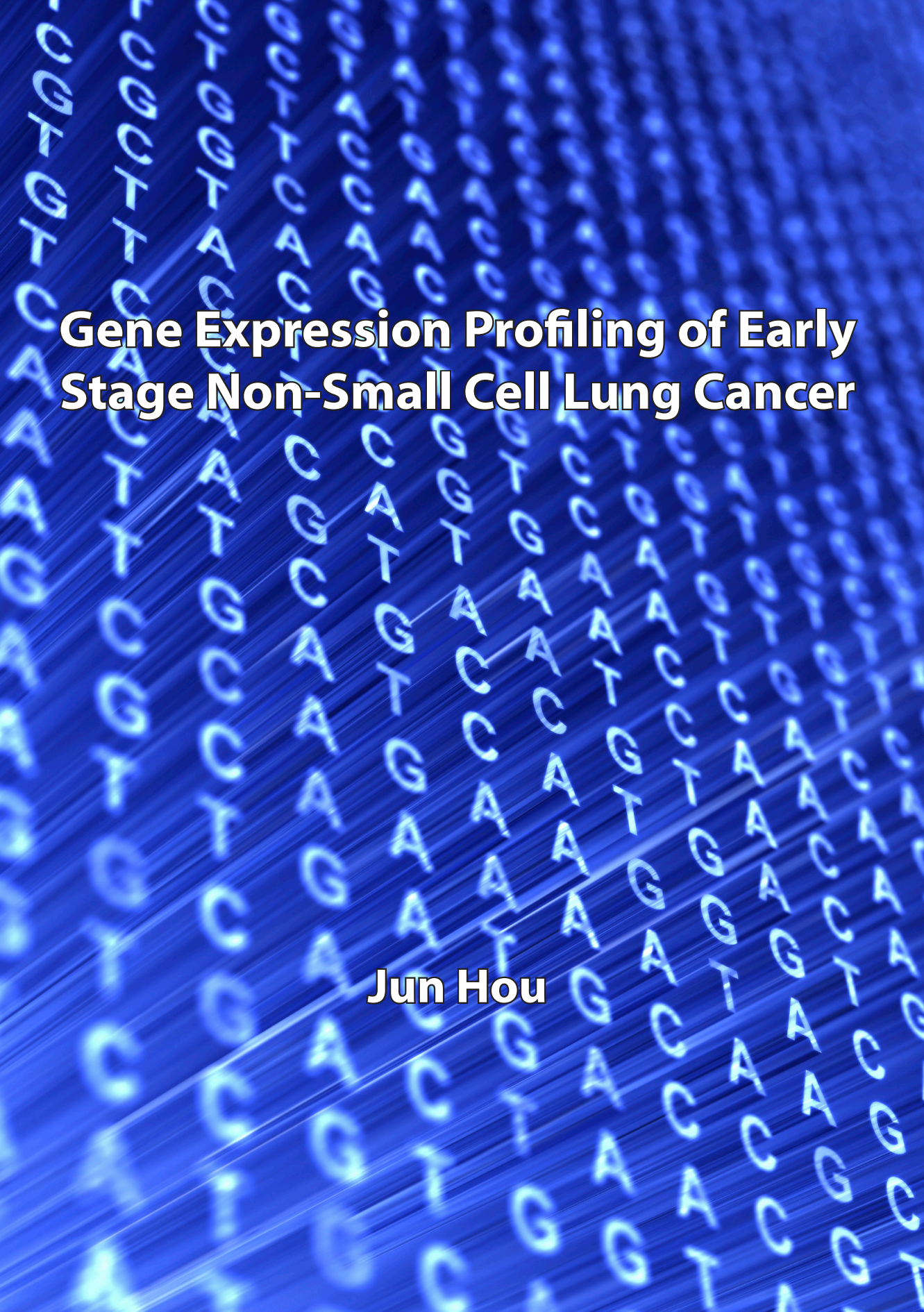


Gene Expression Profiling of Early Stage Non-Small Cell Lung Cancer

Jun Hou

Gene Expression Profiling of Early Stage Non-Small Cell Lung Caner

Jun Hou

侯珺

The work presented in this thesis was performed at the Department of Cell Biology, Erasmus University Medical Center, Rotterdam. The Department is member of the research schools Medisch Genetisch Centrum Zuid-West Nederland (MGC) and Molecular Medicine (MM).

The studies described in this thesis were financially supported by the Netherlands Genomics Initiative (NGI) through the Cancer Genomics Centre (CGC), Netherlands Bioinformatics Centre (NBIC), Netherlands Consortium for System Biology (NCSB) and Netherlands Proteomics Centre (NPC).

Cover photography: http://www.istockphoto.com
Printed by: Repro Océ Erasmus MC

Gene Expression Profiling of Early Stage Non-Small Cell Lung Cancer

Genexpressieprofielen van vroeg stadium niet-kleincellige longcarcinomen

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

Prof.dr. H.Schmidt

and in accordance with the decision of the Doctorate Board

The public defence shall be held on
Wednesday the 13th of January 2010 at 13:30 hrs

by

Jun Hou

born in LanZhou, China

Doctoral Committee

Promotors:	Prof.dr. J.N.J. Philipsen
	Prof.dr. F.G. Grosveld
Other members:	Prof.dr. P.J. van der Spek
	Dr.ir. P.M.J.J. Berns
	Dr.ir. G.W. Jenster
Copromotor:	Dr. J.G.J.V. Aerts

to my parents and Peng

Contents

List of abbreviations

AAH:	Atypical Adenomous Hyperplasia
AC:	Atypical Carcinoid
ADC:	Adenocarcinoma
BAC:	Bronchioalveolar Carcinoma
ChIP:	Chromatin Immunoprecipitation
CIN:	Chromosomal Instability
CIS:	Carcinoma in Situ
CNV:	Copy Number Variation
CT:	Computed Tomography
DIPNECH:	Diffuse Idiopathic Pulmonary Neuroendocrine Cell Hyperplasia
DLCO:	Diffusing Capacity
EGFR-TKI:	EGFR Tyrosine Kinase Inhibitor
FEV1:	Forced Expiratory Volume in 1 Second
IHC:	Immunohistochemistry
IRG:	Internal Reference Gene
LCC:	Large Cell Carcinoma
LCNEC:	Large Cell Neuroendocrine Carcinoma
MM:	Mismatch
MTT:	3-(4, 5-Dimethylthiazol-2-yl)-2, 5-diphenyltetrazolium bromide, a tetrazole that can be reduced to change colors in living cells; used in colorimetric assays
NE:	Neuroendocrine
NR:	Non-Responder
PAM:	Prediction Analysis of Microarrays
PET:	Positron Emission Tomography
PFS:	Progression Free Survival
PM:	Perfect Match
R:	Responder
REV:	Relative Expression Value
QC:	Quality Control

RR:	Response Rate
SAM:	Significant Analysis of Microarray
SCC:	Squamous Cell Carcinoma
SCLC:	Small Cell Carcinoma
SNP:	Single Nucleotide Polymorphism
TC:	Typical Carcinoid
TSG:	Tumor Suppressor Gene
TYMS:	Thymidylate Synthase

Scope of this study

NSCLC is a highly heterogeneous malignancy with a poor prognosis. Treatment for NSCLC is currently based on a combination of pathological staging and histological classification. Recently, gene expression-based NSCLC profiling is proven a superior approach to stratify cancer cases with different prognosis and to sub-classify patients in respect of response to chemotherapy.

The goal of the work described in this thesis is to explore the possible application of expression profiling in oncological practice. The proposed systematical scheme (Fig.1) is to distinguish NSCLC from non-cancerous lung using a Tumor signature; and subsequently to sub-classify NSCLC according to dominant molecular properties represented by a Histology signature. Next, the recognized cancer cases are subjected to be stratified in respect of prognosis according to a Survival signature. The estimated NSCLC behavior and outcome may direct clinicians in evaluating the aggressiveness of treatment and potential benefits the patient might have from that treatment. The application of expression profiling will be further investigated in a study in which NSCLC sensitivity to a chemotherapeutic agent, Pemetrexed is correlated to the expression of particular genes. The common feature of predicted resistant NSCLCs will be studied in the context of conventional histo-pathology and profiling-defined subgroups. The outcome of this study might indicate that molecular attributes of cancer cells improve insight into tumor physiology.

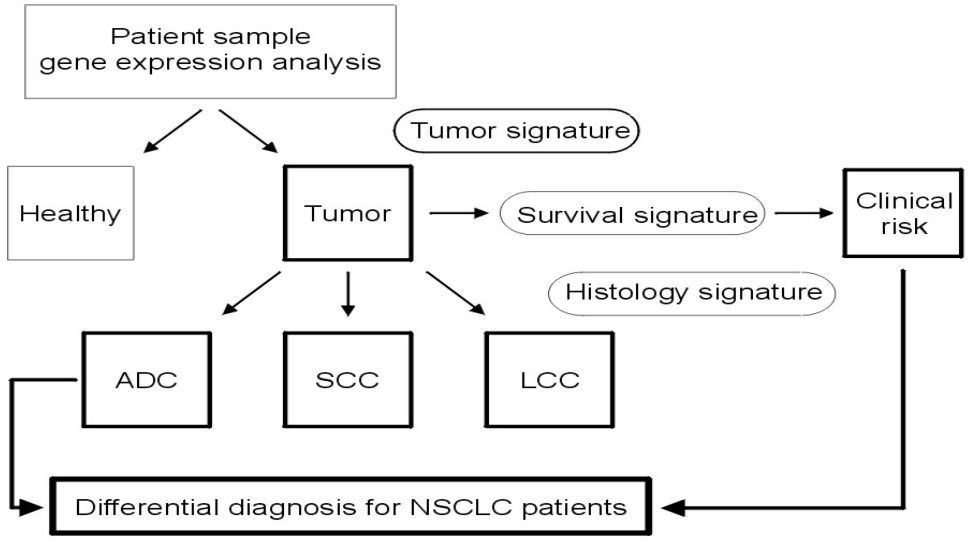


Fig. 1 Systematical scheme for differential diagnosis of NSCLC patients

Chapter 1

General Introduction to Non-small cell lung cancer

Lung cancer is a highly heterogeneous malignancy. The only hope for controlling this disease is to know about its etiology and pathogenesis. This study is trying to seek for new lung cancer diagnostic and prognostic biomarkers, and to understand molecular basis of chemotherapeutic sensitivity in the context of genome-wide expression profiling.

I. Epidemiology of lung cancer

The incidence of lung cancer varies by age, gender, race and smoking habit. Lung cancer is more prevalent in males than in females worldwide. For example, the age-standardized rate is 109 and 20 (per 100,000) between 1993 to 1997 in Dutch males and Dutch females respectively [1]. Worldwide, lung cancer is the most common cancer type in males, and the fourth in females by 2002 [1]. In European countries, it is the second most common cancer type in females after 1990's [2]. Lung cancer is the most common cause of cancer mortality in both males and females [3].

Non small cell lung cancer (NSCLC) accounts for more than 80% of all lung neoplasms in Europe [1]. In general, squamous cell carcinoma (SCC) is the most frequent subtype, followed by adenocarcinoma (ADC) and large cell carcinoma (LCC). However, in females a different trend is seen, with ADC being the most prevalent. About 2 to 4% of NSCLC display mixed features, most commonly composed of squamous and glandular elements.

Tobacco smoking has an established strong correlation with lung cancer development since the early 1960's [4]. It is reported that by 1985 approximately 85-90% of lung cancer cases worldwide were tobacco related [5]. The risk of lung cancer increases with the number of cigarettes smoked per day, the degree of inhalation, and the years spent smoking. Smoking increases the risk of all types of lung cancer, but in SCC the relationship is particularly strong [5]. In European countries, smoking still plays an important role in lung cancer development among females after the 1960's, with the percentage of smokers and incidence rate of lung cancer in females reaching the highest level in history [6]. However, in men, the situation has changed in the past 20 years. It is noted that the percentage of smokers among lung cancer patients has decreased since

the 1960's in men, which is probably also an explanation for the decrease in the incidence of squamous cell carcinoma [1]. The percentage of never smoking lung cancer patients is expected to decrease to 10-20% in the following years. The etiology of those lung cancer cases is ascribed to environmental factors such as population, but the exact mechanisms are unknown.

It is noticed that the relative incidence of the various subtypes of NSCLC has changed over time. SCC was formerly the most frequent subtype of NSCLC. In the past decades, ADC appears to be the predominant type in the United States [7]. The higher proportion of ADC among Americans is likely related to the changes in smoking habits, in particular to the use of filter-tip cigarettes. It has been shown that 97% of smokers were using low-yield filter cigarettes by 1992 [8]. Although filter-tip cigarettes contain less nicotine, smokers using this type of cigarette compensate for the lower delivery of nicotine by inhaling the smoke more deeply and retaining smoke longer. Consequently, the peripheral lung where ADC usually occurs is subjected to a higher deposition of smoke carcinogens [9]. In addition, low-tar filter cigarettes have a higher nitrate content, which induced ADC in laboratory animal models [9].

The higher proportion of women with ADC can also be explained by the smoking behavior discussed above.

A recently published study linked lung cancer incidence with low education, occupation, and low income, addressing the importance of socioeconomic status in lung cancer development [10].

II. **Histology**

The histological classification of lung tumors is based on light microscopic examination and standard histological staining techniques. There are four major types of lung tumor according to the World Health Organization Lung Cancer Classification: squamous cell carcinoma (SCC), large cell carcinoma (LCC), adenocarcinoma (ADC), and small cell carcinoma (SCLC). The complete classification is listed in Table 1. The former three are categorized as non-small cell lung cancer (NSCLC). NSCLC is more common than SCLC, accounting for approximately 80% of all lung cancers.

There is a great variability in biological behavior of NSCLC and SCLC. These different histopathological types of lung cancer grow and spread in different ways and are treated differently. SCLC is a very aggressive malignancy, with frequent widespread metastases at early course. It is characterized by properties such as contra-indication to surgery, poor prognosis, and relatively more sensitive to chemotherapy and radiotherapy, with a response rate more than 50% compared to 15% for NSCLC [11].

SCLC traditionally was not selected for surgical removal due to its propensity to early metastasis. However, more recently some studies suggest that the combination of adjuvant chemotherapy/radiotherapy and surgery improved prognosis of early stage SCLC patients [12].

In contrast, some of the NSCLC generally grows and spreads more slowly. These tumors are then less prone to develop early metastases and are amenable to surgical treatment at early stages.

Therefore, the classification as NSCLC or SCLC is of major clinical importance, as it significantly alters therapy guidelines. However it may be problematic to distinguish between these two major types of lung cancer. This may occur in up to 5 to 7% of cases [13]. Immunohistochemistry for some specific markers, such as Kit/CD117 for SCLC, might be helpful in such difficult diagnostic circumstances [13].

Within the category of NSCLC, histologic subtypes are not recruited in routine practice because firstly until recently the histological subtypes of NSCLC did not alter treatment; and secondly partially due to high inter-observer variability. Although sub-classification of NSCLC did alter little, if any, to the therapeutic strategy at present, the specific subtypes indicate certain patho-physiological and clinical patterns.

1. **Adenocarcinoma**

ADC is the most common subtype of NSCLC in the United States, and constitutes about 20 to 30%, with a increasing trend, of all NSCLC cases in Europe [7]. Being less strongly associated with smoking, adenocarcinoma is a predominant subtype in women in both the US and Europe [7].

It is usually peripheral and arises from cells that have glandular

or secretory properties. Central cavitation is uncommon. On histological examination, ADCs often demonstrate gland formation, papillary structures, or mucin production. On immunohistochemical examination, the tumors usually stain positive for TITF1/ NKX2-1 [14].

ADCs tend to metastasize to regional lymph nodes and to distant sites prior to the development of symptoms secondary to local disease.

According to WHO classification 2004, lung ADC is subtyped mainly to acinar, papillary, and bronchioloalveolar carcinoma. However both between cases and within individual tumors, ADC is the most histologically heterogeneous type of NSCLC. The majority of ADC presents a mixture of histological features that probably implicate different biological behaviours [15].

The bronchioloalveolar carcinoma (BAC) subtype of ADC arises in the periphery of the lung and grows along the alveolar septa. Most patients with BAC don't present with a history of smoking. It may occur as early as the second decade of life and is associated with prior lung fibrotic disease. Histologically, over half of BACs are non-mucin producing tumors and have a slightly better prognosis compared to conventional ADC [16].

Primary mucinous adenocarcinoma is a gather of any types of ADC with mucin secretion, including mucinous BAC, colloid ADC, and solid ADC. Morphologically, they resemble goblet cells. Upon immunohistochemical analysis, nonmucinous and mucinous BAC differ in CK20 and TITF1/ NKX2-1 staining [16].

Patients with adenocarcinomas may have an associated history of chronic lung disease, such as interstitial pneumonitis and recurrent pulmonary infections, and some necrotizing pulmonary diseases. In a small subset of ADC cases, patients present with disorders derived from endocrine and immune system, such as Trousseau's syndrome. Molecules originating from the tumor cells such as cytokines may play a pivotal role in the pathogenesis of concomitant paraneoplastic complications [17].

2. **Squamous Cell Carcinoma**

With a change in smoking habits since the 1960's, the incidence rate of SCC has decreased. The development of SCC is strongly associated with inhaled

carcinogens, as occurs in tobacco usage.

This type of tumor originates from epithelia lying in central bronchi. SCC tumors frequently present with central cavitation from necrosis and tumor cell exfoliation. This characteristic together with its central location makes it possible to detect SCC cells in sputum at an early stage.

In well-differentiated SCC cells, keratin formation, keratin pearl formation, and intercellular bridging between adjacent cells are frequently seen.

In a small proportion of SCC patients, increased secretion of a parathyroid-like hormone leads to hypercalcemia. Because of its central location, SCC tends to cause bronchial obstruction and atelectasis or pneumonia over time.

3. **Large cell carcinoma**

The histological diagnosis of LCC is largely based on exclusion of the other types of NSCLC, in other words, no evidence of squamous or glandular differentiation. Thus LCC is sometimes also referred to as poorly differentiated lung cancer. As a consequence, this subtype of NSCLC is highly heterogeneous in histopathology and clinical presentation. Furthermore, refined histopathological classification may lead to the diagnosis of ADC or SCC in cases that were previously diagnosed as LCC.

LCC accounts for 5 ~ 10% of NSCLC [18]. It is characterized by rapid growth, poor differentiation, and late distant spread. Histologically, these tumors show large cells with large nuclei and prominent nucleoli.

Because of the lack of well-defined classification criteria, LCC is usually considered as ADC with respect to therapeutic strategies [18].

4. **Neuroendocrine carcinomas**

NE includes several histological subtypes: small cell lung cancer (SCLC), large cell neuroendocrine carcinoma (LCNEC), typical carcinoid (TC), and atypical carcinoid (AC). They collectively account for around 20% of lung cancer cases [19]. In a broader view, 10 to 20% of histologically ordinary NSCLCs can be identified with neuroendocrine differentiation by use of immunohistochemistry or electron microscopy. The percentage of NE cells ranges from 3 to 25%, in

admixture with non-NE components in ordinary NSCLC [20, 21].

LCNEC presents most commonly as peripheral tumors as opposed to TC and AC, which generally are central in location. LCNEC is a high-grade malignant neoplasm with a relatively poor prognosis, 15 to 57% 5-year survival, compared to non-NE LCC [19]. Many patients with this type of tumor have developed distant metastases at the time of diagnosis. In contrast to LCNEC, TC is a low-grade and AC a medium-grade malignancy both showing a favorable outcome, with 5-year survival of ~87% in TC and 44 to 78% in AC [19]. Smoking appears to be an important factor for carcinogenesis of AC.

Under the microscope, these tumors present similar morphological appearances, such as organoid and rosette-like growth patterns, low nuclear to cytoplasmic ratio, a high mitotic rate, and neuro-secretory granules [22]. Since there is a considerable overlap in morphology and clinical presentation between NE tumors and other NSCLC, the routine histo-pathological examination is not sufficient and reliable to distinguish these tumors from others. Gene expression profiling might serve as a reliable mode to differentiate NE tumors from morphologically similar tumors, and also to differentiate within this category.

Clinically, these tumors are usually asymptomatic, if presented, the most common symptoms are cough, hemoptysis, and obstructive pneumonitis. There is no established standard therapy regimen for lung NE tumors due to the controversial data on survival advantage of chemotherapy from preclinical trials, and the low incidence of NE tumors [22].

The fact that the majority of lung cancers are histologically heterogeneous has an immediate impact on the prognosis of patients since the histological diagnosis and pathological staging directly guide treatments. This phenomenon is also a main reason for variation in inter- and intra-observer interpretation. Heterogeneity is observed in up to 60% of lung cancer cases. Difference in inter-observer interpretation as to histological subclassification due to histological heterogeneity may occur in up to 38% of examined resections [23].

