker for NSCLC diagnosis and would be a potential medical target for cancer therapy.

RBBP8 (Retinoblastoma-binding protein 8) is a ubiquitously expressed nuclear protein which is binding to the tumor suppressor proteins RB [61] and CtBP [62]. It is also interacting with BRCA1 [63] and is thought to regulate the functions of BRCA1 in transcriptional regulation, DNA damage repair, and G2/M cell cycle checkpoint control [64,65]. RBBP8 is required for DNA double-strand break (DSB) resection, and thereby for recruitment of the protein kinase ATR and replication protein A to DSBs, and promotes ATR activation and homologous recombination [66]. It is reported that DNA repair components were significantly up-regulated including retinoblastoma-binding protein 8 (RBBP8), in lung SCC compared with normal lung tissue, but such up-regulation was not found in lung ACA [67]. As an essential molecular in the cell process DNA damage repair and cell cycle control, RBBP8 has the potential to be a biomarker and therapy target for NSCLC and the mechanism of its distinct expression profile in SCC and ACA deserves further study.

GPC5 (Glypican-5) is a member of the glypican gene family, which is a family of heparan sulphate proteoglycans that are linked to the exocytoplasmic surface of the plasma membrane via glycosyl phosphatidylinositol [68]. The expression level of GPC5 was significantly lower in lung adenocarcinoma tissue than in matched normal lung tissue in never smokers [69]. Yang et al. found deceased expression of GPC5 is correlated with reduced survival in ACA but not in SCC [70]. These all indicate that GPC5 may be a potential tumor suppressor gene in NSCLC, and a candidate bio-marker to discriminate between SCC and ACA.

Conclusion

In this study, we constructed a classifier based on copy number alterations (CNA) to distinguish two subgroups of NSCLC. As a result, 266 CNA probes were selected as the best discriminators. Analysis of genes corresponding to these 266 CNA probes indicate that they were enriched in lung cancer related pathways and enriched in the chromosome regions where CNA usually occur in lung cancer. Some of these genes, such as TP63, SOX2, EPHA4, PPP2R2B, ITGA9, FHIT, RBBP8 and GPC5 are closely related to lung cancer and these candidate genes may provide clues for further research and experiment validation.

Supporting Information

File S1 Clinical information of adenocarcinoma (ACA) and squamous cell carcinoma (SCC) samples.
(DOCX)

File S2 mRMR result for classification. This file contains two sheets. The first one is the MaxRel feature table, which ranked the top 1000 probes according to the relevance between features and class of the samples. The second one is the mRMR feature table, which ranked these 1000 probes according to the redundancy and relevance criteria.
(XLSX)

File S3 The sensitivity (Sn), specificity (Sp), accuracy (Ac), Matthews correlation coefficient (MCC) of each run of IFS for classification.
(XLSX)

File S4 The annotation of the 266 selected probes.
(XLSX)

File S5 The GO enrichment result of CNAs genes.
(XLSX)

Acknowledgments

The authors wish to thank the editor for taking time to edit this paper. The authors would also like to thank the two anonymous reviewers for their constructive comments, which were very helpful for strengthening the presentation of this study.

Author Contributions

Conceived and designed the experiments: TH YDC. Performed the experiments: BQL TH. Analyzed the data: BQL JY. Contributed reagents/materials/analysis tools: TH. Wrote the paper: BQL JY.