Table 1. (2004) WHO Classification of Lung Tumors – Malignant Epithelial Tumors

I	Squamous cell carcinoma Variants: 1. Papillary 2. Clear cell 3. Small cell 4. Basaloid
II	Small-cell carcinoma Variants: 1. Combined small cell carcinoma
III	Adenocarcinoma 1. Acinar 2. Papillary 3. Bronchioloalveolar carcinoma o Non-mucinous (Clara/pneumocyte type II) o Mucinous o Mixed mucinous and non-mucinous 4. Solid adenocarcinoma with mucin 5. Adenocarcinoma with mixed subtypes 6. Variants • Well-differentiated fetal adenocarcinoma • Mucinous (colloid) adenocarcinoma • Mucinous cystadenocarcinoma • Signet-ring adenocarcinoma • Clear cell adenocarcinoma
IV	Large-cell carcinoma Variants: 1. Large cell neuroendocrine carcinoma • Combined large cell neuroendocrine carcinoma 2. Basaloid carcinoma 3. Lymphepithelioma-like carcinoma 4. Clear cell carcinoma 5. Large cell carcinoma with rhabdoid phenotype
V	Adenosquamous carcinoma

VI	Carcinomas with pleomorphic, sarcomatoid or sarcomatous elements 1. Carcinomas with spindle and/or giant cells • Pleomorphic carcinoma • Spindle cell carcinoma • Giant cell carcinoma 2. Carcinosarcoma 3. Pulmonary blastoma 4. others
VII	Carcinoid tumor 1. Typical carcinoid 2. Atypical carcinoid
VIII	Carcinomas of salivary-gland type 1. Mucoepidermoid carcinoma 2. Adenoid cystic carcinoma 3. Others
IX	Unclassified carcinoma

III. **Grading and staging**

Prognosis of NSCLC patients much depends on the staging. Curative interventions – surgery, radiotherapy, chemoradiation therapy, are options only for patients at early stages and without the evidence of distant metastasis, i.e. IA to IIIA.

Accurate staging, which assesses the extent of local and distant disease, is important for properly classifying curable patients and avoiding invasive treatment in those presenting with disseminated cancer cells.

To stage a NSCLC patient, a set of physical examinations and laboratory tests are performed.

- Complete history and physical examination
- CT scan of the chest and upper abdomen (including liver and adrenals)
- PET (Positron Emission Tomography) scan
- Bronchoscopy with transbronchial needle aspiration
- Mediastinal staging (endoscopic ultrasound, mediastinoscopy,

mediastinotomy, or thoracoscopy)
- Head CT in presence of symptoms of metastatic disease or evaluation appearing to be stage IIIA or IIIB
- Bone scanning in presence of bone-associated symptoms, or elevated calcium or alkaline phosphatase level
- Complete blood cell counts
- Liver and kidney functions tests
- Serum electrolytes
- Patient performance status, such as FEV_1 and DLCO, in patients who are candidates for surgical resection

The widely accepted staging of NSCLC is based on the TNM system ("T" for extent of primary tumor, "N" for regional lymph node involvement, and "M" for metastases) (Table 2). For clinical usage, the TNM descriptor is sometimes integrated with the stage grouping schemes of I to III, with A or B subtypes (Table 3).

While the histological classification relies on the observed best differentiation of the tumor under a microscope, NSCLC is graded by their most poorly differentiated portion. For example, a tumor, showing obvious evidence of pearl formation and intercellular bridges is classified as SCC. Moreover, if most of the remaining tumor cells lack these characteristics, the tumor is graded as poorly differentiated. However, in NSCLC it is common to see tumor cells in a variety of differentiation stages, with some cells relatively well differentiated while others being very immature.

IV. Prognosis

In the past few decades, a vast effort has been put into improvement of the prognosis of NCSLC. However, the overall 5-year postoperative survival of NSCLC still stays poor due to the fact that most patients with NSCLC have already developed advanced disease by the time of diagnosis. Secondly, based on current staging techniques patients presenting with only micrometastases cannot be distinguished from others of limited disease. The

Table 2. TNM Descriptors [24]

Descriptors	Definitions	Subgroups*
T	**Primary tumor**	
T0	No primary tumor	
T1	Tumor ≤ 3 cm,† surrounded by lung or visceral pleura, not more proximal than the lobar bronchus	
T1a	Tumor ≤ 2 cm†	T1a
T1b	Tumor > 2 but ≤ 3 cm†	T1b
T2	Tumor > 3 but ≤ 7 cm† or tumor with any of the following‡: Invades visceral pleura, involves main bronchus ≥ 2 cm distal to the carina, atelectasis/obstructive pneumonia extending to	
T2a	Tumor > 3 but ≤ 5 cm†	T2a
T2b	Tumor > 5 but ≤ 7 cm†	T2b
T3	Tumor > 7 cm;	$T3_{>7}$
	or directly invading chest wall, diaphragm, phrenic nerve, mediastinal pleura, or parietal pericardium;	$T3_{Inv}$
	or tumor in the main bronchus < 2 cm distal to the carina§;	$T3_{Centr}$
	or atelectasis/obstructive pneumonitis of entire lung;	$T3_{Centr}$
	or separate tumor nodules in the same lobe	$T3_{Satell}$
T4	Tumor of any size with invasion of heart, great vessels, trachea, recurrent laryngeal nerve, esophagus, vertebral body, or carina;	$T4_{Inv}$
	or separate tumor nodules in a different ipsilateral lobe	$T4_{Ipsi\ Nod}$
N	**Regional lymph nodes**	
N0	No regional node metastasis	
N1	Metastasis in ipsilateral peribronchial and/or perihilar lymph nodes and intrapulmonary nodes, including involvement by	
N2	Metastasis in ipsilateral mediastinal and/or subcarinal lymph nodes	
N3	Metastasis in contralateral mediastinal, contralateral hilar, ipsilateral or contralateral scalene, or supraclavicular lymph nodes	
M	**Distant metastasis**	
M0	No distant metastasis	
M1a	Separate tumor nodules in a contralateral lobe;	$M1a_{Contr\ Nod}$
	or tumor with pleural nodules or malignant pleural	$M1a_{Pl\ Dissem}$
M1b	Distant metastasis	M1b
Special situations		
TX, NX, MX	T, N, or M status not able to be assessed	
Tis	Focus of *in situ* cancer	Tis
T1§	Superficial spreading tumor of any size but confined to the wall of the trachea or mainstem bronchus	$T1_{SS}$

Table 3. Stage Grouping

T/M	Subgroup	N0	N1	N2	N3
T1	T1a	Ia	IIa	IIIa	IIIb
	T1b	Ia	IIa	IIIa	IIIb
T2	T2a	Ib	IIa	IIIa	IIIb
	T2b	IIa	IIb	IIIa	IIIb
T3	$T3_{>7}$	IIb	IIIa	IIIa	IIIb
	$T3_{Inv}$	IIb	IIIa	IIIa	IIIb
	$T3_{Satell}$	IIb	IIIa	IIIa	IIIb
T4	$T4_{Inv}$	IIIa	IIIa	IIIb	IIIb
	$T4_{Ipsi\ Nod}$	IIIa	IIIa	IIIb	IIIb
M1	$M1a_{Contra\ Nod}$	IV	IV	IV	IV
	$M1a_{Pl\ Disem}$	IV	IV	IV	IV
	M1b	IV	IV	IV	IV

poor prognosis also results from significant inter-subtype and intra-subtype variability in tumor progression and response to the treatment.

The prognosis of NSCLC is highly associated with TNM staging of NSCLC. This is to a great extent a result of stage-directed treatment strategy, with I-IIIA stage NSCLC being subject to surgery and more advanced stages to chemotherapy. In general, the estimated 5-year survival rates for each stages are as follows [25]:

- Stage IA – 68.5%
- Stage IB – 66.6%
- Stage IIA – 55.3%
- Stage IIB – 49.0%
- Stage IIIA – 35.8%
- Stage IIIB – 35.4%
- Stage IV - Not defined (3-year survival, 33.1%)

Among patients who have developed metastasis, those with M1 disease localized in the lung have a significant longer median survival than

those with metastasis in other organs. Similarly, patients with tumor cells metastasized to mediastinal nodal lymph nodes have a better prognosis than those with extra-thoracic metastases [25].

The clinical outcome of NSCLC patients also varies with histological subtypes and tumor cell components. BAC cells are relative indolent, hence ADC with BAC components present relatively slower tumor growth, and consequently demonstrate a longer survival compared to ADC without BAC features [26]. In general, ADC is recognized as the type of NSCLC with the highest mortality. Even patients eligible for operation have a median survival of only 24 months from the date of surgery; and up to 71% of the survivors have developed recurrence within 5 years. Median survival is only 7 months from the date of recurrence. In contrast, patients with stage I and II SCC have a better 5-year survival rate in comparison to patients with ADC [25].

Other factors which might associate with prognosis include smoking habits, gender, age, and the differentiation of cancer cells. The techniques and approaches employed to detect and histo-pathologically diagnose NSCLC affect the prognosis as well.

Although prognosis of NSCLC correlates with clinical situations such as pathological stage and histological subclass, the outcomes among patients at the same status can vary dramatically, even in early stage patients who had tumors surgically resected. Part of these patients could have a favorable 5 year postoperative survival, but others relapsed shortly after the initial surgery due to local or distant metastases. One explanation is that at the time of diagnosis and surgery those patients had already developed occult metastases that are undetectable by current routine examination. Therefore, it is vital to develop more sensitive and specific methods for detection of micrometastases.

On the other hand, expression profiling could be a more promising approach to predict patient outcome, and to identify patients who are unlikely to benefit significantly from certain therapeutic agents. A gene expression-based stratification successfully separated early stage NSCLC patients in respect of risk of recurrence into different groups [27], and was applied

through different types of histology. The possibility of refined assessment of clinical outcome for NSCLC patients holds the promise to greatly improve prognosis not only in early stage patients. NSCLC patients at high risk of relapse should receive adjuvant chemotherapy after initial surgery resection, while patients unlikely to develop disease progression would require less invasive treatment to preclude treatment related mortality and morbidity.

VI. Chemotherapy

The main objective of treatment of advanced NSCLC where surgery is not possible is to palliate symptoms and to improve quality of life, and also presently to modestly increase overall survival. Currently, chemotherapy and radiotherapy are commonly used treatments for advanced NSCLC.

Chemotherapy is the use of cytotoxic compounds to kill cancer cells or make them less active. Chemotherapeutic agents work by destroying rapidly dividing cells. Since cancer cells divide more frequently than most normal cells, they are particularly targeted by such drugs. Certain normal cells, such as those in hair follicles, the stomach epithelia, and bone marrow, divide at a comparable rate; therefore chemotherapy will eliminate such normal cells while killing cancer cells. The damage to normal cells results in the most common side-effects of chemotherapy, including hair loss, nausea, diarrhea, anemia, hemorrhage, and myelosuppression.

Since different chemotherapeutic agents target cells at different stages of cell division (Fig. 1), usually two or more agents are given at the same time to get the most efficient anti-rumor effect.

1. Timing of Chemotherapy

The timing of administration of chemotherapy differs depending on the specific aim. *Neoadjuvant chemotherapy* is given before the main treatment, such as surgery, to reduce the size of the tumor so as to facilitate the surgery more effectively. Conversely, *adjuvant chemotherapy* is given after the surgical removal of tumor, to reduce the risk of developing relapse due to occult cancer cells. In *concomitant chemotherapy* chemotherapy is administered at the same time as other therapies such as radiotherapy.

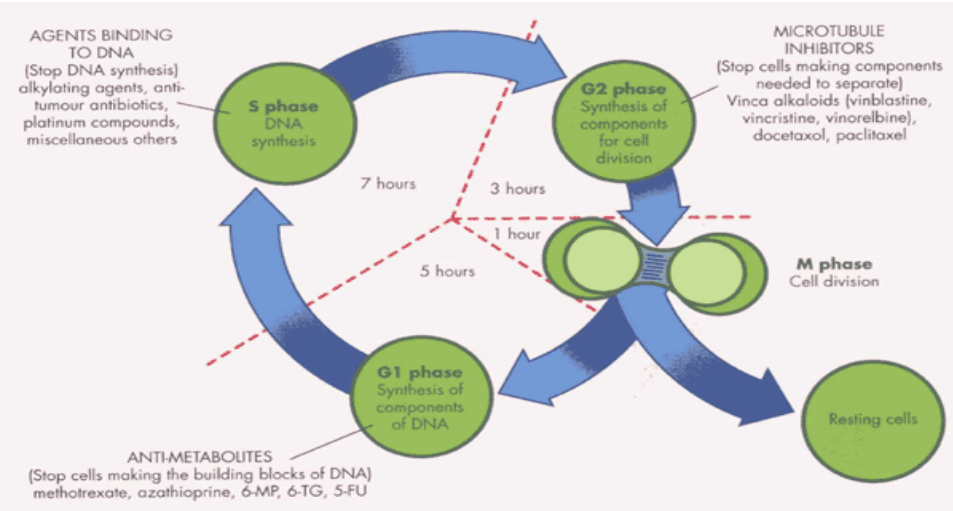


Fig. 1. Different chemotherapeutic agents target different stages of cell cycle (from www.drugdevelopment-technology.com)

First-line treatment represents the standard treatment (the "gold standard") given when a patient is diagnosed with a particular disease or condition, such as advanced NSCLC. Platinum-based doublet regimen has been recommended as the first line chemotherapy for NSCLC, including cisplatin + paclitaxel, cisplatin + gemcitabine, or cisplatin + docetaxel.

Despite the improvements in first-line chemotherapy regimens and the development of new drugs, over 60% of advanced patients experience progression and relapse. For patients with good clinical performance (PS of 0 or 1 by ECOG score), *second-line chemotherapy* is recommended. Docetaxel and more recently Pemetrexed and EGFR-TKI (Gefitinib and Erlotinib) are mostly recommended after the first-line failed in stage III to IV NSCLC patients.

Performance status (PS, ranging from 0 to 5) is a system that estimates a patient's general well-being by assessing his/her daily activities. Patients with PS score 0 or 1 is usually regarded as having good clinical performance status who are asymptomatic or symptomatic but completely ambulatory.

2. **Types of chemotherapy**

The majority of chemotherapeutic agents can be classified according to their

mechanism of action. Some agents target cells at a specific stage of cell cycle, such as MTX, 5-FU, bleomycin, paclitaxel, while others kill cells at any stage of the cell cycle. According to their site of action, chemotherapeutic agents can be divided into three main groups (Table 4). In general, chemotherapeutic agents are broad acting and display little specificity.

Table 4: Chemotherapeutic agents classification

Action site	medication	function
nucleotide synthesis	MTX, 5-FU, Pemetrexed	Thymidine synthesis
	6-MP, Pemetrexed	Purine synthesis
	cytarabine	Damage DNA; inhibit DNA synthesis
DNA, mRNA	Alkylating agents, platinum-based	DNA cross-links
	Dactinomycin, doxorubicin	DNA intercalation
	Bleomycin	strand breakage, DNA intercalation
	Etoposide	strand breakage
	steroids	may trigger apoptosis
Protein	Tamoxifen	steroid receptor antagonist
	Vinca alkaloids	inhibit microtubule formation
	Paclitaxel	inhibit microtubule disassembly

3. **Targeted therapy**

With the significant advances achieved in molecular biology, more critical genes in pathogenesis of NSCLC were identified that directed the development of a new class of therapeutic agents which manipulate specific biological pathways forming the basis of carcinogenesis of NSCLC. Among those agents include chemicals selectively working on particular genes, either established chemotherapeutic targets like thymidylate synthase or novel targets such as EGFR, as well as monoclonal antibodies against important proteins/enzymes from an oncogenic pathway.

Pemetrexed, a multi-target antimetabolite, is an anti-folate

pyrrolopyridimine-based chemical that exerts its anti-neoplastic activity by disrupting folate-dependent metabolic process such as pyrimidine biosynthesis, essential for all cell replication (Fig. 2). It is transported into cells via the reduced folate carrier (SLC19A1), or membrane folate binding proteins including the folate receptor (FOLR1). The penta-glutamated Pemetrexed is the active and predominant intracellular form which has a more potent inhibition of its targets, and increased half-life in cancer cells. Within the cell, Pemetrexed prevents the formation of DNA and RNA via inhibiting at least 3 enzymes in folate-dependent pathways: TYMS, DHFR, and GART, thereby blocking the proliferation of cells. Pemetrexed has been tested in advanced NSCLC patients from several clinical trials. The combination of Pemetrexed and cisplatin reached the highest response rate (RR) of 39% and 45% in two phase II studies, accompanying a median survival of 10.9 and 8.9 months. If folic acid or vitamin B12 was used as a supplement, the drug-related toxicity was kept at a very low level. The observations from these clinical studies indicate Pemetrexed as a promising agent for NSCLC treatment [28].

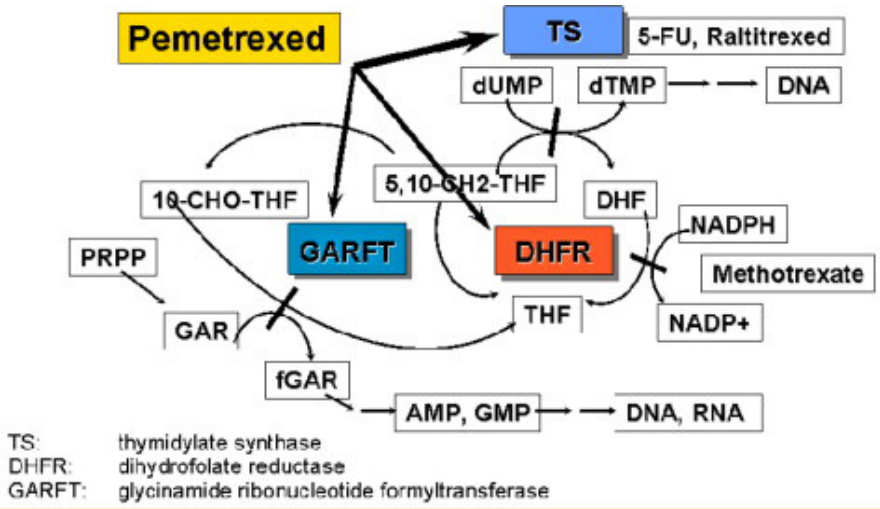


Fig. 2. Mechanism of action of Pemetrexed (from [29])

Epidermal growth factor receptor tyrosine kinase inhibitors (EGFR-TKIs), gefitinib and erlotinib, are used as first-line therapy in combination with other agents. In addition, EGFR-TKI showed equal activity in second

line compared to standard chemotherapy (Fig. 3). They have an acceptable tolerability profile, with an 18% RR, prolonged median survival, and a significant improvement in disease-related symptoms and quality of life [30]. Large-scale pilot studies confirmed that addition of EGFR-TKI's to standard chemotherapy has not contributed to a survival increase in advanced NSCLC patients. Recent evidence shows, however, that maintenance treatement of Erlotinib was associated with a survival befefit in NSCLC patients.

Overexpression of EGFR is observed in smoking related NSCLC and is associated with poor prognosis [31]. In vitro studies showed that EGFR tyrosine kinase inhibitors blocked the growth of human NSCLC cell lines via inhibition of the phosphorylation of the intracellular part of receptor and downstream proteins [30].

In addition, monoclonal antibodies against VEGF as well as PKC inhibitors reached a median PFS of 6.7 months in recurrent NSCLC, when combined with chemotherapy in preclinical trials [32].

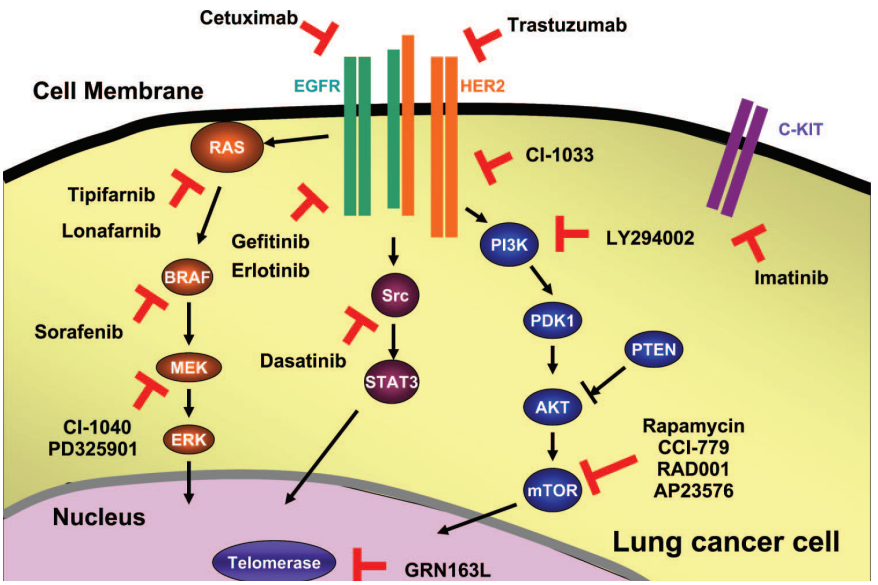


Fig. 3. Cancer drugs targeting major growth transduction pathways involved in lung cancer pathogenesis (from [33])

4. Strategy of choosing chemotherapeutic regimen

The traditional development of therapeutic regimens is largely empirical.

Although with the appearance of new medications the treatment approach is more selective, it is still far away from personalized therapy targeting distinctive properties of cancer cells.

The administration of Pemetrexed is dependent on the histology of NSCLC. Clinical trials showed that Pemetrexed is unlikely to give better remission rates in SCC patients compared to standard regimens, while improved responses were seen in ADC and LCC cases. The response to Pemetrexed might be dependent on differential expression of TYMS between histology types [34]. Accordingly, this multi-target antifolate agent is approved for treatment of non-SCC NSCLC where a higher RR, improved overall survival, and milder drug toxicity were seen [34].

Similarly, targeted chemotherapy for example EGFR-TKI, holds promise to improve the overall survival of patients with advanced NSCLC. However, the response to EGFR-TKI or EGFR antibodies varies widely among patients with comparable tumor histopathology. Preclinical studies showed contradictory effects of EGFR-TKI treatment in advanced NSCLC patients [35]. Two large-scale studies of first-line therapy for NSCLC where the EGFR antibody cetuximab was given in combination with carboplatin-taxane or cisplatin-vinorelbine indicated that RR as well as induced PFS and overall survival were significantly improved among patients with EGFR mutations [36]. The similar results were also found in first-line treatment with EGFR-TKI in advanced NSCLC patients [37, 38]. The lack of efficacy in the former studies therefore suggests that EGFR-TKI treatment may have little or no effect in first-line treatment in unselected patients.

A molecular rationale is required to explain the observed low RR and large variation in patient response. This probably resides in differentially activated oncogenic pathways, not necessarily connected to histological tumor subtype.

5. **Molecular prediction of response to chemotherapy**

The chemosensitivity profile has been studied for decades. To date, most of predictive factors used in practice for chemotherapy outcome are clinical characteristics, for instance gender, ethnic background, and tumor stage.

It has been suggested that the performance of molecular predictors may exceed conventional clinical parameters. The level of DNA repair enzymes was correlated with the efficacy of chemotherapeutic agents acting directly on the DNA, including alkylating agents and platinum-based compounds. Furuta et al., found that ERCC1, a nucleotide excision repair enzyme, is overexpressed in cisplatin-resistant cancer patients, indicating that the efficiency of removing cisplatin-induced DNA adducts by nucleotide repair enzymes plays a central role in cisplatin resistance [39]. Similarly, the expression level of RRM1, a rate-limiting enzyme in DNA synthesis, was linked to NSCLC sensitivity to gemcitabine, which exerts its antitumor activity by incorporation into DNA and subsequently arresting cell growth [40].

Expression studies on the response of NSCLC to Pemetrexed indicated that TYMS and associated genes, including DHFR, GART, and MRP4, are important determinants for RR, and expression levels correlated negatively with disease-free time to progression and overall survival [41-44].

6. **Future developments: pharmacogenomics**

With the identification of global patterns of gene expression, the application of chemotherapy is expected to be tailored to the genetic changes of individual tumors in order to obtain maximal efficacy and minimal toxicity.

New technologies, such as microarrays and high throughput sequencing, allow for rapid identification of association between transcriptome abnormalities and tumor behavior to different therapeutic medications. Hsu et al., studied the chemosensitivity profiles of NSCLC cell lines in the NCI-60 screening program and discovered a panel of 45 genes as determinants for response of NSCLC cell lines to Pemetrexed [44]. In another study by Hanauske et al., the expression of targets of Pemetrexed as well as the molecules involved in in vivo metabolism of this drug was measured and their relationship with tumor reaction to Pemetrexed treatment was determined [43].

The predictive role of TYMS and MRP4 for NSCLC response to antifolate metabolites was revealed in this study, raising the potential to apply expression of these proteins to guide decisions on Pemetrexed administration in clinical practice.

Table 5. Genomic methodologies and their application in pharmacogenomics (adapted from [45])

Technology	Application	Advantages	Disadvantages
CGH Array	Scanning for DNA deletion, amplification	Can analyze thousands of precisely mapped loci	Tedious custom manufacture from BAC clones
SNP array	Detection of DNA polymorphisms	Detects single base changes, correlation of alleles with phenotype	PCR-based technologies cumbersome
Spotted cDNA microarray	Quantitate the expression of mRNAs	Inexpensive, ability to customize	Lesser density, infrastructure costs
Oligonucleotide microarray	Quantitate the expression of mRNAs	High density, commercial support, off shelf software	Expensive, high startup costs, inability to customize
Proteomics – 2D gels	Quantitate the expression and modification of proteins	Discerns large range of protein molecular weights	Large cumbersome gels, large amount of protein, difficult to align spots across gels, only approximate molecular weights
Proteomics – mass spectrometry	Quantitate the expression and modification of proteins	High sensitivity, highly accurate molecular weight determination	High equipment cost, limited molecular weight range.
Tissue microarray	Rapid screening of antibodies or FISH probes in paraffin sections	Inexpensive, can analyze hundreds of samples, gives cell relationships	Requires paraffin blocks, difficult to score, limited number of sections from one block

References

1. Janssen-Heijnen, M.L. and J.W. Coebergh, *The changing epidemiology of lung cancer in Europe.* Lung Cancer, 2003. **41**(3): p. 245-58.
2. Parkin, D.M., et al., *Global cancer statistics, 2002.* CA Cancer J Clin, 2005. **55**(2): p. 74-108.
3. Kosacka, M. and R. Jankowska, *[The epidemiology of lung cancer].* Pneumonol Alergol Pol, 2007. **75**(1): p. 76-80.
4. Doll, R. and A.B. Hill, *Smoking and carcinoma of the lung; preliminary report.* Br Med J, 1950. **2**(4682): p. 739-48.
5. Tyczynski, J.E., F. Bray, and D.M. Parkin, *Lung cancer in Europe in 2000: epidemiology, prevention, and early detection.* Lancet Oncol, 2003. **4**(1): p. 45-55.
6. Bray, F., J.E. Tyczynski, and D.M. Parkin, *Going up or coming down? The changing phases of the lung cancer epidemic from 1967 to 1999 in the 15 European Union countries.* Eur J Cancer, 2004. **40**(1): p. 96-125.
7. Charloux, A., et al., *International differences in epidemiology of lung adenocarcinoma.* Lung Cancer, 1997. **16**(2-3): p. 133-43.
8. Wynder, E.L. and J.E. Muscat, *The changing epidemiology of smoking and lung cancer histology.* Environ Health Perspect, 1995. **103 Suppl 8**: p. 143-8.
9. Thun, M.J., et al., *Cigarette smoking and changes in the histopathology of lung cancer.* J Natl Cancer Inst, 1997. **89**(21): p. 1580-6.
10. Sidorchuk, A., et al., *Socioeconomic differences in lung cancer incidence: a systematic review and meta-analysis.* Cancer Causes Control, 2009. **20**(4): p. 459-71.
11. Spiro, S.G. and G.A. Silvestri, *One hundred years of lung cancer.* Am J Respir Crit Care Med, 2005. **172**(5): p. 523-9.
12. Panagopoulos, N., et al., *Low incidence of bronchopleural fistula after pneumonectomy for lung cancer.* Interact Cardiovasc Thorac Surg, 2009.
13. Butnor, K.J., et al., *The spectrum of Kit (CD117) immunoreactivity in lung and pleural tumors: a study of 96 cases using a single-source antibody with a review of the literature.* Arch Pathol Lab Med, 2004. **128**(5): p. 538-43.
14. Beasley, M.B., *Immunohistochemistry of pulmonary and pleural neoplasia.* Arch Pathol Lab Med, 2008. **132**(7): p. 1062-72.
15. *Pathology of lung cancer* http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=cmed.section.20772. (Date last updated: 2000.).
16. Yousem, S.A. and M.B. Beasley, *Bronchioloalveolar carcinoma: a review of current concepts and evolving issues.* Arch Pathol Lab Med, 2007. **131**(7): p. 1027-32.
17. Sato, T., et al., *Trousseau's syndrome associated with tissue factor produced by pulmonary adenocarcinoma.* Thorax, 2006. **61**(11): p. 1009-10.
18. Tiseo, M., et al., *First-line treatment in advanced non-small-cell lung cancer: the emerging role of the histologic subtype.* Expert Rev Anticancer Ther, 2009. **9**(4): p. 425-35.
19. Gustafsson, B.I., et al., *Bronchopulmonary neuroendocrine tumors.* Cancer, 2008. **113**(1): p. 5-21.
20. Brownson, R.C., et al., *Lung cancer in nonsmoking women. Histology and survival patterns.* Cancer, 1995. **75**(1): p. 29-33.
21. Righi, L., et al., *Neuro-endocrine tumours of the lung. A review of relevant pathological and molecular data.* Virchows Arch, 2007. **451 Suppl 1**: p. S51-9.

22. Fernandez, F.G. and R.J. Battafarano, *Large-cell neuroendocrine carcinoma of the lung.* Cancer Control, 2006. **13**(4): p. 270-5.
23. Roggli, V.L., et al., *Lung cancer heterogeneity: a blinded and randomized study of 100 consecutive cases.* Hum Pathol, 1985. **16**(6): p. 569-79.
24. Detterbeck, F.C., D.J. Boffa, and L.T. Tanoue, *The new lung cancer staging system.* Chest, 2009. **136**(1): p. 260-71.
25. Pfannschmidt, J., et al., *Prognostic assessment after surgical resection for non-small cell lung cancer: experiences in 2083 patients.* Lung Cancer, 2007. **55**(3): p. 371-7.
26. Lee, K.S., et al., *T1 non-small cell lung cancer: imaging and histopathologic findings and their prognostic implications.* Radiographics, 2004. **24**(6): p. 1617-36; discussion 1632-6.
27. Potti, A., et al., *A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer.* N Engl J Med, 2006. **355**(6): p. 570-80.
28. Rosell, R., et al., *The biology of non-small-cell lung cancer: identifying new targets for rational therapy.* Lung Cancer, 2004. **46**(2): p. 135-48.
29. Hanauske, A.R., et al., *Pemetrexed disodium: a novel antifolate clinically active against multiple solid tumors.* Oncologist, 2001. **6**(4): p. 363-73.
30. Silvestri, G.A. and M.P. Rivera, *Targeted therapy for the treatment of advanced non-small cell lung cancer: a review of the epidermal growth factor receptor antagonists.* Chest, 2005. **128**(6): p. 3975-84.
31. Hirsch, F.R., et al., *Epidermal growth factor receptor in non-small-cell lung carcinomas: correlation between gene copy number and protein expression and impact on prognosis.* J Clin Oncol, 2003. **21**(20): p. 3798-807.
32. Sculier, J.P. and D. Moro-Sibilot, *First- and second-line therapy for advanced nonsmall cell lung cancer.* Eur Respir J, 2009. **33**(4): p. 915-30.
33. Sato, M., et al., *A translational view of the molecular pathogenesis of lung cancer.* J Thorac Oncol, 2007. **2**(4): p. 327-43.
34. Rossi, A., et al., *Pemetrexed in the treatment of advanced non-squamous lung cancer.* Lung Cancer, 2009. **66**(2): p. 141-9.
35. Socinski, M.A., *Antibodies to the epidermal growth factor receptor in non small cell lung cancer: current status of matuzumab and panitumumab.* Clin Cancer Res, 2007. **13**(15 Pt 2): p. s4597-601.
36. Tiseo, M., et al., *Predictors of gefitinib outcomes in advanced non-small cell lung cancer (NSCLC): Study of a comprehensive panel of molecular markers.* Lung Cancer, 2009.
37. Jiang, H., *Overview of gefitinib in non-small cell lung cancer: an Asian perspective.* Jpn J Clin Oncol, 2009. **39**(3): p. 137-50.
38. Jackman, D.M., et al., *Impact of epidermal growth factor receptor and KRAS mutations on clinical outcomes in previously untreated non-small cell lung cancer patients: results of an online tumor registry of clinical trials.* Clin Cancer Res, 2009. **15**(16): p. 5267-73.
39. Furuta, T., et al., *Transcription-coupled nucleotide excision repair as a determinant of cisplatin sensitivity of human cells.* Cancer Res, 2002. **62**(17): p. 4899-902.
40. Davidson, J.D., et al., *An increase in the expression of ribonucleotide reductase large subunit 1 is associated with gemcitabine resistance in non-small cell lung cancer cell lines.* Cancer Res, 2004. **64**(11): p. 3761-6.
41. Bepler, G., et al., *Clinical efficacy and predictive molecular markers of neoadjuvant gemcitabine and pemetrexed in resectable non-small cell lung cancer.* J Thorac Oncol,

2008. **3**(10): p. 1112-8.
42. Giovannetti, E., et al., *Cellular and pharmacogenetics foundation of synergistic interaction of pemetrexed and gemcitabine in human non-small-cell lung cancer cells.* Mol Pharmacol, 2005. **68**(1): p. 110-8.
43. Hanauske, A.R., et al., *In vitro chemosensitivity of freshly explanted tumor cells to pemetrexed is correlated with target gene expression.* Invest New Drugs, 2007. **25**(5): p. 417-23.
44. Hsu, D.S., et al., *Pharmacogenomic strategies provide a rational approach to the treatment of cisplatin-resistant patients with advanced cancer.* J Clin Oncol, 2007. **25**(28): p. 4350-7.
45. Franklin, W.A. and D.P. Carbone, *Molecular staging and pharmacogenomics. Clinical implications: from lab to patients and back.* Lung Cancer, 2003. **41 Suppl 1**: p. S147-54.

Chapter 2

NSCLC Carcinogenesis

The abnormalities contributing to cancer initiation, development, and progression include genomic instability, such as large-scale chromosomal instability (CIN) and microsatellite instability (MSI); altered expression of oncogenes and tumor suppressor genes (TSG), like EGFR, KRAS, MYC, TP53, and RB; and resultant dysregulation of downstream signal pathways, including MAPK, PKC, and PI3K/AKT signalling; epigenetic changes - promoter hypermethylation of TSG and histone deacetylation, sustained angiogenesis and high rate of mitosis.

Underlying the transformation of normal phenotypes into malignant phenotypes is the accumulation of multiple genetic and epigenetic alterations. Critical changes include oncogene activation, repression of suppressor genes, and loss of cell cycle- and apoptosis control. In addition, tumor development is characterized by genomic instability, such as chromosomal translocation. Previous studies showed that although some abnormalities are widely seen in all types of NSCLC, there are certain molecular alterations that are distinct for different histologic types.

I. Morphologic preneoplastic changes

As with other epithelial malignancies, the development of NSCLC involves a series of progressive morphologic changes in respiratory epithelium. According to the histological classification of pre-invasive lung lesions by the WHO, the three major morphologic preneoplastic conditions include [1].

1. Squamous dysplasia and carcinoma in situ (CIS),
2. Atypical adenomatous hyperplasia (AAH),
3. Diffuse idiopathic pulmonary neuroendocrine cell hyperplasia (DIPNECH).

However, these lesions only account for the development of a subset of NSCLC.

The sequential premalignant lesions in central bronchi epithelium are well defined [2] (Fig.1),

1. Reserve cell hyperplasia,
2. Squamous metaplasia,
3. Low-grade dysplasia (mild dysplasia),

4. High-grade dysplasia (moderate or severe dysplasia),
5. Carcinoma in situ (CIS)

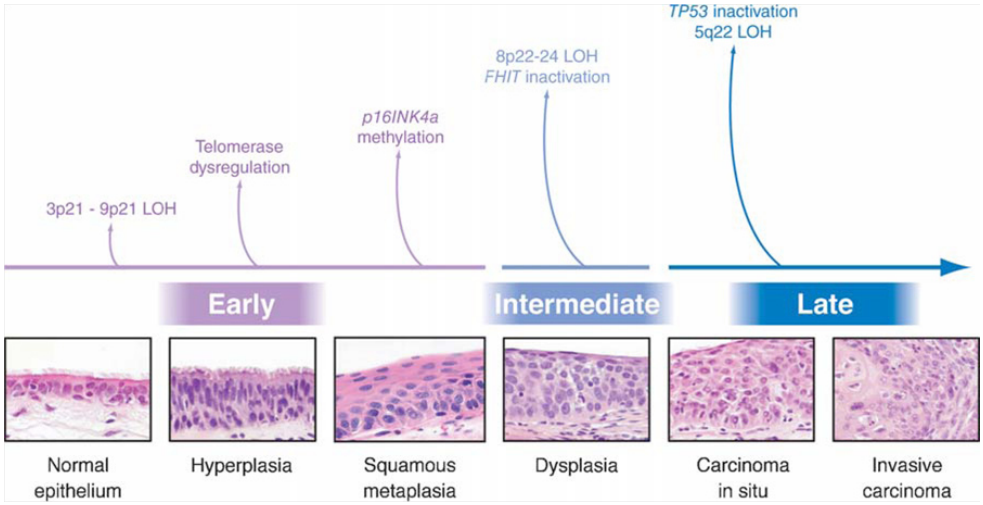


Fig. 1. Sequential histopathological and molecular changes during the pathogenesis of squamous cell carcinoma of the lung [3].

Hyperplasia of the bronchial epithelium and squamous metaplasia are regarded reversible, whereas displasia and CIS frequently proceed to invasive squamous cell carcinoma. About 25% of dysplasia and 50% of CIS were found to progress to invasive SCC within 30 to 36 months [2].

The carcinogenesis sequence for other types of NSCLC, however, has been poorly documented. AAH in peripheral airways, where ADC always arises, is supposed as the initial morphologic change in glandular neoplasm development [1]. In most cases, AAH arises in alveoli and are lined by cuboidal or columnar cells, however, the lesions are frequently found heterogeneous and might present multiple atypical proliferations [1]. In addition, little is known about the specific cell types involved in the tumorigenesis of most lung adenocarcinomas. If AAHs cannot regress, they progress to bronchioloalveolar carcinoma (BAC) and invasive peripheral ADC (Fig. 2).

Although the precursor lesions are unknown for lung NEU, DIPNECH has been proposed as the premalignant change of other pulmonary neuroendocrine tumors, including TC and AC [1]. The hypothesis is based

on observations in some patients bearing carinoid tumors. In those patients, hyperplasia of neuroendocrine cells is always seen concurrently with carcinoid tumors. Thus, DIPNECH could be a preinvasive lesion, which might give rise to atypical or typical carcinoid. The presumed development of carcinoid tumors is following sequential steps: hyperplasia of airway neuroendocrine cells; hyperplasia invading the epithelial basement membrane; further interstitially-extended lesion associated with fibrous tissues, also characterized as carcinoid tumorlet; carcinoid tumors, defined as tumorlets exceeding 5 mm in diameter.

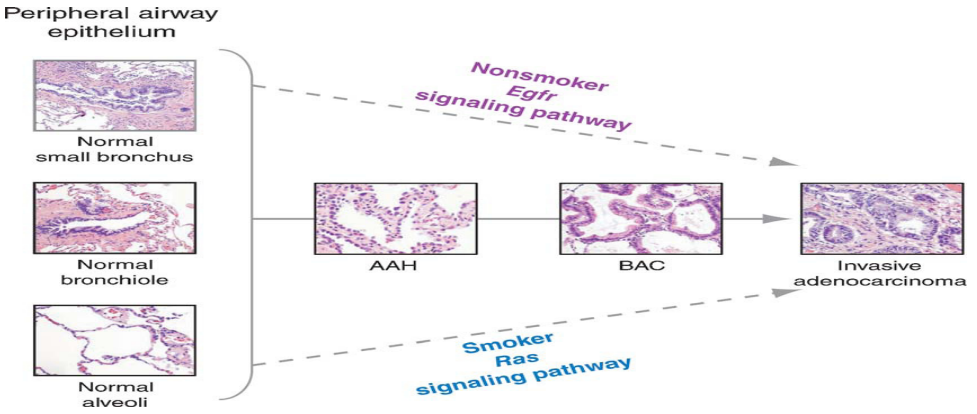


Fig. 2. Two molecular pathways involved in the development of lung adenocarcinoma [3]

II. **Cell origin**

The classical model of carcinogenesis is a multiple step process involving successive accumulation of multiple genetic and epigenetic alterations, leading to transformation of normal epithelial cells to preneoplastic cells. This hypothesis is supported by extensive evidences that NSCLCs harbour multiple genetic and epigenetic abnormalities. In addition, preneoplastic cells and histologically normal bronchial epithelia present with many same types of abnormalities. Therefore, it is suggested that lung cancer develops from normal epithelial cells acquiring multiple molecular alterations.

Variable locations of NSCLC may also indicate NSCLC originates from local normal epithelia cells (Fig. 3). SCC usually arises centrally from major bronchi where airways are lined by columnar ciliated epithelia and goblet cells. Considering there are no squamous cells in the normal airways, basal cells

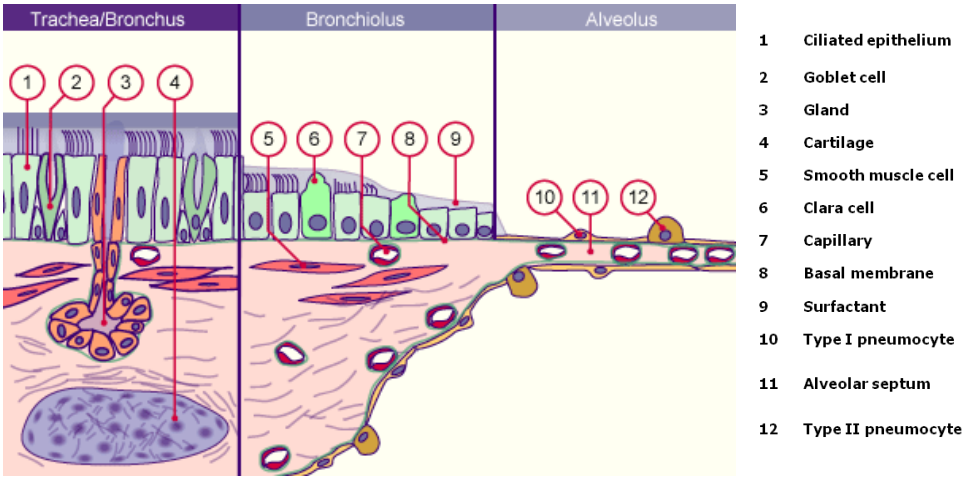


Fig. 3. Epithelial cells present in a normal adult lung (from www.embryology.ch)

are presumed progenitors of preneoplastic epithelium for central located SCC [3]. ADC and LCC usually arise peripherally from small bronchi, bronchioles, or alveoli of distant airways. Clara cells, and type I and II pneumocytes are predominant in bronchioles and alveoli. The Clara cells and type II pneumocytes are responsible for the synthesis of surfactant protein A (SP-A), which is used as specific markers for these cell types. Type II cells are believed to be progenitor cells from which ADCs originate [4]. The hypothesis is based on the observation that some premalignant lesions and ADC cells showed phenotypes of type II cells. Specific markers for type II cells were present as well in almost all AAH, ADC using IHC staining [4].