References

1. Siegel R, Naishadham D, Jemal A Cancer statistics, 2012. CA: A Cancer Journal for Clinicians 62: 10–29.
2. Travis WD, Brambilla E, Noguchi M, Nicholson AG, Geisinger KR, et al. (2011) International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. J Thorac Oncol 6: 244–285.
3. Perez-Moreno P, Brambilla E, Thomas R, Soria J-C Squamous Cell Carcinoma of the Lung: Molecular Subtypes and Therapeutic Opportunities. Clinical Cancer Research 18: 2443–2451.
4. Subramanian J, Govindan R (2007) Lung Cancer in Never Smokers: A Review. Journal of Clinical Oncology 25: 561–570.
5. Travis WD (2011) Pathology of Lung Cancer. Clinics in chest medicine 32: 669–692.
6. Kenfield SA, Wei EK, Stampfer MJ, Rosner BA, Colditz GA (2008) Comparison of aspects of smoking among the four histological types of lung cancer. Tobacco Control 17: 198–204.
7. Scagliotti G, Hanna N, Fossella F, Sugarman K, Blatter J, et al. (2009) The Differential Efficacy of Pemetrexed According to NSCLC Histology: A Review of Two Phase III Studies. The Oncologist 14: 253–263.
8. Cohen MH, Gootenberg J, Keegan P, Pazdur R (2007) FDA drug approval summary: bevacizumab (Avastin) plus Carboplatin and Paclitaxel as first-line treatment of advanced/metastatic recurrent nonsquamous non-small cell lung cancer. Oncologist 12: 713–718.
9. Sandler A, Gray R, Perry MC, Brahmer J, Schiller JH, et al. (2006) Paclitaxel–Carboplatin Alone or with Bevacizumab for Non–Small-Cell Lung Cancer. New England Journal of Medicine 355: 2542–2550.
10. Johnson DH, Fehrenbacher L, Novotny WF, Herbst RS, Nemunaitis JJ, et al. (2004) Randomized phase II trial comparing bevacizumab plus carboplatin and paclitaxel with carboplatin and paclitaxel alone in previously untreated locally advanced or metastatic non-small-cell lung cancer. J Clin Oncol 22: 2184–2191.
11. Travis WD, Rekhtman N, Riley GJ, Geisinger KR, Asamura H, et al. (2010) Pathologic Diagnosis of Advanced Lung Cancer Based on Small Biopsies and Cytology: A Paradigm Shift. Journal of Thoracic Oncology 5: 411–414 410.1097/JTO.1090b1013e3181d1057f1096e.
12. Kargi A, Gurel D, Tuna B (2007) The Diagnostic Value of TTF-1, CK 5/6, and p63 Immunostaining in Classification of Lung Carcinomas. Applied Immunohistochemistry & Molecular Morphology 15: 415–420 410.1097/PAI.1090b1013e31802fab31875.
13. Pao W, Chmielecki J (2010) Rational, biologically based treatment of EGFR-mutant non-small-cell lung cancer. Nat Rev Cancer 10: 760–774.
14. Kwak EL, Bang YJ, Camidge DR, Shaw AT, Solomon B, et al. (2010) Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. N Engl J Med 363: 1693–1703.
15. Baudis M (2007) Genomic imbalances in 5918 malignant epithelial tumors: an explorative meta-analysis of chromosomal CGH data. BMC Cancer 7: 226.
16. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, et al. (2010) The landscape of somatic copy-number alteration across human cancers. Nature 463: 899–905.
17. Huang Y-T, Lin X, Chirieac LR, McGovern R, Wain JC, et al. (2011) Impact on Disease Development, Genomic Location and Biological Function of Copy Number Alterations in Non-Small Cell Lung Cancer. PLoS ONE 6: e22961.
18. Tonon G, Wong K-K, Maulik G, Brennan C, Feng B, et al. (2005) High-resolution genomic profiles of human lung cancer. Proceedings of the National Academy of Sciences of the United States of America 102: 9625–9630.
19. Huang YT, Lin X, Chirieac LR, McGovern R, Wain JC, et al. (2011) Impact on disease development, genomic location and biological function of copy number alterations in non-small cell lung cancer. Plos One 6: e22961.
20. Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27: 1226–1238.
21. Friedman JH, Baskett F, Shustek LJ (1975) An algorithm for finding nearest neighbors. IEEE Transaction on Information Theory C-24: 1000–1006.
22. Denoeux T (1995) A k-nearest neighbor classification rule based on Dempster-Shafer theory. IEEE Transactions on Systems, Man and Cybernetics 25.: 804–813.
23. Li BQ, Huang T, Liu L, Cai YD, Chou KC (2012) Identification of colorectal cancer related genes with mRMR and shortest path in protein-protein interaction network. PLoS ONE 7: e33393.
24. Li B-Q, Hu L-L, Niu S, Cai Y-D, Chou K-C (2012) Predict and analyze S-nitrosylation modification sites with the mRMR and IFS approaches. Journal of Proteomics 75: 1654–1665.
25. Huang T, Chen L, Cai Y-D, Chou K-C (2011) Classification and analysis of regulatory pathways using graph property, biochemical and physicochemical property, and functional property. PLoS ONE 6: e25297.
26. Huang T, Cui W, Hu L, Feng K, Li YX, et al. (2009) Prediction of pharmacological and xenobiotic responses to drugs based on time course gene expression profiles. PLoS ONE 4: e8126.
27. Gao Y-F, Li B-Q, Cai Y-D, Feng K-Y, Li Z-D, et al. (2013) Prediction of active sites of enzymes by maximum relevance minimum redundancy (mRMR) feature selection. Molecular BioSystems.