A fundamentally different hypothesis, cancer stem cell model, has become more accredited recently. This hypothesis advocates that a cancer stem cell or pluripotent cell is a common precursor for all types of lung tumors. Cancer stem cells possess stem cell like properties, including self-renewal and multilineage differentiation. As a consequence, tumors arisen from them retain the ability to undergo cell division to give rise to more cancer stem cells, subsequently differentiate to phenotypically diverse tumor cell population. Evidence for existence of cancer stem cells was reported in hematologic malignancies, breast cancer, and CNS tumors. Lung cancer stem cells have not been reported yet. However, a stem cell population which had the ability

of self-renewal and differentiation was isolated from the bronchioalveolar duct junction in the Kras mouse model [5]. Although more work is needed to prove the existence of lung cancer stem cells and to characterize their role in carcinogenesis, the heterogeneity and molecular complexity found in various types of lung cancer, particularly in ADC, LCC and SCLCs, indirectly support the cancer stem cell hypothesis.

Two hypotheses are not necessary mutually exclusive, they might coexist in and/or cooperatively function for lung carcinogenesis. The majority of lung cancer patients are smokers whose airway epithelia are continuously exposed to inhaled smoke containing carcinogens and irritants and hence undergo chronic injury, inflammation, regeneration. In those lung cancer patients, the accumulation of multiple somatic mutations finally results in the transformation of a small portion of normal lung epithelia to cancer cells, which acquire aberrant over-capacity of cell proliferation, cell growth, and cell migration. Despite of the dominant prevalence of smokers, there is an increasing incidence of lung cancer among non-smokers who have no an apparent history of smoking or exposure to environmental carcinogens. Those patients usually have distinct clinical presentations compared with smoking lung cancer patients, such as early onset of cancer, infrequent genetic mutations, aggressive properties, and a high heterogeneity. Cancer stem cells are likely to be the cell origin of those lung cancers. They are arrested at different stages of differentiation that thus lead to a histological composition of multiple cell types in tumor tissue. Due to the low degree of differentiation and self-renew capacity, cancer stem cells have the potential to escape from chemotherapeutic toxicity and serve as the source of recurrence.

III. **Tumor Suppressor Gene (TSG) inactivation**

1. *TP53*

TP53 plays a role in cell cycle control through regulating the expression of its downstream genes, such as CDKN1A, BAX, BCL2, and CDC2. TP53 is a DNA binding protein. When it binds to binding sites harboring by BAX and CDKN1A, it induces the expression of these genes, resulting in apoptosis

or G1 phase arrest. Thus, TP53 functions as a tumor suppressor gene (TSG). Inactivating mutations of TP53 are one of the most common genetic alterations in diverse human cancers, including ~ 60% of lung cancer [6]. In such cases, G1 arrest cannot be achieved and abnormal cells proceed to S phase. The genetic damage is further propagated in descendant cells, which may lead to cancer.
In NSCLC, G:C to T:A mutations are the most frequent mutation of TP53, unlike in other solid tumors where A:T transitions are more common. And the G → T mutation in TP53 codon 157 is found being lung cancer specific [6]. In addition, TP53 mutation has been associated with tobacco exposure based on the finding that cigarette smoke carcinogens, benzo[a] pyrene diol epoxide (BPDE) adducts, are distributed more often in TP53 mutation hotspots [6].

Although, the mutation frequency of TP53 does not show significant difference between smokers and non-smokers in the other study, the mutation pattern in different cell types is different, such as between SCC and ADC [7]. The different observations might due to different study populations involved in two studies and different codons screened. The former is conducted in the United States, while the later contains only Asian people, and codon 157 was not screened in the second study. According to Gao et al, among Asian people, the mutation pattern instead of transition frequency of TP53 varies between smokers and non-smokers and among different histology types of lung cancer patients. The more common G → A and T → C transitions in the TP53 gene were found in non-smokers, while A to G and A to C transversions are more frequent in SCC than in ADC [7].

It is reported by Ikeda et al that in NSCLC patients the presence of TP53 mutation had an adverse impact on overall survival, at the same time, such patients showed significant benefits from adjuvant chemotherapy than those with TP53 wild type NSCLC [8].

2. *RB*

RB gene was the first described TSG, and originally detected in hereditary retinoblastoma, hence it is named. Aberrant expression of RB gene later was found in other types of tumor, including lung cancer.

RB has multiple functions relating to cell cycle progression, especially

in G1 to S phase transition, proliferation, differentiation, and apoptosis, thus it plays an important role in tumorigenesis. The disruption of the RB-Cyclin(CCN) D1-p16 pathway has been shown to be crucial in the development of many solid tumors, including lung cancer. The mechanism of RB mediated tumor suppression is achieved by controlling the transcription of E2F-responsive genes.

RB has a repressive role in regulating the transcription of E2F1 responsive genes. The binding of RB to the binding site of E2F1 inhibits transcription of downstream target genes. The inhibition of RB can be reversed by phosporylation. Each subset of RB can be phosphorylated by different cyclin-CDK complexes, such as CCND1-CDK4/CDK6 and CCNE10-CDK2, resulting in the release of sequestrated E2F transcription factors [9]. Free E2F complexes drive the transcription of responsive genes including important protein encoding genes required for DNA synthesis, such as CCNA2, TK1, and DFHR. The diminishment of RB function can also be induced by mediated by gene mutation, homozygous/heterozygous deletion, and promoter hypermethylation. [9]

RB function in NSCLC was correlated with p16/CDKN2A, p53/TP53, and MDM2 - a reciprocal expression between RB and MDM2, p53 and p16 was observed in a subset of NSCLC [10, 11]. Simultaneous abnormal expression of RB and p16 indicated that the loss of RB function is essential for cell cycle proceeding initiated from p16 inactivation. In this respect, the tumor possessing an abnormal p16 further acquires RB mutation, consequently releases the inhibition to E2F1 transcription factor. If RB keeps functional, p16 inactivation alone is unlikely to prompt tumor cell growth. On the other hand, the coincidence of aberration of p16-RB and p53-MDM2 in a proportion of NSCLC (43%) suggested a synergistic relationship between two pathways in carcinogenesis. The negative regulation of RB by MDM2 might mediate the interrelation among these proteins [10].

The expression of RB was significantly decreased in MDM2-high tumor tissues compared with those in MDM2-low tissues. In vitro and in vivo studies suggested that MDM2, a ubiquitin ligase, inhibited the function of RB

by a ubiquitin-dependent degradation, thus promoting cell cycle progression from G1 to S phase.

Aberrant or absent expression of RB gene is more common in patients with neuroendocrine tumor, SCLC and LCNEC, up to 90% of patients with this type of tumor presented absent RB expression [10]. On the contrary, only 25% of tumor specimens from ADC and SCC had eliminated RB expression, and the premaligant lesions and TC rarely exhibited abnormal RB expression. Different RB expression pattern is accompanied by the different RB alterations in subtypes of lung cancer. In SCLC and LCNEC, RB alteration might result from point mutations, deletion, or chromosome loss (up to 30%, 90%, and 58% respectively), but not promoter hypermethylation; while in NSCLC LOH and RB mutation is more frequent (up to 75% and 33%).

IV. **Genetic instability**

1. Chr. 3p

The allelic losses and homozygous deletions of 3p are detected in various types of cancers, including uterine cervix, renal, breast and lung cancers. In NSCLC, 96% of patients presented LOH of at least one locus [12].

Allelic loss was the most frequently detected in SCLC (85 ~ 100%), followed by ADC (80 ~95%). It is reported that 86% of current and former smokers displayed LOH in their bronchial mucosa, and 50% of the histologically normal specimens from smokers showed LOH, while never smokers had no allelic loss. One of the most frequently observed deletions is located at 3p14. In the region of 3p14.2, the frequency of LOH is increasing from 0 to 40%, and 100% along with the progression of malignancy in histological changes [12]. Given these investigations, the allelic losses of chromosome 3p are assumed the earliest genetic events in lung carcinogenesis caused by exposure to smoke carcinogens.

The progressive alteration was also observed in other Chr3p regions, such as 3p21.3. Wistuba et al. showed in SCC a significant increasing of 3p21 allelic losses was detected with the progression of carcinogenesis, with LOH frequencies below 20% in normal and reversible lesions, 47 to 83% in dysplasia

and CIS, and 90% in invasive carcinoma [2].

Moreover, a correlation between FRL, Fractional Regional Loss, and lung cancer development was established by Wistuba and his colleagues [2]. Out of 24 microsatellite markers spanning Chr3p region, FAL index was significantly smaller and discrete in histologically normal lung epithelium and mildly abnormal lesions, while in CIS and invasive carcinoma, FAL reached up 0.74 and 0.81. The extent of Chr3p deletion was intermediate in epithelial dysplasia [2].

2. Chr. 9p

Like chromosome 3p, alterations of 9p have been proposed as an early event in NSCLC formation. Allelic loss was detected in approximately 20% of histologically normal tissues from SCC who had a smoking history. And the frequency of LOH at 9p region increases progressively along with the multi-step SCC development [2]. One well-studied TSG, p16/CDKN2A, is mapped to chromosome 9p21. Its protein products function as an inhibitor of cyclin-CDK4 complex and interact with p53-MDM2 pathway via RB, then play a role in cell cycle G1 control through p16-RB and p53-MDM2 networks [10].

The loss of expression of 9p21 was found in over 80% of CIS and invasive SCC. By contrast, less than 20% normal and hyperplasia, around 35% dysplasia showed the loss of 9p21 [2]. In respect of CDKN2A, the loss of this gene was already seen in pre-invasive lesions of patients with CDKN2A negative invasive carcinoma. In patients who had retained CDKN2A expression in invasive carcinoma, no CDKN2A loss was observed in preinvasive lesions [13]. It suggests that the loss of CDKN2A expression is a critical step in early carcinogenesis, and is maintained during tumor progression.

Multiple mechanisms are responsible for the loss of CDKN2A expression, including homozygous deletions and point mutations within the coding region of the CDKN2A gene and aberrant promoter hypermethylation [14]. Methylated CDKN2A could be detected in patient's sputum up to 35 months before the clinical manifestation of SCC. In addition to sputum, approximately 63% of corresponding bronchoalveolar lavage showed concordant methylation of CDKN2A. Considering that non-cancerous patients

rarely present methylation of CDKN2A in serum, sputum, or bronchial lavage, the presence of CDKN2A methylation in sputum or lavage fluid might serve as a highly specific marker for early diagnosis of SCC.

V. **Oncogene activation**

1. *RAS family and downstream effectors*

The RAS small GTPase super family is named because of their low molecular weight (~20-35 kDa) and their intrinsic GTPase activity. This superfamily consists of more than 100 members in humans. They are ubiquitously involved in signal transduction and membrane trafficking. There are 6 families in this superfamily, RAS, RHO, ARF, RAB, RAN and RAD [15]. Oncogenic mutations in RAS genes are detected in nearly 30% of human cancers, with KRAS activating mutation frequently found in NSCLC [16].

The RAS family contains 5 RAS – H-RAS, K-RAS, M-RAS, N-RAS, and R-RAS, 4 RAP – RAP1A, RAP1B, RAP2A, and RAP2B, and 2 RAL – RALA and RALB proteins. These RAS proteins are associated with diverse biological processes, including cell cycle progress, apoptosis regulation, and cell adhesion. Another important RAS family is RHO family, consisting of 7 RHO, including RHOA to E, G and H, 3 RAC, including RAC1 to 3, and CDC 42 proteins. The RHO proteins commonly act as key regulators in relation to cell cycle progression, actin cytoskeleton dynamics, and mobility.

Ras protiens are membrane-bound and cycle between the GDP- and GTP-bound forms. In the quiescent stat, Ras exists in a RAS-GDP form. The sequential events, including the binding of an external ligand such as EGF to its receptor, phosphorylation of intracellular receptor tyrosine residue, the formation of adaptor complex, the interaction of adaptor complex and RAS, induces a change in RAS conformation and dissociation of GDP from RAS-GDP. Subsequently, RAS binds to GTP and become active. The function of RAS-GTP is realized mainly via three downstream signaling pathways, mitogen-activated protein kinase (MAPK) cascades, signal transducer and activator of transcription (STAT), and phosphatidylinositol 3-kinases (PI3Ks).

The activation of the MAPK or PI3K pathway induces the elevated

expression of multiple transcription factors, such as c-JUN, c-MYC, and c-FOS. The binding of these transcription factors to the CCND1 promoter results in increased transcription of genes associated with cell proliferation. On the other hand, RAS downstream MAPK pathway can promote cell cycle progression by decreasing p27 protein level through enhanced degradation. The high level of p27 is able to inhibit cyclin-CDK complexes and lead to cell cycle arrest [16]. RAS is involved in apoptosis, and cancer cell invasion via another downstream effector PI3K. RAS-GTP binding to a subunit of PI3K leads to the activation of RAC or the suppression of c-Myc. The final effect of PI3K pathway is the uncapping of actin filaments at plus-end or decreased expression of several proteins, including pro-apoptotic protein BAD, caspase 9, and forkhead protein, then eventually induces membrane ruffling, high cell motility, and inhibition of apoptosis [16].

2. *RAS mutation in NSCLC*

The activity of RAS is normally transient because of its intrinsic guanine triphosphatease (GTPase) capability which hydrolyzes the bound GTP resulting in the dissociation of RAS-GTP active form. Mutant variants of RAS have defective GTPase activities, and consequently are constitutively activated, resulting in the sustained activation of downstream pathways.

KRAS point mutation was found in 48% of screened NSCLC by Gao and his colleagues. Since smoking is a common feature among lung cancer patients similar as KRAS mutation, therefore an association between smoking exposure and KRAS mutation it is assumed. However, in Gao's study, no difference in KRAS mutation was found between smoker and non-smoker NSCLC patients [7]. Although no significant difference in KRAS mutation frequency observed, a different mutation pattern in non-smoker was detected by the same study, with a higher number of G → A transitions in non-smoking NSCLC patients. Among NSCLC types, a higher number of KRAS mutation was seen in SCC compared to ADC (66% vs 21%), but the difference was not statistically significant. Similarly, among SCC patients an A → T transversion was significantly more frequent in comparison with ADC [7]. These observations indicated that smoking status is not a reason-effect factor of KRAS mutations,

or not all KRAS mutations are associated with tobacco exposure.

The relationship between KRAS mutation and tumor progression was studied using ADC and its corresponding premalignant lesions [17]. A decreasing frequency of KRAS mutations along progression of ADC was observed, 26.7% in AAH, 16.7% in BAC, and 10% in invasive ADC [17]. This result suggested that KRAS mutation must work synergistically with other alterations to induce NSCLC progression otherwise they would remain indolent.

3. *EGFR*

EGFR is located on chromosome 7p12, and is a member of the ERBB family of tyrosine kinase receptor proteins. These family members have intrinsic tyrosine kinase activity, and are structurally similar in the extracelluar domain. The receptor exists as an inactive monomer. Upon binding to the ligands, including growth factors, the receptor undergoes a conformational change, which induces the dimerization and autophosphorylation of intracellular tyrosine domain. Activated EGFR subsequently results in the activation of RAS, which functions as a central distributor of the signal to downstream pathways important in cell survival and proliferation. Stimulated effector pathways include AKT signaling via PI3K, and MAPK signaling via RAF.

The most common mutations in EGFR gene include: in-fram deletion found in exon 19, resulting in four amino acid elimination (E746-A750) in tyrosine kinase domain of the encoding protein; and point muation in exon 21, leading to leucine-arginine transversion (L858R) [18].

These mutants confer ligand-independent activation to EGFR and prolong receptor kinase activity after ligand stimulation, therefore constitutively prompt the activity of downstream effectors. In vitro kinase activity studies, L858R mutant presented a ~20-fold higher catalytic efficiency than that of the wild-type EGFR [19]. Similarly, mouse model studies showed that expression of either exon 19 deletions or L858R mutant lung epithelia leads to formation of tumors analogous to human lung cancers [20].

EGFR mutation is correlated with non-smoker, female, Asian population and BAC subtype [18]. In a screening for EGFR mutation in a NSCLC cohort

(nr.=154), mutations within exon 18 to 21 were examined using PCR assay. A significant difference in EGFR mutation frequency was found between female and male, non-smoker and former/current smoker, ADC with and without BAC component, and ADC with or without solid component [18].

However, at protein level, the overexpression of EGFR had the highest frequency (51%) in SCC, followed by 40% in BAC, compared to 23% and 20% in other ADC and LCC. The high EGFR protein level was correlated to high gene copy numbers in SCC rather than in non-SCC [21]. A similar overexpression of EGFR protein was also seen in BAC, but it was more associated to low gene copy number per cell. In contrast to SCC and BAC, low EGFR protein level and related balanced disomy and trisomy were observed in ADC [21].

Tyrosine kinase inhibitors (TKI), including erlotinib, gefitinib, and cetuximab, can bind to EGFR and then inhibit the autophosphorylation of mutant EGFR, inducing apoptosis of tumor cells habouring EGFR mutations. Recent studies have shown that clinical response to TKIs in NSCLC patients is significantly associated with mutations in EGFR. As reviewed by Riely et al., the response rate to erlotinib or gefitinib for patients with tumors habouring EGFR mutations is around 75%, comparing with 10% for those whose tumors have wildtype EGFR [22](Fig. 4). Nevertheless, EGFR mutations are not equivalent in respect of clinical response to TKI therapy. Highly responsive NSCLC contains E746-A750 deletion in exon 19. For patients with E746-A750 deletions, the time to progression after initial TKI treatment was 12 months, compared with 5 months for patients with L858R; the median survival was 34 months for patients with E746-A750 deletions versus 8 months for those with L858R [23].

Other mechanisms resulting in the activation of the EGFR pathway include gene amplification, amplification of a dinucleotide repeat in the promoter, and enhanced signaling due to heterodimerization with other members of the EGFR family, including HER2/HER3.

VI. **Growth stimulation**

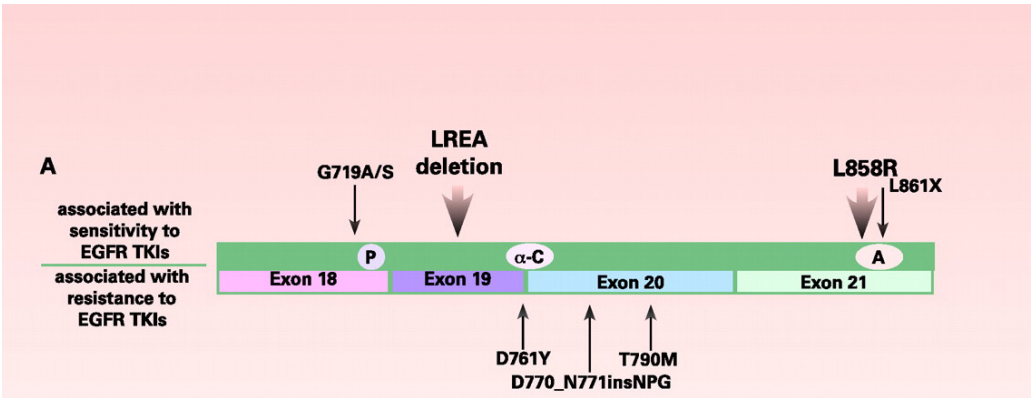


Fig. 4. EGFR mutations and their association with sensitivity to EGFR-TKIs [22]

1. *Apoptosis and cell cycle*

Expansion and progression of premalignant cells toward malignancy depend on the interaction of multiple biological processes, overwhelming tumorigenesis and disabled anti-neoplastic defenses being a common consequence.

The most important capabilities potential tumor cells acquire during tumor development is to escape the physiological pathway of programmed cell death and to overthrow the cell cycle arrest.

Normal cells will undergo apoptosis under some conditions, such as DNA damage, oncogene activiation, and hypoxia. However, tumor cells are protected against apoptosis through diverse molecular pathways which function independently or cooperatively. The mechanism is initiated by the loss or inactivation of TSGs, such as P53, PTEN, and FHIT, or the overexpression of oncogenes, including EGFR and KRAS. The deregulation of these genes triggers PI3K-AKT signaling pathway, leading to a reduced apoptosis via the inhibition of caspases; or MAPK pathway leading to enhanced cell proliferation; or p53 pathway leading to impaired G1 to S phase arrest.