28. Chou KC (2011) Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). Journal of Theoretical Biology 273: 236–247
29. Zhang N, Li B-Q, Gao S, Ruan J-S, Cai Y-D (2012) Computational prediction and analysis of protein [gamma]-carboxylation sites based on a random forest method. Molecular BioSystems 8: 2946–2955.
30. Huang T, Xu Z, Chen L, Cai YD, Kong X (2011) Computational analysis of HIV-1 resistance based on gene expression profiles and the virus-host interaction network. Plos One 6: e17291.
31. Li B-Q, Feng K-Y, Chen L, Huang T, Cai Y-D (2012) Prediction of Protein-Protein Interaction Sites by Random Forest Algorithm with mRMR and IFS. PLoS ONE 7: e43927.
32. Li B-Q, Hu L-L, Chen L, Feng K-Y, Cai Y-D, et al. (2012) Prediction of Protein Domain with mRMR Feature Selection and Analysis. PLoS One 7: e39308.
33. Li B-Q, Cai Y-D, Feng K-Y, Zhao G-J (2012) Prediction of Protein Cleavage Site with Feature Selection by Random Forest. PLoS One 7: e45854.
34. Chang JT, Nevins JR (2006) GATHER: a systems approach to interpreting genomic signatures. Bioinformatics 22: 2926–2933.
35. Uematsu K, He B, You L, Xu Z, McCormick F, et al. (0000) Activation of the Wnt pathway in non small cell lung cancer: evidence of dishevelled overexpression. Oncogene 22: 7218–7221.
36. You L, He B, Xu Z, Uematsu K, Mazieres J, et al. (2004) Inhibition of Wnt-2-mediated signaling induces programmed cell death in non-small-cell lung cancer cells. Oncogene 23: 6170–6174.
37. Berrier AL, Yamada KM (2007) Cell–matrix adhesion. Journal of Cellular Physiology 213: 565–573.
38. Frisch SM, Vuori K, Ruoslahti E, Chan-Hui PY (1996) Control of adhesion-dependent cell survival by focal adhesion kinase. J Cell Biol 134: 793–799.
39. Carelli S, Zadra G, Vaira V, Falleni M, Bottiglieri L, et al. (2006) Up-regulation of focal adhesion kinase in non-small cell lung cancer. Lung Cancer 53: 263–271.
40. Hanahan D, Weinberg Robert A (2011) Hallmarks of Cancer: The Next Generation. Cell 144: 646–674.
41. Pei J, Balsara BR, Li W, Litwin S, Gabrielson E, et al. (2001) Genomic imbalances in human lung adenocarcinomas and squamous cell carcinomas. Genes, Chromosomes and Cancer 31: 282–287.
42. Lockwood WW, Chari R, Coe BP, Thu KL, Garnis C, et al. Integrative genomic analyses identify BRF2 as a novel lineage-specific oncogene in lung squamous cell carcinoma. PLoS medicine 7: e1000315.
43. Flores ER, Tsai KY, Crowley D, Sengupta S, Yang A, et al. (2002) p63 and p73 are required for p53-dependent apoptosis in response to DNA damage. Nature 416: 560–564.
44. Kim MJ, Shin HC, Shin KC, Ro JY Best immunohistochemical panel in distinguishing adenocarcinoma from squamous cell carcinoma of lung: tissue microarray assay in resected lung cancer specimens. Annals of Diagnostic Pathology.
45. Pasquale EB Eph receptors and ephrins in cancer: bidirectional signalling and beyond. Nat Rev Cancer 10: 165–180.
46. Fukai J, Yokote H, Yamanaka R, Arao T, Nishio K, et al. (2008) EphA4 promotes cell proliferation and migration through a novel EphA4-FGFR1 signaling pathway in the human glioma U251 cell line. Molecular Cancer Therapeutics 7: 2768–2778.
47. Brannan JM, Dong W, Prudkin L, Behrens C, Lotan R, et al. (2009) Expression of the Receptor Tyrosine Kinase EphA2 Is Increased in Smokers and Predicts Poor Survival in Non-small Cell Lung Cancer. Clinical Cancer Research 15: 4423–4430.
48. Faoro L, Singleton PA, Cervantes GM, Lennon FE, Choong NW, et al. EphA2 Mutation in Lung Squamous Cell Carcinoma Promotes Increased Cell Survival, Cell Invasion, Focal Adhesions, and Mammalian Target of Rapamycin Activation. Journal of Biological Chemistry 285: 18575–18585.
49. Bennin DA, Don ASA, Brake T, McKenzie JL, Rosenbaum H, et al. (2002) Cyclin G2 Associates with Protein Phosphatase 2A Catalytic and Regulatory B′Subunits in Active Complexes and Induces Nuclear Aberrations and a G1/S Phase Cell Cycle Arrest. Journal of Biological Chemistry 277: 27449–27467.
50. Dong J, Hu Z, Wu C, Guo H, Zhou B, et al. Association analyses identify multiple new lung cancer susceptibility loci and their interactions with smoking in the Chinese population. Nat Genet 44: 895–899.
51. Høye AM, Couchman JR, Wewer UM, Fukami K, Yoneda A The newcomer in the integrin family: Integrin α9 in biology and cancer. Advances in Biological Regulation 52: 326–339.
52. Anedchenko EA, Dmitriev AA, Krasnov GS, Kondrat'eva TT, Kopantsev EP, et al. (2008) Down-regulation of RBSP3/CTDSPL, NPRL2/G21, RASSF1A, ITGA9, HYAL1 and HYAL2 genes in non-small cell lung cancer. Mol Biol (Mosk) 42: 965–976.
53. Dmitriev AA, Kashuba VI, Haraldson K, Senchenko VN, Pavlova TV, et al. Genetic and epigenetic analysis of non-small cell lung cancer with NotI-microarrays. Epigenetics 7: 502–513.
54. Hussenet T, Dali S, Exinger J, Monga B, Jost B, et al. SOX2 Is an Oncogene Activated by Recurrent 3q26.3 Amplifications in Human Lung Squamous Cell Carcinomas. Plos One 5: e8960.