In normal cells, cell cycle arrest in G1 phase is achieved by RB inhibiting expression of E2F responsive genes. This growth inhibition is reversed by the phosphorylation of RB induced by CCND1-CDK4/6. The formation of CCND1-CDK complexes is on the other hand inhibited by members of INK CDKI family, including CDKN2A, CDKN1A, and PSMD9/p27. In tumor cells, CCND1-CDK4/6 is activated by MAPK signaling pathway, or

activated indirectly as the consequence of the down-regulation of CDKN2A or CDKN1A. The MAPK signaling pathway is promoted by the overexpression of oncogenes, such as K-RAS and EGFR. In contrast, the down-regulation of TSGs, e.g. P53, and enhanced PI3K-AKT signaling lead to inactivation of CDKN1A/PSMD9.

2. *Angiogenesis*

To sustain growth and metastasize, tumors need sufficient vasculature to supply increased demand for nutrition and to facilitate spread of tumor cells. VEGF, an angiogenic factor, has a mitogenic effect on vascular endothelial cell. The expression of VEGF is upregulated in response to hypoxia and by activation of RAS. In such cases, diverse growth factors are produced by tumor cells to induce the formation of new blood vessels. Overexpression of VEGF is found in over 80% of NSCLC and premalignant lesions, indicating that VEGF plays a role in the early carcinogenic process [24].

Table 5. Frequently Involved genetic alterations in Human Lung Cancer
Oncogenes
MYC, MYCN, MYCL (deregulated expression) K-RAS, H-RAS, N-RAS (activating mutation) HER-2/neu (deregulated expression) EGFR
Tumor-suppressor genes
3p14 (FHIT deletion) 3p21.3 (region of the GNAI2 gene) 3p24-25 (region of the Von Hippel Lindau gene) 5q (FAP, MCC gene cluster) 9p (interferon gene cluster) 11p15, ~ 11p13 13q14 (retinoblastoma gene, RB1) 17p13 (TP*53* gene)

References

1. Kerr, K.M., *Pulmonary preinvasive neoplasia.* J Clin Pathol, 2001. **54**(4): p. 257-71.
2. Wistuba, II, et al., *Sequential molecular abnormalities are involved in the multistage development of squamous cell lung carcinoma.* Oncogene, 1999. **18**(3): p. 643-50.
3. Wistuba, II and A.F. Gazdar, *Lung cancer preneoplasia.* Annu Rev Pathol, 2006. **1**: p. 331-48.
4. Ten Have-Opbroek, A.A., et al., *The alveolar type II cell is a pluripotential stem cell in the genesis of human adenocarcinomas and squamous cell carcinomas.* Histol Histopathol, 1997. **12**(2): p. 319-36.
5. Kim, C.F., et al., *Identification of bronchioalveolar stem cells in normal lung and lung cancer.* Cell, 2005. **121**(6): p. 823-35.
6. Denissenko, M.F., et al., *Preferential formation of benzo[a]pyrene adducts at lung cancer mutational hotspots in P53.* Science, 1996. **274**(5286): p. 430-2.
7. Gao, H.G., et al., *Distribution of p53 and K-ras mutations in human lung cancer tissues.* Carcinogenesis, 1997. **18**(3): p. 473-8.
8. Ikeda, N., S. Nagase, and T. Ohira, *Individualized adjuvant chemotherapy for surgically resected lung cancer and the roles of biomarkers.* Ann Thorac Cardiovasc Surg, 2009. **15**(3): p. 144-9.
9. Wikman, H. and E. Kettunen, *Regulation of the G1/S phase of the cell cycle and alterations in the RB pathway in human lung cancer.* Expert Rev Anticancer Ther, 2006. **6**(4): p. 515-30.
10. Gorgoulis, V.G., et al., *Alterations of the p16-pRb pathway and the chromosome locus 9p21-22 in non-small-cell lung carcinomas: relationship with p53 and MDM2 protein expression.* Am J Pathol, 1998. **153**(6): p. 1749-65.
11. Leversha, M.A., et al., *Expression of p53, pRB, and p16 in lung tumours: a validation study on tissue microarrays.* J Pathol, 2003. **200**(5): p. 610-9.
12. Wistuba, II, et al., *High resolution chromosome 3p allelotyping of human lung cancer and preneoplastic/preinvasive bronchial epithelium reveals multiple, discontinuous sites of 3p allele loss and three regions of frequent breakpoints.* Cancer Res, 2000. **60**(7): p. 1949-60.
13. Brambilla, E., et al., *Alterations of Rb pathway (Rb-p16INK4-cyclin D1) in preinvasive bronchial lesions.* Clin Cancer Res, 1999. **5**(2): p. 243-50.
14. Gazzeri, S., et al., *Mechanisms of p16INK4A inactivation in non small-cell lung cancers.* Oncogene, 1998. **16**(4): p. 497-504.
15. Riely, G.J., J. Marks, and W. Pao, *KRAS mutations in non-small cell lung cancer.* Proc Am Thorac Soc, 2009. **6**(2): p. 201-5.
16. Adjei, A.A., *Blocking oncogenic Ras signaling for cancer therapy.* J Natl Cancer Inst, 2001. **93**(14): p. 1062-74.
17. Yoshida, Y., et al., *Mutations of the epidermal growth factor receptor gene in atypical adenomatous hyperplasia and bronchioloalveolar carcinoma of the lung.* Lung Cancer, 2005. **50**(1): p. 1-8.
18. Sonobe, M., et al., *Mutations in the epidermal growth factor receptor gene are linked to smoking-independent, lung adenocarcinoma.* Br J Cancer, 2005. **93**(3): p. 355-63.
19. Zhang, X., et al., *An allosteric mechanism for activation of the kinase domain of*

epidermal growth factor receptor. Cell, 2006. **125**(6): p. 1137-49.
20. Ji, H., et al., *The impact of human EGFR kinase domain mutations on lung tumorigenesis and in vivo sensitivity to EGFR-targeted therapies.* Cancer Cell, 2006. **9**(6): p. 485-95.
21. Hirsch, F.R., et al., *Epidermal growth factor receptor in non-small-cell lung carcinomas: correlation between gene copy number and protein expression and impact on prognosis.* J Clin Oncol, 2003. **21**(20): p. 3798-807.
22. Riely, G.J., et al., *Update on epidermal growth factor receptor mutations in non-small cell lung cancer.* Clin Cancer Res, 2006. **12**(24): p. 7232-41.
23. Riely, G.J., et al., *Clinical course of patients with non-small cell lung cancer and epidermal growth factor receptor exon 19 and exon 21 mutations treated with gefitinib or erlotinib.* Clin Cancer Res, 2006. **12**(3 Pt 1): p. 839-44.
24. Lantuejoul, S., et al., *Expression of VEGF, semaphorin SEMA3F, and their common receptors neuropilins NP1 and NP2 in preinvasive bronchial lesions, lung tumours, and cell lines.* J Pathol, 2003. **200**(3): p. 336-47.

Chapter 3

High-throughput assesment of RNA expression by oligonucleotide microarray

Gene expression has been studied for decades by various techniques. Traditional gene-by-gene approaches are far insufficient to understand complex cancer biology. New techniques with capability of profiling global gene expression are widely used in post-genome era.

I. DNA microarray technology

The high throughput microarray assay is widely used to perform genome-wide studies of gene expression. This technique is developed based on the principles utilized by Southern or Northern blotting. Instead of a few genes targeted by the blotting assays, the whole genome can be studied simultaneously on a single chip.

Affymetrix gene expression arrays utilize a standardized biotin-labeling protocol (Fig.1). To generate expression profiles by gene microarray, total RNAs is extracted firstly from samples of diverse resources, such as surgery resections, cell lines, and biopsies. Considering that RNA is highly susceptible to degradation, the quality of RNA has to be checked before it is processed further. The following sample processing includes mRNA reverse transcription, amplification, labeling, and chip hybridization. Taking Affymetrix expression assays as an example, probes on these arrays target for cRNAs. Therefore, the isolated mRNA must be converted to cDNA by reverse transcriptase with a poly-T primer. After the conversion, cDNA is transcripted by using T7-polymerase in the presence of biotin-UTP and biotin-CTP. The biotin-labeled cRNA is fragmented before hybridized to Affymetrix chips. The ready-made arrays are washed and stained using GeneChip fluidics. Hybridization patterns are detected by a laser scanner reading out fluorescence emitted from cRNA hybridized to probes on the microarray. The intensity of emitted fluorescence represents the degree of expression of the mRNA or the targeted genes. The output intensity files are ready for further normalization and analyses.

1. Affymetrix Genechip design

Affymetrix Human gene expression chip is able to study over 30,000 human genes simultaneously. This type of array is designed in such a way that each gene is targeted by one or more probe-sets each comprising a set (~ 20) of 25-

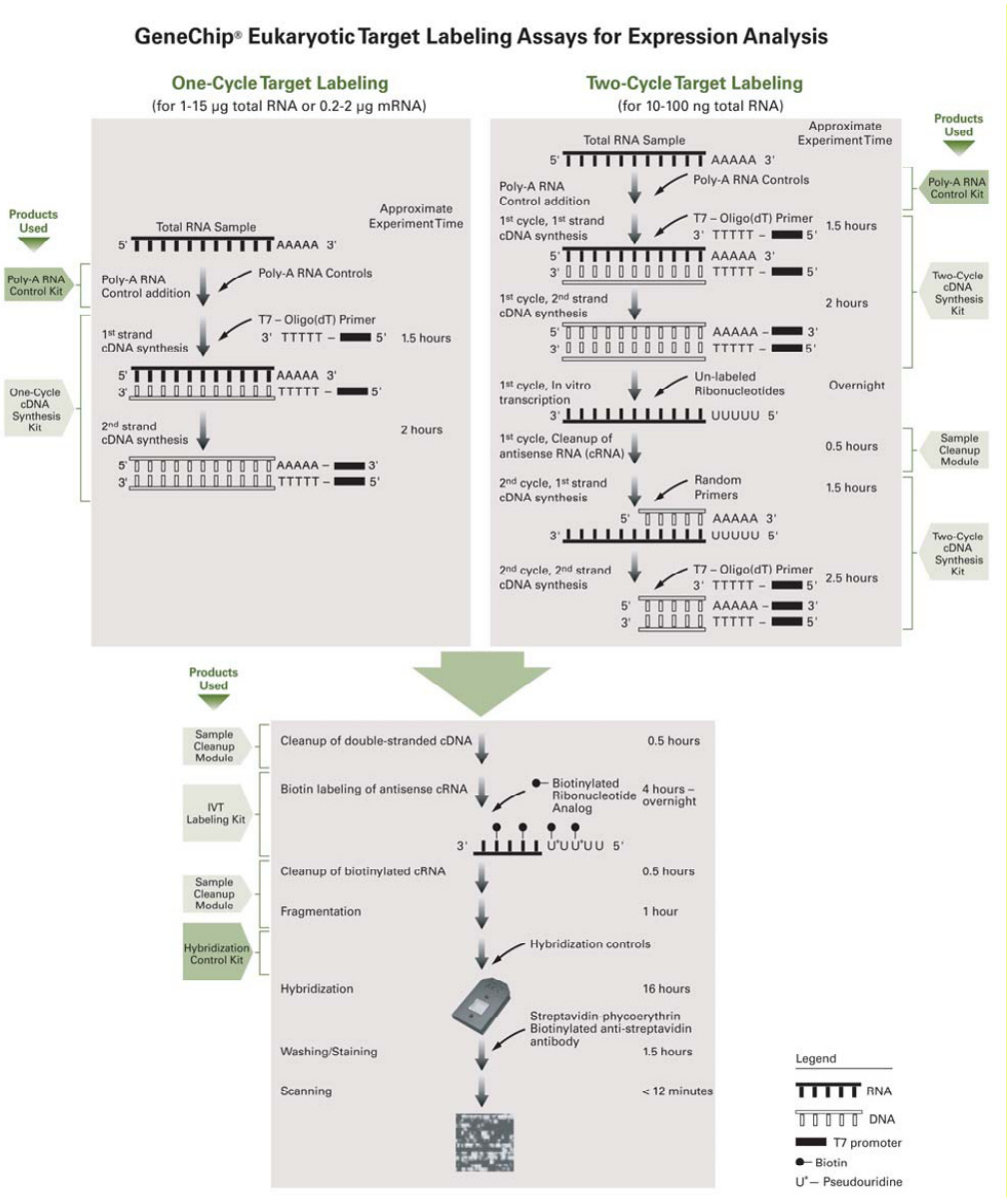


Fig 1. Affymetrix GeneChip labeling assays for eukaryotic expression microarray (from www.affymetrix.com)

base long DNA strands – termed probes, which are complementary sequences to a certain proportion of corresponding gene sequences and have a minimal possibility to match or cross-hybridize to the rest of the human genome.

The principle of how microarray works is the nature chemical attraction

between DNA and RNA molecules, single strand DNA and RNA with complementary sequences can match to each other by base-pairing between A and U, or C and G (Fig. 2). Thus, when sample RNA/cDNA is washed over a microarray, matches between RNA/cDNA and their complementary probes are realized, and the presence of fluorescence emission reflects the expression of probe-targeted genes. With knowledge of predetermined location of each gene probe on the chip, the expression status and expression level of individual genes in a specific sample can be quantified by studying positional intensities of fluorescence.

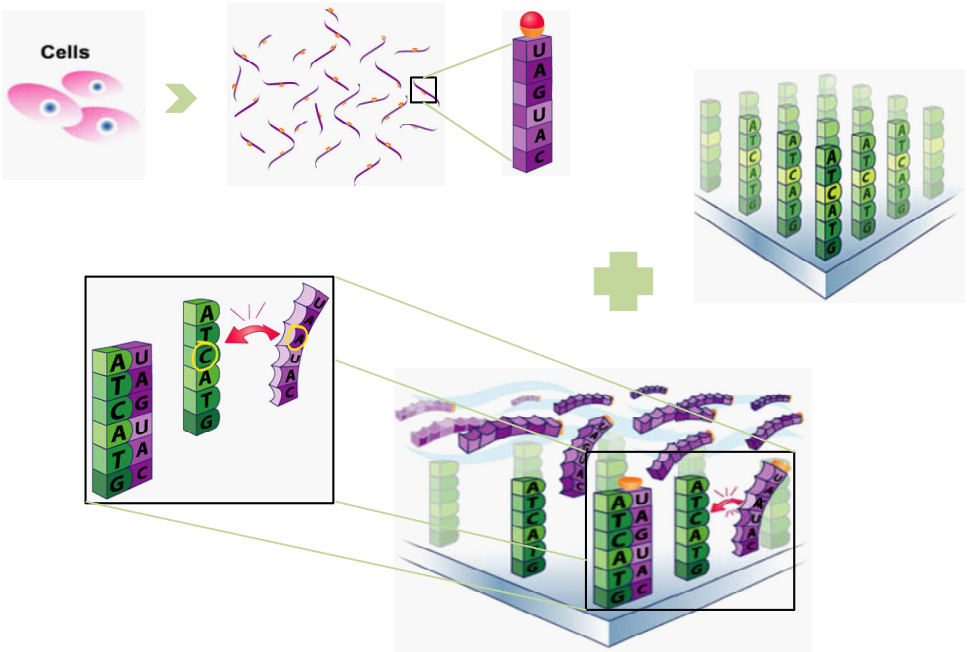


Fig. 2. The design of Affymetrix expression GeneChip (adapted from www.affymetrix.com)

To evaluate and control noise such as non-specific binding of cRNA fragments to probes, Mismatch (MM) probes are introduced which are identical to Perfect Match (PM) described above with an exception for the middle (13th) nucleotide in the probe sequence. The real expression of gene probes can be estimated by using the relative expression of PM to MM.

2. **Microarray data analysis**

Gene expression data can be analyzed by a large variety of approaches.

There is still a lack of a standardized pipeline for processing and analyzing array data derived from different platforms. A few main steps, however, are essential and indispensable in low-level analyses for any type of microarray data in order to diminish noise introduced by and procedures during array preparation and eliminate non-biological variations. Those processes include data preprocessing, normalization, and filtering.

The algorithm developed by Affymetrix, MAS5.0 or GCOS, is widely used for Affymetrix expression data preprocessing. With this algorithm, the expression intensities are measured from both PM and MM. the final intensity values for each probe-set are calculated using the differences between each pair of PM and MM, and then combined among all probes of one probe-set by using one-step Turkey biweight algorithm. Afterward, all intensities on a chip are global scaled to make different experiments comparable. Such processed data are subjected to pre-filtering to remove uninformative and redundant information that comes from genes consistently expressed among the samples. The resulting data set is the basis for subsequent higher level analyses.

The using of mismatch expression might induce spurious estimation of non-specific binding. Therefore, alternative algorithms were developed in which only PM expression intensities were used. An example is RMA normalization standing for Robust Multi-array Averaging. RMA is an integrated algorithm comprising background adjustment, quantile normalization, and expression summarization by median polish. The intensities were background-corrected in such a way that all corrected values must be positive. The RMA algorithm utilized quantile normalization in which the signal value of individual probes was substituted by the average of all probes with the same rank of intensity on each chip. Finally Tukey's median polish algorithm was used to obtain the estimates of expression for normalized probe intensities. GeneChip RMA (GC-RMA) is an improved form of RMA that utilizes the sequence-specific affinities of the probes on chip to attain more accurate gene expression values.

3. Microarray data high level analysis

The most common application of microarray study is to develop gene profiles for relevant characteristics. It is realized by using supervised analysis,

with clinical information or biological phenotypes taken into account. For example, in a study to identify cancer related genes, gene expression from tumor microarrays is compared with that from normal tissues. Genes that show different expression patterns are supposed to be cancer initiation and progression associated, and are candidates for specific cancer markers.

The identified crude signatures can be optimized and trained to build refined models, which then are able to predict the presence of certain properties for new cases, such as the potential recurrence of disease. The power of models, e.g. if the model or signature reflects the biologic features rather than being observed by chance, usually needed to be assessed by applying the model to independent data sets, termed validation sets. The accuracy of prediction then can be evaluated using a percentage of correct classification or prediction. In case no independent validation sets available, "leave-one-out" cross-validation can be applied within the training data set.

Furthermore, the analysis can be performed without taking any external information into account, a strategy termed as unsupervised analysis. The most commonly application is expression profile clustering to find groups of co-regulated genes, or samples sharing certain phenotypes. The strong point of unsupervised analysis is that it allows identification of underlying novel and unexpected patterns, for example new cancer subtypes. Such clustering is based on a distance metric that calculates the similarity between genes or samples. Similarity can be measured either by correlation coefficient or by Euclidean distance.

II. Application of gene expression profiles in oncology

Gene expression profiling has been utilized widely to address diverse biological characteristics, including malignant phenotypes for patient stratification. It has also been shown that this approach has specific power to recapitulate tumor histopathologic features, including identifying new cancer subtypes and to predict clinical outcomes using expression pattern of a set of genes. In addition, gene signatures have been used to predict sensitivity or resistance to certain chemotherapeutic agents; to identify deregulated oncogenic pathways

to instruct the use of targeted therapies.

1. **Histological feature recapitulation**

Gene expression profiles obtained by microarray studies to a great extent reflect tissue histological features. When this approach is applied to lung cancer research, NSCLC, SCLC, and non-cancerous tissues were distinguished from each other by clustering [1-4]. Within the NSCLC cluster, subgroups were evidently recognizable, which recapitulate major subtypes of NSCLC - ADC, SCC and LCC. A distinct type of NSCLC - CAR, a type of benign tumor with neuroendocrine feature, was distinguished from other malignant NSCLC in a separated cluster [1, 4]. It was positioned closely to other neuroendocrine tumors, SCLC and LCNEC by hierarchical clustering, but its expression profiles showed dissimilar patterns [1, 4]. Similarly, the heterogeneity of ADC was recapitulated by expression profiling. ADC shows a high degree of histologic complexity and is further divided into 6 subtypes according to the WHO classification of lung cancer. The microarray data-based clustering of ADC generated multiple subgroups showing diverse expression models at molecular level. BAC, a subtype of ADC with a distinct pathologic feature and relative better prognosis, presented a different profile from other types of ADC. This type of samples was predominant in one of few subgroups and was proved to be associated with good prognosis in subsequent survival analysis [1, 5]. The strong correlation between histologic cell types and gene profile-based clusters implicated that cell type differentiation is the major variance among tumors. It also implicated that gene profiling is a promising tool to discern dominant oncogenic characters in highly heterogenous tumors which are indistinguishable by current routine histopathology.

2. **Clinical outcome prediction**

The most common way to study the association between gene expression profiles and patient survival is to start with predefined classes, e.g. short versus long survival, metastasized versus localized tumor, or relapse versus relapse-free cases. This approach is actually a group comparison and identifies a signature or gene panel that is differentially expressed between two groups [6, 7]. The genes identified in such studies are highly related with the risk of post-

operation recurrence, and play a role in cancer cell motility and invasion.

Another approach that has become more and more accepted does not use predefined groups. By contrast, an unsupervised clustering is performed. It distinguishes aggressive tumors from ones with relatively good prognosis solely on the global gene expression patterns regardless of their histologic types, stages of disease, etc. [6, 8]. The gene signatures that are highly related with clinical outcome identified by this approach might not show extreme difference in expression magnitude individually. The interaction among those genes in a network might confer cancer cells properties favoring aggressive progression and early recurrence.

3. **New cancer subtypes identification**

It has been noticed that routine histopathological classification system is insufficient to address potency of tumor progression and prognosis. Gene expression profiling is an alternative to reveal tumor features that are clinical and pathological outcome related. Currently, the sub-classification based on differentiation grade and histological features of SCC and ADC is poorly correlated with prognosis. Post-surgery survival varied intensively among patients with the same performance status or same tumor histopathology. Two profiling studies using DNA microarry divided SCC into two distinct groups with a significant prognostic difference [3, 9]. Genes involved in cell proliferation, cell cycle transition, MAPKK cascade, and protein modification are differentially expressed in two subclasses. Similarly, the subdivision of ADC based on gene expression patterns was to a great extent correlated to clinical outcomes. ADC subgroups presenting favorite overall survival were defined [1, 2, 10] (Fig. 3).

4. **Identification of deregulated oncogenic pathways**

The development and progression of tumor comprises the accumulation of multiple somatic mutations and epigenetic abnormalities that lead to deregulation of more than one signaling pathways important in controlling cell cycle progression, proliferation, and apoptosis. A strategy to study the role of deregulated signaling pathways in driving tumor formation is to compare different phenotypes in a quiescent state, which is assumed to possess inactive oncogenic pathways, and in an active state in which a specified pathway is

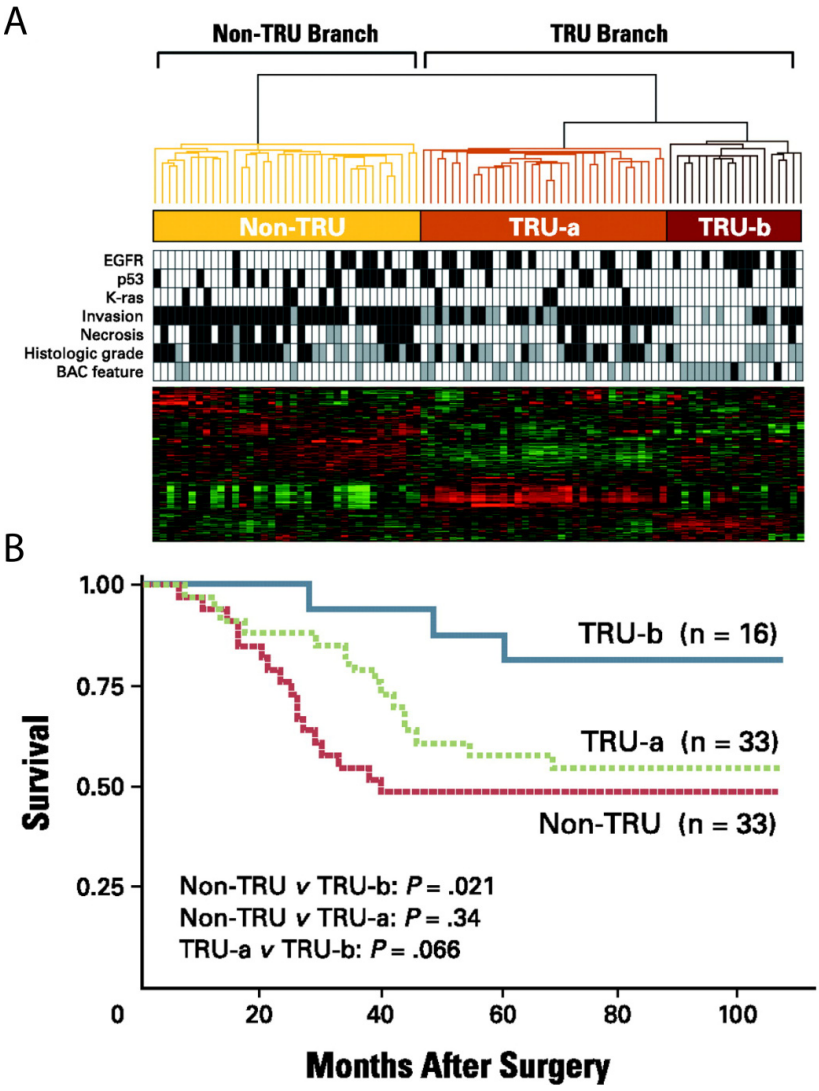


Fig. 3. Unsupervised hierarchical clustering identified three subgroups of adenocarcinoma. These subgroups of ADC differed in genetic changes, tumor behaviors, and patient prognosis (adapted from [2]).

activated by the infection of retrovirus expressing certain gene(s). Bild et al. used gene expression profiling to investigate the correlation between the activated oncogenic pathways and tumor biology and outcome. They firstly identified gene signatures that reflect the activities of several well-known pathways, including MYC, RAS, E2F3, SRC, and β-catenin, central to

oncogenesis in solid cancers. The status of pathways, such as Ras pathway, was clearly correlated with ADC but not other NSCLC subtypes. When signatures for multiple key pathways were combined, they were able to improve the categorization of NSCLCs, breast cancers, and ovarian cancers in terms of prognosis and response to chemotherapeutic agents [11]. These observations indicated that major oncogenic pathways function differentially in the development of different tumor types or subtypes. And the synergic interaction between oncogenic pathways instead of working singularly is the essential determinant for cancer behavior.

5. **Prediction for sensitivity or resistance to clinical treatment**

NSCLC has a high degree histopathological heterogeneity and response dramatically different to therapeutic agents. Taking EGFR tyrosine kinase inhibitors (TKI), an approved target therapeutic agent, as an example, these agents only produced objective responses in 9 to 26% of advanced NSCLC patients in clinical trials, and about half of all treated patients developed relapse within 2 months of initiating therapy. Cell lines that are highly sensitive or highly resistant to gefitinib (EGFR-TKI) were compared to uncover 415 probe-sets/genes associated with the sensitivity to this agent [12]. The prediction for gefitinib sensitivity by this gene signature on independent cohorts appeared to be quite promising: eleven out of twelve cell lines are correctly assigned to be either sensitive or resistant. Another smaller set of genes were identified as independent chemoresistant factors by using microarray expression profiling of trans-bronchial biopsies [13]. These genes include AIF1, CD74, and HLA-DRB1. The expression of these genes was significantly higher in non-responders to cisplatin than that in responders, indicating correlations between their expression levels and tumor response to platinum-based chemotherapy.

Given the intrinsic resistance of NSCLC to various anti-cancer agents, it is essential to understand the primary determinants for chemoresistance so to design more appropriate therapeutic regimens reaching maximal response within the range of tolerable drug toxicity. However, presently it is still a challenge to match the right drug to the right patient for oncologists. Expression profiling provides an opportunity to determine the activity of multiple oncogenic

pathways in individual patients and consequently to predict the sensitivity to certain therapeutic agents that specifically target given pathways. The practical application of this strategy makes it possible to direct individualized therapy, e.g. by combination of medication targeting multiple deregulated pathways, with the aim to improve the response rate in NSCLC chemotherapy. Moreover, the detection of relevant genetic heterogeneity amongst cancers open the door for testing combined chemotherapies in a standardized manner.

References:

1. Bhattacharjee, A., et al., *Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses.* Proc Natl Acad Sci U S A, 2001. **98**(24): p. 13790-5.
2. Takeuchi, T., et al., *Expression profile-defined classification of lung adenocarcinoma shows close relationship with underlying major genetic changes and clinicopathologic behaviors.* J Clin Oncol, 2006. **24**(11): p. 1679-88.
3. Inamura, K., et al., *Two subclasses of lung squamous cell carcinoma with different gene expression profiles and prognosis identified by hierarchical clustering and non-negative matrix factorization.* Oncogene, 2005. **24**(47): p. 7105-13.
4. Jones, M.H., et al., *Two prognostically significant subtypes of high-grade lung neuroendocrine tumours independent of small-cell and large-cell neuroendocrine carcinomas identified by gene expression profiles.* Lancet, 2004. **363**(9411): p. 775-81.
5. Beer, D.G., et al., *Gene-expression profiles predict survival of patients with lung adenocarcinoma.* Nat Med, 2002. **8**(8): p. 816-24.
6. Sun, Z., et al., *Can gene expression profiling predict survival for patients with squamous cell carcinoma of the lung?* Mol Cancer, 2004. **3**(1): p. 35.
7. Potti, A., et al., *A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer.* N Engl J Med, 2006. **355**(6): p. 570-80.
8. Wigle, D.A., et al., *Molecular profiling of non-small cell lung cancer and correlation with disease-free survival.* Cancer Res, 2002. **62**(11): p. 3005-8.
9. Raponi, M., et al., *Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung.* Cancer Res, 2006. **66**(15): p. 7466-72.
10. Garber, M.E., et al., *Diversity of gene expression in adenocarcinoma of the lung.* Proc Natl Acad Sci U S A, 2001. **98**(24): p. 13784-9.
11. Bild, A.H., et al., *Oncogenic pathway signatures in human cancers as a guide to targeted therapies.* Nature, 2006. **439**(7074): p. 353-7.
12. Coldren, C.D., et al., *Baseline gene expression predicts sensitivity to gefitinib in non-small cell lung cancer cell lines.* Mol Cancer Res, 2006. **4**(8): p. 521-8.
13. Oshita, F., et al., *Genomic-wide cDNA microarray screening to correlate gene expression profile with chemoresistance in patients with advanced lung cancer.* J Exp Ther Oncol, 2004. **4**(2): p. 155-60.

Chapter 4

RNA expression-based classification of non-small cell lung carcinomas and survival prediction

Hou J, Aerts J, den Hamer B, van IJcken W, den Bakker M, Riegman P, van der Leest C, van der Spek P, Foekens JA, Hoogsteden HC, Grosveld F, Philipsen S

PloS ONE, submitted, 2009

Abstract

Current clinical therapy of non-small cell lung cancer (NSCLC) depends on histo-pathological classification. This approach poorly predicts clinical outcome for individual patients. Gene expression profiling holds promise to improve clinical stratification, thus paving the way for individualized therapy. A genome-wide gene expression analysis was performed on 91 NSCLC- and 65 adjacent normal lung tissue samples. We defined sets of predictor genes with the expression profiles. The power of predictor genes was evaluated using independent cohorts of 96 NSCLC- and 6 normal lung samples. We identified a 5-gene tumor signature that aggregates the NSCLC and normal lung samples into the expected groups. We also identified a 75-gene histology signature, which classifies the samples in the major subtypes of NSCLC. Correlation analysis identified 17 genes showing the best association with post-surgery survival time. This signature stratified all patients in two risk groups with a significant difference in post-surgery survival time (p=5.6E-6). Compared to previously published prognostic signatures for NSCLC, the 17-gene signature performed well on the Erasmus MC- and validation cohorts. The gene signatures identified are promising tools for histo-pathological classification of NSCLC, and may improve the prediction of clinical outcome.

Introduction

Lung cancer is the most frequent cause of cancer deaths in the North America and Europe. In Europe alone, there were 386,300 new lung cancer cases in 2006, with an estimated 334,800 deaths. This accounts for 13.5% of all cancer deaths [1]. Based on histo-pathological presentation, lung cancer is sub-divided into four major histological subtypes: small cell lung cancer (SCLC), squamous cell carcinoma (SCC), adenocarcinoma (ADC), and large cell carcinoma (LCC). The latter three, collectively referred to as non-small cell lung cancer (NSCLC), account for almost 80% of lung cancers [2]. At present, treatment of NSCLC is based on histo-pathological features and staging. However, pathologically similar tumors with comparable stage show dramatically different response to the same therapy. Common features at the molecular level may be able to predict such outcome discrepancies among patients more reliably. For instance,

the efficacy of epidermal growth factor receptor (EGFR) antagonists has been shown to depend on expression of its target -EGFR- in the tumor [3]. Thus, improved classification of NSCLC is of considerable clinical interest.

Recent advances in microarray technology enable researchers to recapitulate molecular properties of NSCLC at the level of individual genes [4-8]. However, the reproducibility of gene expression signatures to predict high-risk patients is rarely reported. Therefore, it is highly desirable to identify molecular classifiers that can reliably predict specific subgroups of high- and low-risk patients. This would be helpful to select the most appropriate therapy for individual patients.

In this study, we performed gene expression profiling on NSCLC tumors and simultaneously collected normal lung tissue samples in order to determine histo-pathological classifier genes and high-risk index genes.

Materials and Methods

A detailed description is provided in Supplementary Methods.

Patient enrolment

Ninety-one NSCLC patients treated at Erasmus MC were included in this study. Patient and tumor characteristics are listed in Table 1. All studies were approved by the local medical ethical committee. We used two independent validation sets: 6 normal lung tissues from GSE3526, and NSCLC samples from the Duke University cohort [12].

Pathological analysis

Tumor samples were typed by two independent routine pathological reviews, according to WHO guidelines [17].

RNA Isolation and gene expression profiling

Dissected tumors and adjacent normal tissue were snap-frozen in liquid nitrogen within two hours after surgical resection, and stored at –80 °C until RNA extraction. 5 μg of total RNA was processed for analysis on Affymetrix U133 plus 2.0 arrays using standard protocols.

Bioinformatics analyses

Table 1. Characteristics of patients and samples

		Training set			Validation set		
Heathy tissue (n)		36			29		
Tumor (n)		44			47		
Mean age (years)		62.3±10.81			63.5±10.73		
Sex-%	Female	27			34		
	Male	73			66		
Race-%	Caucasian	90			89		
	other	5			3		
	unknown	5			8		
Smoking history-%							
	None	-			-		
	≤ 30 yr	20			24		
	31-49 yr	20			18		
	≥ 50 yr	18			18		
	unknown	41			39		
Tumor type (n)	Path. review	1st	2nd	consistent	1st	2nd	consistent
	ADC	19	14	14	13	10	8
	SCC	16	8	8	11	8	8
	LCC	7	13	6	6	11	3
	other	2	9	1	8	9	1
	unknown	0	0		9	9	
Stage-%	Path. review	1st			1st		
	IA	18			16		
	IB	45			42		
	IIA	2			-		
	IIB	30			21		
	IIIA	2			16		
	IIIB	-			-		
	IV	2			5		
Status-%	Alive	34			29		
	Deceased	61			63		
	unknown	5			8		
Cause of death-%	Lung cancer	27			34		
	other	18			18		
	unknown	55			47		

Table 1. Clinical characteristics of NSCLC cohort.

Standard QC methods were used to control the overall quality of arrays. The final intensity value of probe sets was summarized as the deviation to the geometric mean of that probe set among all arrays. Uninformative probe sets were eliminated and the remaining probe sets were used for subsequent analyses.

Class comparison

Two-group comparisons were performed by Significance Analysis of Microarrays [18]. This supervised analysis correlates gene expression with a

clinical variable based on a score calculated using the change in expression and the standard deviation across all samples.

Class prediction

All primary signatures were optimized to identify subgroups of genes that maintain the capacity in distinguishing different groups maximally [9]. The performance of optimized signatures was validated by "leave-one-out" cross validation within the training set firstly, then with the validation set [10]. Hierarchical clustering was performed using the Spotfire Decision Site.

Survival analysis

We used a step-wise approach based on gene expression profiles to classify NSCLC with respect to prognostic outcome. Firstly, the Wald test in the Cox proportional hazards model was used to identify prognostic genes most likely associated with overall survival [11]. Candidate genes were selected based on p-values (< 0.001) computed from 1000 random permutations. The resulting candidate survival genes were subjected to a supervised analysis [12], which comprises computation of principal components, Cox proportional hazards regression analysis using these principal components, and finally prognostic predictor calculation by fitting the predictive prognosis model derived from the Cox regression. The predictive value of the prognosis model was evaluated by "leave-one-out" cross-validation [12, 13]. The prognostic value of the predictor relative to clinical variables, such as age, tumor cell content (%), tumor size, pack years, Forced Expiratory Volume 1, gender, histology, and tumor grade was tested by the Wald test (Supplementary Table 7). The correlation between the survival signature and clinical parameters is summarized in Supplementary Table 6.

Other NSCLC prognostic classifiers

The 20- and 6-probe set predictors were developed by Lee et al [14]. Additional survival related signatures include one derived from Affymetrix U133A chips [15], one from Affymetrix HuGeneFL chips [16, 17], two from other types of oligonucleotide arrays [18, 19], and one from RT-PCR assays [20] (Supplementary Table 8).

Results

Study design

Tumors (n=91) and unaffected lung tissue samples (n=65) were collected from NSCLC patients undergoing lung resection at Erasmus MC. Tissue specimens were snap frozen in liquid nitrogen and stored at –80 °C until further processing. Clinical parameters of the patients are summarized in Table 1. Paraffin sections of the tumors were scored by routine pathology and an independent pathologist (MdB). We isolated RNA from 25 μm cryostat sections of the snap-frozen specimens and used this for labelling and hybridisation to Affymetrix U133 2.0 plus arrays. Tumor cell content was determined from 10 μm sections taken at the start and end of cryostat cutting. The samples were divided into training and validation sets (Table 1; see Supplementary Materials for criteria).

Signature genes distinguish NSCLC from normal lung tissue

By unsupervised Pearson's correlation analysis, tumor samples were clearly separated from healthy lung samples (Fig. 1). We therefore first sought to derive a minimized signature gene set that could distinguish tumors from healthy lung tissue. To this end, we compared gene expression profiles from 44 tumors with that from 36 healthy lung tissues. By supervised analysis, we identified 187 genes that were differentially expressed in the NSCLC samples (Fig. 2A and Supplementary Table 1). Using Prediction Analysis of Microarrays we found that a 5-gene signature distinguished healthy tissue from NSCLC with an accuracy of 98%, (Fig. 2B and Supplementary Table 2). Two tumor and three non-cancerous lung samples were incorrectly classified by the 5-gene signature. Of these, one presented with an uncertain histological diagnosis, and two were from patients with multiple primary tumors. Weconclude that the 5-gene signature accurately distinguishes NSCLC from healthy lung tissue, regardless of NSCLC subtype.

NSCLC are sub-classified by histology signature genes

As NSCLC are tumors with a high degree of heterogeneity, genes characterizing histological features were identified using strictly selected tumor samples. Firstly, the histological diagnosis had to be consistent between the two

independent pathology reviews, resulting in forty-nine cases left. Secondly, the samples should not display apparent tumor cell heterogeneity. Thirdly, the content of cancer cells should be above 60%. This super-training set consisted twenty-three cases. We compared the gene expression profiles of each NSCLC subtype to those of the remaining samples, and identified 518 genes representing the three major subtypes of NSCLC: ADC, SCC, and LCC (Supplementary Table 3). Using "leave-one-out" cross validation, we found that the percentage of correct classification by Prediction Analysis of Microarrays was 96% (22 out of 23) in the training samples (Fig. 3A). When this signature was applied to classify the validation samples, we found that the three carcinoid (CAR) samples, which were not involved in deriving the signature, and one LCC sample were separated from the other tumors, thus representing a unique group (Fig. 3B). We note that the LCC sample in this group was classified as CAR by the second pathology review. The optimized signature gene set consisted of 75 genes (Supplementary Table 4). This optimized signature classified the training samples with 100% accuracy (Fig. 3A). This signature was applied to predict the histology subtype of the samples with conflicting pathology diagnoses (n=18). Of these 18 samples, one had an ambiguous diagnosis due to unsatisfactory histology, and three had a tumor cell content of less than 20%. We note that over 60% (n=11) of these 18 samples presented with apparent tumor cell type heterogeneity. With three exceptions, the ambiguously classified LCCs (n=11) were classified as ADC or SCC, and this was consistent with the primary diagnosis (Fig. 4A). We conclude that the 75-gene histology signature may aid in assigning the correct histological classification in ambiguous NSCLC cases.