55. Bass AJ, Watanabe H, Mermel CH, Yu S, Perner S, et al. (2009) SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas. Nat Genet 41: 1238–1242.
56. Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, et al. (2007) Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors. Cell 131: 861–872.
57. Barnes LD, Garrison PN, Siprashvili Z, Guranowski A, Robinson AK, et al. (1996) Fhit, a Putative Tumor Suppressor in Humans, Is a dinucleoside 5′,5‴-P1,P3-triphosphate Hydrolase. Biochemistry 35: 11529–11535.
58. Roz L, Andriani F, Ferreira CG, Giaccone G, Sozzi G (2004) The apoptotic pathway triggered by the Fhit protein in lung cancer cell lines is not affected by Bcl-2 or Bcl-x(L) overexpression. Oncogene 23: 9102–9110.
59. Tseng JE, Kemp BL, Khuri FR, Kurie JM, Lee JS, et al. (1999) Loss of Fhit Is Frequent in Stage I Non-Small Cell Lung Cancer and in the Lungs of Chronic Smokers. Cancer Research 59: 4798–4803.
60. Toledo G, Sola JJ, Lozano MD, Soria E, Pardo J (2004) Loss of FHIT protein expression is related to high proliferation, low apoptosis and worse prognosis in non-small-cell lung cancer. Mod Pathol 17: 440–448.
61. Fusco C, Reymond A, Zervos AS (1998) Molecular cloning and characterization of a novel retinoblastoma-binding protein. Genomics 51: 351–358.
62. Schaeper U, Subramanian T, Lim L, Boyd JM, Chinnadurai G (1998) Interaction between a cellular protein that binds to the C-terminal region of adenovirus E1A (CtBP) and a novel cellular protein is disrupted by E1A through a conserved PLDLS motif. J Biol Chem 273: 8549–8552.
63. Yu X, Chen J (2004) DNA damage-induced cell cycle checkpoint control requires CtIP, a phosphorylation-dependent binding partner of BRCA1 C-terminal domains. Mol Cell Biol 24: 9478–9486.
64. Yu X, Fu S, Lai M, Baer R, Chen J (2006) BRCA1 ubiquitinates its phosphorylation-dependent binding partner CtIP. Genes Dev 20: 1721–1726.
65. Greenberg RA, Sobhian B, Pathania S, Cantor SB, Nakatani Y, et al. (2006) Multifactorial contributions to an acute DNA damage response by BRCA1/BARD1-containing complexes. Genes Dev 20: 34–46.
66. Sartori AA, Lukas C, Coates J, Mistrik M, Fu S, et al. (2007) Human CtIP promotes DNA end resection. Nature 450: 509–514.
67. Daraselia N, Wang Y, Budoff A, Lituev A, Potapova O, et al. (2012) Molecular signature and pathway analysis of human primary squamous and adenocarcinoma lung cancers. Am J Cancer Res 2: 93–103.
68. Veugelers M, Vermeesch J, Reekmans G, Steinfeld R, Marynen P, et al. (1997) Characterization of glypican-5 and chromosomal localization of human GPC5, a new member of the glypican gene family. Genomics 40: 24–30.
69. Li Y, Sheu CC, Ye Y, de Andrade M, Wang L, et al. (2010) Genetic variants and risk of lung cancer in never smokers: a genome-wide association study. Lancet Oncol 11: 321–330.
70. Yang X, Zhang Z, Qiu M, Hu J, Fan X, et al. Glypican-5 is a novel metastasis suppressor gene in non-small cell lung cancer. Cancer Letters.