Survival risk prediction

Starting with the 11,515 probe sets remaining from the data filtering process, we used the Wald test from the Cox proportional hazards model to identify the genes that were best correlated with survival time. The principal components computed from the expression of these genes were subjected to Cox proportional hazard regression analysis, and built up a model for predicting a prognostic probability for each NSCLC case. The predictive value of the prognosis

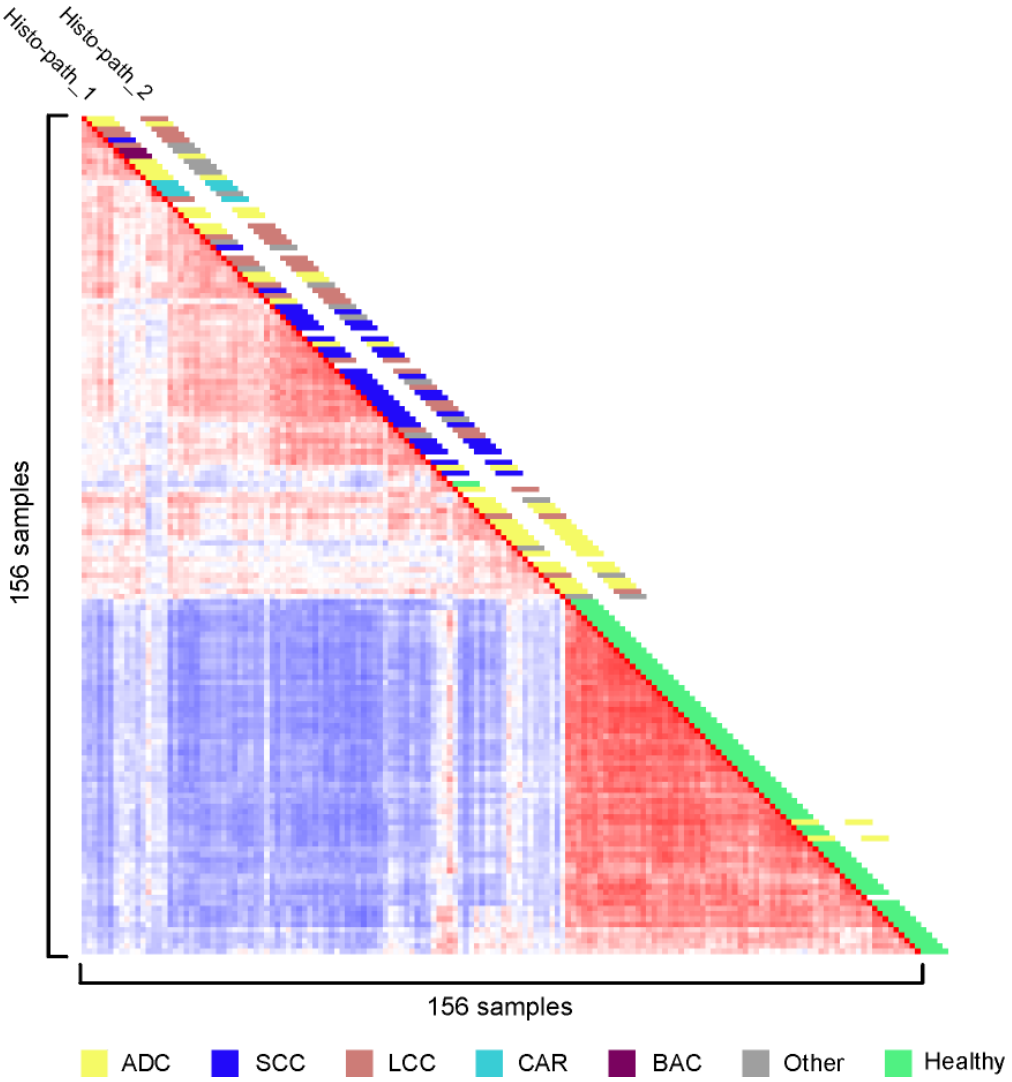


Figure 1. Correlation view of 156 samples from patients with NSCLC.
Pairwise correlations between any two samples are displayed, based on 4791 informative probe sets. The colors of the cells represent Pearson's correlation coefficient values, with deeper red indicating higher positive and deeper blue lower negative correlations. The red diagonal line displays the self-to-self comparison of each sample. Histological classification of the samples is depicted along the diagonal; the key to the color code is shown at the bottom. Histo-path_1 & Histo-path_2: initial and second histo-pathological review.

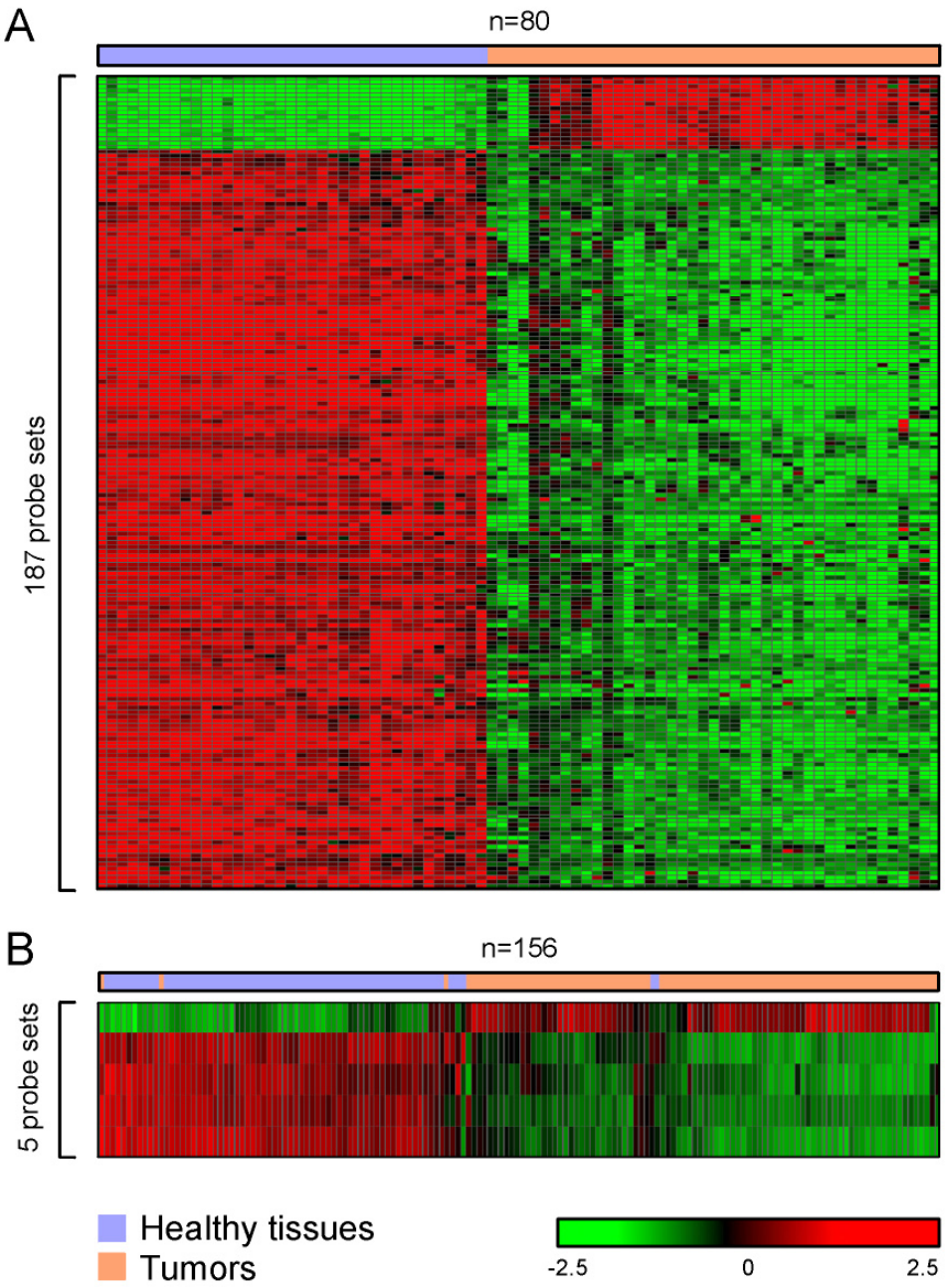


Figure 2. Hierarchical clustering distinguishes tumors from healthy lung tissue. A. Two-dimensional hierarchical clustering of 80 training samples, including tumors and healthy lung samples, was performed with 187 genes. The relative expression to the overall mean for each probe set (rows) in each sample (columns) is indicated by a color code. B. Hierarchical clustering of 156 tissue samples with 5 genes yields 2 groups, tumor and normal lung

model was evaluated by "leave-one-out" cross-validation. This resulted in an optimized model consisting of 17 genes. A risk percentile cut-off of 60% was used to define two risk groups, which were distinguished at significance p-value = 5.6E-6 by log-rank test. A Kaplan-Meier curve of overall survival from these two risk groups is shown (Fig. 5A).

The association between the prognosis profile and clinical parameters was studied. The prognosis profile was significantly associated with age ($p<0.023$), pack years ($p<0.014$), gender ($p<0.012$) and Forced Expiratory Volume 1 ($p<0.009$), but not with tumor stage, cell content, histology and size (Supplementary Table 6). We performed multivariate proportional hazard regression analysis to evaluate the predictive value of the prognostic predictor for patient outcome in comparison with other clinical parameters. No evidence of relation was found between relative hazard ratio and age, gender, pack years, tumor cell content, Forced Expiratory Volume 1, tumor histology and tumor size. Supplementary Table 7 shows the Wald statistics and significance for each variable tested. Tumor stage and the 17-gene prognostic predictor were significantly related to the hazard of death. However, the prognostic predictor presented the highest importance: 21.68 compared to 3.80 from tumor stage. Moreover, the relative hazard ratio predicted by the prognostic predictor was 2.47 (95% confidence interval, 1.69 to 3.60, $p<1.5E-06$), the highest among all tested risks (Table 2).

	HAZARD RATIO	(95% confidence interval)	Change in -2 log likelihood	Significance
Age	1.03	(0.99 - 1.07)	10.35	0.001293
Tumor cell %	1.01	(0.99 - 1.03)	2.16	0.141500
Stage	1.32	(1.00 - 1.74)	3.90	0.048425
Gender	1.00	(0.44 - 2.27)	2.78	0.095444
Smoking years	1.00	(0.97 - 1.04)	1.13	0.286797
Forced Expiratory Volume 1	1.01	(0.99 - 1.03)	0.51	0.476836
Tumor size	1.00	(0.98 - 1.03)	0.00	0.979352
Histology	0.91	(0.81 - 1.02)	3.49	0.061814
Prognostic predictor	2.47	(1.69 - 3.6)	19.55	0.000010

Table 2. Multivariable proportional hazard analysis of the risk of death

Similarly, inclusion of the prognostic predictor into the predictive model improved model performance to 19.5, in terms of -2 log likelihood, with a p-value of 9.8E-06, compared to 24.3 and p-value 2.0E-03 without it. Thus,

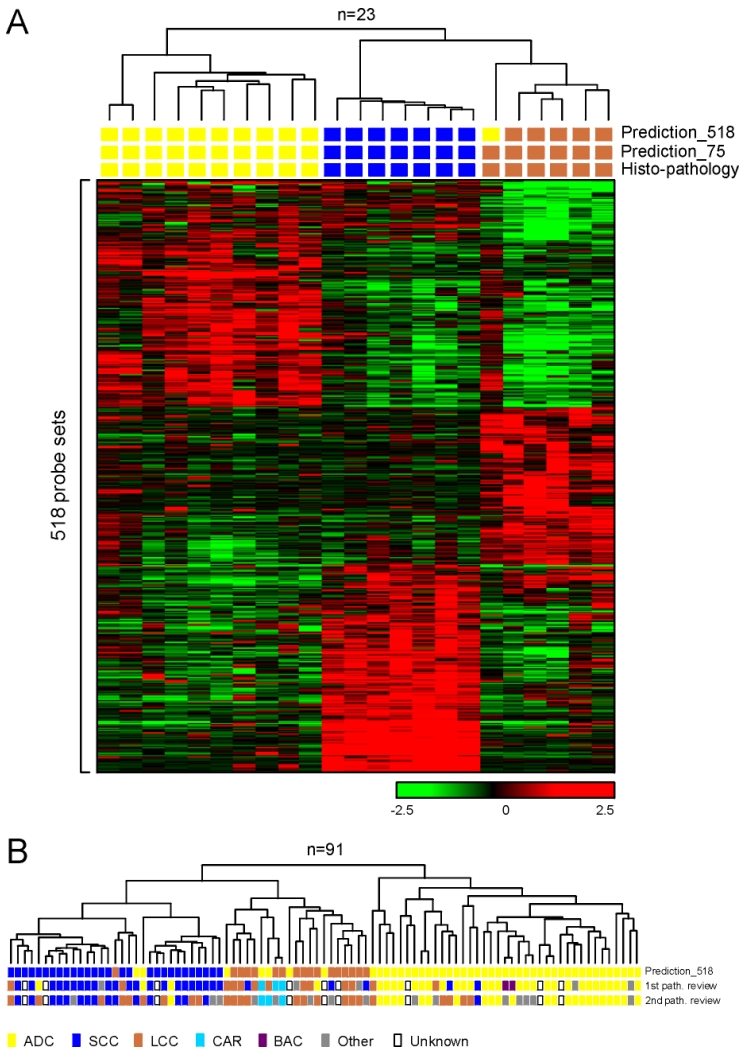


Figure 3. Clustering analysis of NSCLC tumors with the 518 probe set histology signature

A. agglomerative hierarchical clustering of 23 NSCLC samples using the 518 probe set histology signature. The relative expression to the overall mean for each probe set (rows) in each sample (columns) is indicated by a color code. Correlation between the samples is depicted by the dendrogram. Histo-pathological diagnosis and predictions of histology subtype by Prediction Analysis of Microarrays, using the 518 and 75 probe set signatures, are shown by colored blocks. **B.** correlation dendrogram generated by agglomerative hierarchical clustering of all 91 Erasmus MC NSCLC samples using the 518 probe set signature. Histo-pathological diagnosis of the initial and second review, and prediction of histology subtype by Prediction Analysis of Microarrays using the 518 probe set signature, are shown by colored blocks.

multivariate proportional hazard analysis indicates that the 17-gene signature is the strongest survival predictor.

Validation of gene signatures

We studied the expression patterns of all signatures in two independent sets of microarray data collected in the United States, consisting of the NSCLC cohort from Duke University (n = 96) [21], and 6 normal lung specimens (GSE3526). These were chosen because 1) they were also analyzed on the Affymetrix U133 plus 2.0 arrays, and 2) the original .CEL files were available (i.e. raw rather than pre-normalized data). The 5-gene tumor signature performed on the validation set with an accuracy of 97%: 93 out of 96 NSCLC were correctly classified as 'tumor' and all normal lung specimens were correctly classified as 'healthy'. Since there were no LCC or other types of NSCLC in the Duke University data set, we only used the ADC and SCC signature genes for histological classification. For 84% of Duke University samples, the prediction by the 68-gene ADC/SCC signature was consistent with the reported histology diagnosis. When the LCC signature was included in the prediction analysis, this percentage decreased to 83%: 2 samples were classified as LCC (Fig. 4B). Follow-up data were available for 89 of 96 patients in the Duke University cohort, and we calculated the prognostic predictor for these patients using the 17-gene survival signature and the predictive model. The difference in the hazard of death between the patient groups with a predicted good prognosis and the group with a poor prognosis was 2.44-fold, with a significance of p-value = 1.9E-03 by log-rank test. A Kaplan-Meier curve of overall survival is shown in Fig. 5B. If the Erasmus MC patient cohort is combined with the Duke University cohort, the p-value reduces to 2.6E-7.

Comparison with other prognostic gene expression signatures

A number of gene expression profiling-derived prognostic predictors have been previously reported for NSCLC [14-20]. These signatures were derived from a wide variety of technological approaches (Supplementary Table 8). We assessed the performance of 14 signatures from 6 different publications on the Erasmus MC and Duke University data sets (Supplementary Methods and Supplementary Table 8). For each publication, the results obtained with the

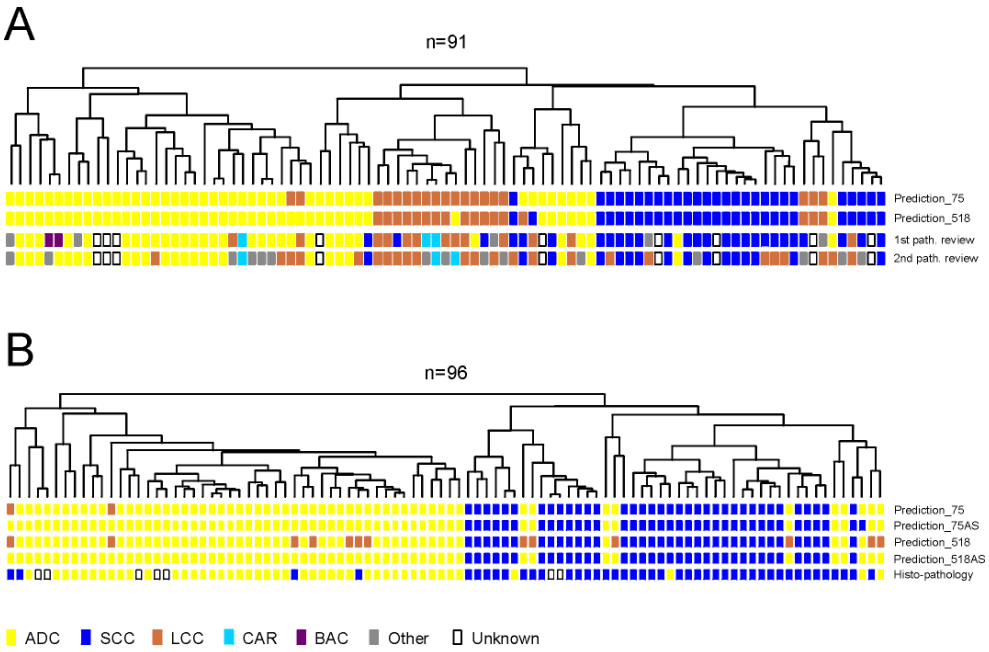


Figure 4. Prediction of histology subtype of Erasmus MC and Duke University NSCLC samples

A. correlation dendrogram generated by agglomerative hierarchical clustering of all 91 Erasmus MC NSCLC samples using the 75 probe set histology signature. Histo-pathological diagnosis of the initial and second review, and prediction of histology subtype by Prediction Analysis of Microarrays using the 75- and 518 probe set histology signatures, are shown by colored blocks. **B.** correlation dendrogram generated by agglomerative hierarchical clustering of all 96 Duke University NSCLC samples using the 75 probe set histology signature. The reported histo-pathological diagnosis, and prediction of histology subtype by Prediction Analysis of Microarrays using the 75- and 518 probe set histology signatures, are shown by colored blocks. 75AS and 518AS: prediction without the LCC genes in the histology signatures, using 68 and 329 genes respectively (see Supplementary Tables 3 and 4).

signature yielding the best stratification are displayed in Kaplan-Meier curves (Supplementary Fig. 1). Performance of the 6-gene signature of Boutros et al [20] was reasonable on the Duke University cohort (p-value 0.016) but not on the Erasmus MC cohort (p-value 0.69). The 41-gene signature of Shedden et al. was developed for ADC samples [15]. It performed unsatisfactory on the complete Erasmus MC and Duke University cohorts (p-values 0.113 and 0.158 respectively). However, if the analysis was limited to samples classified as ADC by our histology signature, this was the only prognostic signature that

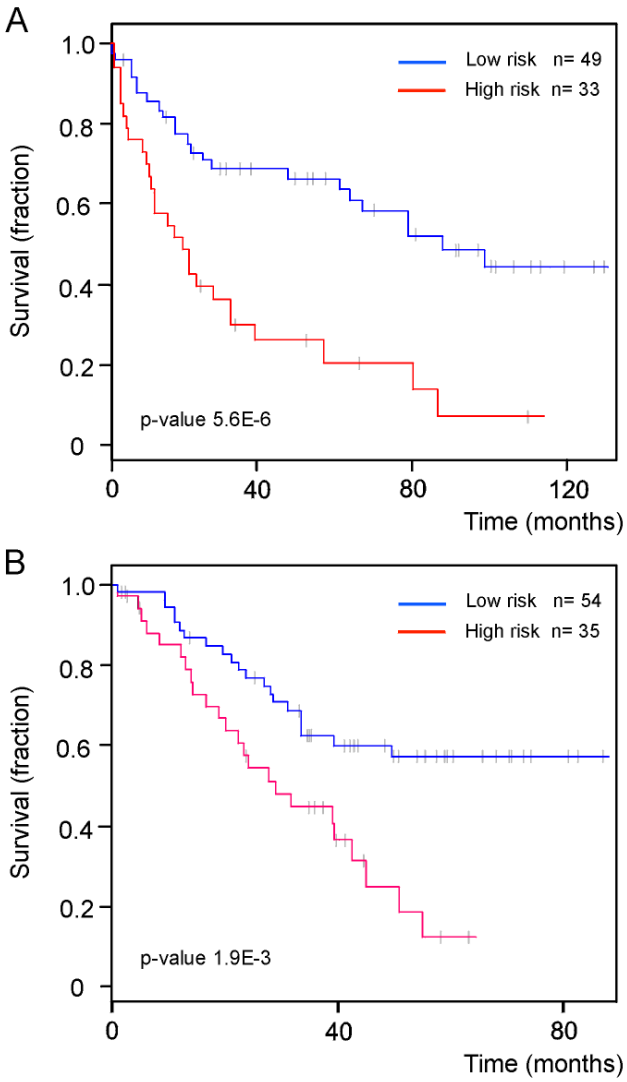


Figure 5. A 17 probe set signature predicts patient survival time

Kaplan-Meier curves for **A.** 82 Erasmus MC NSCLC patients and **B.** 89 Duke University NSCLC patients fitted by their risk assignments based on the 17 probe set survival signature. The high- and low-risk groups differ significantly, indicated by the p-values. Grey bars indicate patients at last follow-up, still alive.

performed well on both cohorts (Erasmus MC p-value 0.016, Duke University p-value 0.019).

Discussion

Here, we defined a set of molecular classifiers for NSCLC. These classifiers

were developed with the Erasmus MC cohort of NSCLC patients, and validated using the independent Duke University cohort. The tumor signature gene set can be used to distinguish NSCLC from unaffected lung tissue. The histology signature gene set may aid in the histo-pathological classification of NSCLC. In addition, we identified a survival signature gene set that predicts overall patient survival.

Potential for improved NSCLC classification

The histological diagnosis of LCC is based on exclusion of the other types of NSCLC. As a result, this subtype of NSCLC is highly heterogeneous in histopathology and clinical presentation. LCC accounts for about 16% of NSCLC cases. By applying special stains and electron microscopy it has been shown that many cases of LCC are poorly differentiated ADC or SCC [22]. The difficulty in distinguishing LCC from other NSCLC by routine histopathology results in considerable variation in the classification of NSCLC cases. In contrast, molecularly defined NSCLC subtypes display distinct gene expression profiles. For instance, a number of well-known SCC markers, such as TP63, PERP, keratins, and SERPINB, were uniformly expressed among a subset of the LCC samples, suggesting that these were actually SCC. In addition, expression profiling revealed that some of the tumors diagnosed as SCC display neuroendocrine characteristics, indicating that these were neuroendocrine tumors. Thus, the molecular signatures reveal specific features of the tumors. This could be used to improve the classification of NSCLC tumors, especially in histologically heterogeneous tumors where the signatures would identify the most characteristic molecular features of the samples.

A 17-gene signature set predicts survival

We have identified a small set of genes that predicts survival time independent of histo-pathological tumor type. Multivariate proportional hazard analysis that included age, pack years, gender, Forced Expiratory Volume 1, tumor stage, tumor cell content, tumor histology, and tumor size indicates that the 17-gene signature set is the strongest predictor of the likelihood of death. Importantly, the performance of this molecular predictor was similar in an independent NSCLC patient cohort, indicating its reproducibility and potential for practical

application in the clinical setting.

Divergence of prognostic gene expression signatures

Potti et al. [21] developed a metagene model to predict the risk of recurrence for individual patients. The model was predictive for the major types of NSCLC – ADC and SCC, and performed reasonably satisfactory in two independent patient cohorts. Confounding components of the metagene models contain over 100 genes. These attributes complicate the direct comparison of the metagenes to survival signatures derived from other studies. As such, the genes in the metagene model have no predictive power for survival analysis (data not shown).

It has been noted before that there is very little, if any, overlap between the reported prognostic signatures for NSCLC [19, 23]. Remarkably, there is not a single gene shared by the 7 signatures tested here (the 6 best performing previously reported signatures and our 17-gene signature). This has been attributed to the notion that the space from which such minimized signatures can be derived is large [19, 20] and hence there are many different possible outcomes depending on the particular dataset and bioinformatics approaches taken. For instance, although outcome signatures make predictions beyond histological subtype, it is still possible that genes in the signature are histology-related. When these signatures are applied to other datasets with different tumor composition, they do not necessarily reflect clinical risk. The 41-gene prognostic signature of Shedden et al. [15] was developed with ADC samples. We found that stratification of the Erasmus MC and Duke University cohorts by this signature is histology-dependent, since it only performs satisfactorily on the ADC samples. For this analysis, we assigned tumor types in the Erasmus MC and Duke University cohorts with our histology signature. Thus, a scenario emerges where application of a histology signature is followed by analysis with a tumor type-specific prognostic classifier. Clearly, it is important to test whether prognostic classifiers of NSCLC are operative beyond histological criteria.

Alternatively, prognostic classifiers transcending tumor histology would be more straightforward to use. To develop these, different tumor types

and subtypes should be included in the experimental set-up. Our dataset covers a relatively broad spectrum of NSCLC, and we have validated the signatures using independent samples profiled on the identical platform [21]. The lack of availability of raw microarray data (.CEL files) precludes validation of our signatures using more independent NSCLC cohorts; the complex issue of cross-platform meta-analysis [16, 24] is beyond the scope of this paper. Nonetheless, our survival signature performed well compared to those previously reported [14-16, 18-20] when tested on the Erasmus MC and Duke University cohorts. We note that although the Duke University samples are clearly separated from the Erasmus MC samples in unsupervised analysis (Supplementary Fig. 2) our signatures still perform well on the Duke University data (e.g. Figs. 4B and 5B), indicating that they are robust.

In conclusion, the sets of molecular markers identified in this report reveal histo-pathological attributes of NSCLC. These gene signatures might provide clinically relevant information for NSCLC, transcending traditional histological classification and patient outcome prediction.

Acknowledgements

This work was supported by the Netherlands Genomics Initiative.

References

1. Ferlay J, Autier P, Boniol M, Heanue M, Colombet M, Boyle P. Estimates of the cancer incidence and mortality in Europe in 2006. *Ann Oncol* 2007: 18(3): 581-592.
2. Pretreatment evaluation of non-small-cell lung cancer. The American Thoracic Society and The European Respiratory Society. *Am J Respir Crit Care Med* 1997: 156(1): 320-332.
3. Sequist LV, Bell DW, Lynch TJ, Haber DA. Molecular predictors of response to epidermal growth factor receptor antagonists in non-small-cell lung cancer. *J Clin Oncol* 2007: 25(5): 587-595.
4. Fujii T, Dracheva T, Player A, Chacko S, Clifford R, Strausberg RL, Buetow K, Azumi N, Travis WD, Jen J. A preliminary transcriptome map of non-small cell lung cancer. *Cancer Res* 2002: 62(12): 3340-3346.

5. Kikuchi T, Daigo Y, Katagiri T, Tsunoda T, Okada K, Kakiuchi S, Zembutsu H, Furukawa Y, Kawamura M, Kobayashi K, Imai K, Nakamura Y. Expression profiles of non-small cell lung cancers on cDNA microarrays: identification of genes for prediction of lymph-node metastasis and sensitivity to anti-cancer drugs. *Oncogene* 2003: 22(14): 2192-2205.
6. Yao R, Wang Y, Lubet RA, You M. Differentially expressed genes associated with mouse lung tumor progression. *Oncogene* 2002: 21(37): 5814-5821.
7. Jones MH, Virtanen C, Honjoh D, Miyoshi T, Satoh Y, Okumura S, Nakagawa K, Nomura H, Ishikawa Y. Two prognostically significant subtypes of high-grade lung neuroendocrine tumours independent of small-cell and large-cell neuroendocrine carcinomas identified by gene expression profiles. *Lancet* 2004: 363(9411): 775-781.
8. Kobayashi K, Nishioka M, Kohno T, Nakamoto M, Maeshima A, Aoyagi K, Sasaki H, Takenoshita S, Sugimura H, Yokota J. Identification of genes whose expression is upregulated in lung adenocarcinoma cells in comparison with type II alveolar cells and bronchiolar epithelial cells in vivo. *Oncogene* 2004: 23(17): 3089-3096.
9. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America* 2002: 99(10): 6567-6572.
10. Golub T, Slonim D, Tamayo P, Huard C, Gaasenbeek M, Mesirov J, Coller H, Loh M, Downing J, Caligiuri M, Bloomfield C, Lander E. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science (New York, NY* 1999: 286: 531-536.
11. Cox DR. Regression models and life-tables. *J R Stat Soc* 1972: 34: 187-220.
12. Bair E, Tibshirani R. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol* 2004: 2(4): E108.
13. Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute* 2003: 95(1): 14-18.
14. Lee ES, Son DS, Kim SH, Lee J, Jo J, Han J, Kim H, Lee HJ, Choi HY, Jung Y, Park M, Lim YS, Kim K, Shim Y, Kim BC, Lee K, Huh N, Ko C, Park K, Lee JW, Choi YS, Kim J. Prediction of recurrence-free survival in postoperative non-small cell lung cancer patients by using an integrated model of clinical information and gene expression. *Clin Cancer Res* 2008: 14(22): 7397-7404.
15. Shedden K, Taylor JM, Enkemann SA, Tsao MS, Yeatman TJ, Gerald WL, Eschrich S, Jurisica I, Giordano TJ, Misek DE, Chang AC, Zhu CQ, Strumpf D, Hanash S, Shepherd FA, Ding K, Seymour L, Naoki K, Pennell N, Weir B, Verhaak R, Ladd-Acosta C, Golub T, Gruidl M, Sharma A, Szoke J, Zakowski M, Rusch V, Kris M, Viale A, Motoi N, Travis W, Conley B, Seshan VE, Meyerson M, Kuick R, Dobbin KK, Lively T, Jacobson JW, Beer DG. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat Med* 2008: 14(8): 822-827.
16. Guo NL, Wan Y-W, Tosun K, Lin H, Msiska Z, Flynn DC, Remick SC, Vallyathan V, Dowlati A, Shi X, Castranova V, Beer DG, Qian Y. Confirmation of Gene Expression-Based Prediction of Survival in Non-Small Cell Lung Cancer. *Clin Cancer Res* 2008: 14(24): 8213-8220.
17. Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizyness ML, Kuick R, Hayasaka S, Taylor JM, Iannettoni MD, Orringer MB, Hanash S. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med* 2002: 8(8): 816-824.
18. Chen HY, Yu SL, Chen CH, Chang GC, Chen CY, Yuan A, Cheng CL, Wang CH,

Terng HJ, Kao SF, Chan WK, Li HN, Liu CC, Singh S, Chen WJ, Chen JJ, Yang PC. A five-gene signature and clinical outcome in non-small-cell lung cancer. *N Engl J Med* 2007: 356(1): 11-20.

19. Roepman P, Jassem J, Smit EF, Muley T, Niklinski J, van de Velde T, Witteveen AT, Rzyman W, Floore A, Burgers S, Giaccone G, Meister M, Dienemann H, Skrzypski M, Kozlowski M, Mooi WJ, van Zandwijk N. An immune response enriched 72-gene prognostic profile for early-stage non-small-cell lung cancer. *Clin Cancer Res* 2009: 15(1): 284-290.

20. Boutros PC, Lau SK, Pintilie M, Liu N, Shepherd FA, Der SD, Tsao M-S, Penn LZ, Jurisica I. Prognostic gene signatures for non-small-cell lung cancer. *Proceedings of the National Academy of Sciences* 2009: 106(8): 2824-2828.

21. Potti A, Mukherjee S, Petersen R, Dressman HK, Bild A, Koontz J, Kratzke R, Watson MA, Kelley M, Ginsburg GS, West M, Harpole DH, Jr., Nevins JR. A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer. *N Engl J Med* 2006: 355(6): 570-580.

22. Pathology of lung cancer http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=cmed.section.20772. (Date last updated: 2000.).

23. Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* 2005: 365(9458): 488-492.

24. Jarvinen AK, Hautaniemi S, Edgren H, Auvinen P, Saarela J, Kallioniemi OP, Monni O. Are data from different gene expression microarray platforms comparable? *Genomics* 2004: 83(6): 1164-1168.

Supplementary Methods

Patient enrolment

Samples from patients recruited in this study were obtained from two Erasmus MC collections: the Tissue Bank and the Department of Internal Oncology. All lung tumor samples and adjacent non-cancerous specimens were collected from patients who had undergone curative surgical resection between 1992 and 1998 (Internal Oncology), or between 1996 and 2004 (Tissue Bank) at the Erasmus MC. Tissues were collected and studied under an anonymous tissue protocol approved by the local medical ethical committee of the institution. There were 91 patients with NSCLC included in our analysis.

The study comprised two independent validation sets. The first set included 6 normal lung tissues which were transcriptionally profiled by Affymetrix U133 plus 2.0 array (GSE3526). The second set comprised a cohort of 96 NSCLC patients collected at Duke University, including 50 ADC and 46 SCC samples [12]. Patient characteristics and the original microarray .CEL files were downloaded from http://data.genome.duke.edu/LungPotti.php. Eighty-nine out of those samples had relevant follow-up data available and were used for validating the performance of our survival signature.

Histopathological analysis

All tumor samples were independently reviewed by two pathologists. The cohort included 32/24 adenocarcinomas (ADC), 27/16 squamous cell carcinomas (SCC), and 13/24 large cell carcinomas. The remaining patients presented with rarer types of lung tumors, such as bronchioloalveolar (BAC), carcinoid (CAR), mixed adeno-squamous, or unknown. In the cohort of patients, over 57 percent had a known smoking history, with an average of 36.7 pack years. Of the 91 NSCLC patients, 51 were at stage I, 21 were at stage II, and 10 were at either stage III or IV. Three patients displayed distal metastases at the time of diagnosis. In addition, eight patients developed multiple primary tumors at different sites originating from the same cell type or different cell types, either synchronously or non-synchronously. Three had undergone neo-adjuvant radiation or chemotherapy before the surgery. Patient and tumor characteristics are listed in Table 1.

Definining the training and validation sets

All samples were divided into two subsets, the training set and the validation set, and the former was used to identify NSCLC related molecular signatures. As a result, thirty-six 'core' normal tissues were included in the training set, which showed strong similarities in global gene expression profile with each other and appeared in the core of normal lung cluster in an unsupervised clustering. Tumor samples were divided according to the two independent histopathological reviews, cancer cell contents, and degree of tumor differentiation.

For tumor samples, those from patients with a complete clinical record were assigned into the training set. Samples were excluded from the training set if they fell into anyone of the following cases,

1. From patients who developed multiple primary tumors;
2. Received chemotherapy or radiotherapy prior to the surgery;
3. Tumor cell content < 60%;

To develop histology signatures, additional criteria were employed to create a super training set to sketch a precise histological profile. Tumor samples had to meet below conditions:

1. Consistent classification between two histopathological reviews;
2. No cell type heterogeneity;

As a result, forty-four tumor samples were included in the training set, and twenty-three composed the super training set.

The remaining samples were used as a separate dataset for validating gene signatures identified by the training set. They were either from patients lacking complete clinic information or rejected by the above inclusion criteria, including eight LCC and five of rare types of NSCLC samples with a high level of cell type heterogeneity, and 19 percent (17 out of 91) of tumor samples had a discrepancy in histopathological classification.s

Total RNA isolation

The samples used in this study were fresh frozen tissues. Dissected tumors and adjacent normal tissues were snap-frozen in liquid nitrogen precooled isopentane immediately after the surgical resection, and stored at –196°°C or

–80 °C until RNA extraction. Specimens were sectioned in Cryostat into slices of 25 μm thick for RNA extraction. For each specimen, two thinner sections (10 μm) were taken at the start and end of collection, and used to determine the percentage of tumor cells. Samples were homogenized with a mortar and pestle in TRI Reagent (Invitrogen, Carlsbad, CA), and then incubated at room temperature for 5 minutes before adding 0.2 μl of chloroform for each 1ml sample. After centrifuging at full speed (12000 rpm) for 20 minutes, the supernatant containing the RNA was precipitated and centrifuged with isopropanol. The resultant RNA pellets were washed with 75% ethanol and solved in RNase-free water. If applicable, they were stored at –80 °C for further usage.

Assessment of RNA quality and concentration

The integrity if the isolated total RNA was verified on the Agilent 2100 BioAnalyzer (Agilent Technologies, Palo Alto, CA). Samples were kept for further processes if the 28s/18s ratio of its RNA was lower than 1.2. The concentrations of the RNAs were measured with a NanoDrop ND-111 UV-VIS spectrophotometer.

cRNA amplification and labelling

Double strand (ds) cDNA synthesis was performed according to the standardized protocol for One-Cycle cDNA synthesis from Affymetrix (Santa Clara, CA). Approximately 5 μg of total RNA was first converted to single strand cDNA in a 20 μl First-Strand Reaction Mix, containing poly-A control RNA, 100 μmol T7-Oligo Primer, 1x first strand buffer, 0.2 mol DTT 10 mmol dNTP mix and SuperScript II. In detail, the sample RNA, the poly-A control RNA and the T7-Oligo Primer were mixed and incubated for 10 min at 70 °C. Secondly, the first strand buffer, the DTT and the dNTP mix were added and incubated for 2 min at 42 °C, followed by adding SuperScript II and incubation of 1 hour at 42 °C. The ds cDNA was prepared from the resultant First-Strand Reaction Mix, mixed with 1x second strand reaction buffer, 30 mmol dNTP mix, E.coli DNA ligase, E.coli DNA Polymerase I and RNaseH. The mix was incubated for 2 hours at 16 °C, then supplemented with T4 DNA Polymerase, and then incubated for another 5 minnutes at 16 °C. The reaction was stopped by the addition of EDTA

to a final concentration of 5 μM. The Sample Cleanup Module and GeneChip IVT Labeling Kit from Affymetrix were used to purify the synthesized ds cDNA, which was used to generate biotin-labeled cRNA, in the presence of 1x IVT Labeling buffer, IVT Labeling NTP Mix, IVT Labeling Enzyme Mix and RNase-free water in a total volume of 40 μl. After an incubation of 16 hours at 37 °C, the concentration and quality of the labelled cRNA were checked with NanoDrop ND-1000 UV-VIS spectrophotometer. An A_{260}/A_{280} ratio between 1.9 and 2.1 was considered acceptable. Approximately 20 μg cRNA per array was fragmented to an average size of 35-200 nucleotides by heating at 94 °C for 35 min, in the presence of a 1x Fragmentation Buffer in a total volume of 40 μl. The undiluted, fragmented samples were stored at –20 °C before being subjected to hybridization.

Hybridization

Hybridization was conducted following Affymetrix instruction for GeneChip® Human Genome U133 plus 2.0 array. The GeneArray scanner 3000 (Affymetrix) was then employed to detect the hybridization signals.

Preprocessing microarray data

Array Quality Control

Microarrays that did not pass the quality assessment were removed from further analyses. The quality metrics used to exclude microarrays was the statistics summary calculated by the GCOS algorithm during the processing of probe-level data. The primary inclusion criteria include: all arrays had to have comparable noise values (Raw Q, measurement for the pixel-to-pixel variation of probe cells on the chip); background values were within the range of 20 to 100; percent of present probe sets on the array should not be below 45%. The other criteria were: arrays with extremely high or low values for any of these parameters, e.g. values beyond the range of standard deviation ± median, were excluded; signal ratio of ≤3 of the 3' / 5' probe sets for GAPDH and Actin were used as a cut-off; labelling and hybridization were controlled by using standard spike-in controls according to the Affymetrix protocol; if global scaling was applied, the scaling factors for each array were within a three-fold range.

Data normalization

Microarray data was processed at two levels: probe level and probe set level.

1. **At probe level by quantile normalization**

RMA (Robust Multi-Array average) is an integrated algorithm comprising background adjustment, quantile normalization, and expression summarization by median polish [1]. The intensities of mismatch probes were entirely ignored due to their spurious estimation of non-specific binding. The intensities were background-corrected in such a way that all corrected values must be positive. The RMA algorithm utilized quantile normalization in which the signal value of individual probes was substituted by the average of all probes with the same rank of intensity on each chip/array. Finally Tukey's median polish algorithm was used to obtain the estimates of expression for normalized probe intensities.

2. **At probe set level by Global Scaling (GCOS v1.4)**

This algorithm was a summary method embedded in GeneChip Operating Software (GCOS) from Affymetrix, and fully described in the data_analysis_fundamentals_manual. The signal intensity of each probe was firstly corrected by the overall background. The differences between perfect match (PM) and mismatch (MM) probes were examined by using background-adjusted intensities for each probe pair. The significance of the differences between PM and MM probe sets was reflected by a p-value calculated by one-sided Wilcoxon-signed rank test. The final signal for a probe set was assigned as the one-step biweight estimate of the combined differences of all probe pairs belonging to one probe set. The trimmed mean signal of each array was then scaled to the same Target Intensity (e.g. 250) by a global method to minimize technique-derived discrepancies.

Other transformations

Intensities of probe sets lower than 30 were reset to 30. The geometric mean for each probe set was calculated across all samples or for each subgroup of samples firstly and then across all samples (OmniViz). The intensity values of individual probe sets in each sample were then displayed as the log 2 of the deviations to the calculated geometric means.

Probe sets filtering

Probe sets were involved in further analysis only if their expression levels deviated from the overall mean in at least one array by a minimum factor of 2.5, because the remaining data were unlikely to be informative. The result was that 43,160 probe sets were eliminated, and 11,515 probe sets remained for further analysis.

Unsupervised clustering and visualization of gene/sample similarity

Clustering was performed without taking into account any external information such as histology subtypes and tumor stages, with each of the selected 11,515 probe sets using the K-means algorithm (OmniViz). Similarities were measured by magnitude and shape (Euclidean distance). Pair-wised similarities between samples were sorted and visualized by the Pearson Correlation Matrix (OmniViz). The order of clusters and individual samples within each cluster was sorted according to the Pearson Correlation Coefficient.

Statistical analysis

The resulting 11,515 probe sets from the filtering step was the starting point for all supervised analyses which, for instance, correlated gene expression with the clinical variables such as the histological subtype. Two-Class comparison analysis was performed by using Significance Analysis of Microarray (SAM), integrated in OminiViz version 5.1. Class prediction analysis was performed with the use of Prediction Analysis of Microarrays (PAM) software, integrated in BRBArray version 3.8. Clustering was performed using the Spotfire DecisionSite software (TIBCO, Palo Alto, CA). The samples were clustered with various signatures using the Weighted Pair-Group Method algorithm and similarity measured by Euclidean distance or correlation.

Class comparison

SAM discovered differentially expressed genes among different sample classes, e.g. between non-cancerous tissues and tumors or between a particalur histology subtype and the remaining samples [2]. Firstly this algorithm calculated the different expression for each gene between classes relative to the variation expected in the mean difference. To correct multiple testing, false discovery rate (FDR) was controlled by randomly permutating the classes of

samples 100 times. Signature probe sets for assigned classes were selected by a change factor of 2 and a FDR of less than 1 percent. The class comparisons were performed with both RMA- and GCOS-processed data. The common probe sets identified by both sets of data were selected as the final signatures.

Class prediction

The resultant signatures from Class Comparison were tested by the nearest shrunken centroids algorithm (PAM) to identify subgroups of genes that best characterized the predefined classes [3]. The prediction accuracy of optimized signatures was determined by performing "leave-one-out" cross validation within the training set, with one sample omitted each time and class label being predicted with other samples for the omitted sample [4]. The predictive models generated by the optimal subsets were subsequently applied to make predictions of classes for samples in the validation set, which were not involved in the corresponding class comparisons. The prediction accuracy on validation samples was calculated by comparing predicted class labels with the clinical histopathological diagnoses for those samples; samples without histopathological records were excluded from the calculations.

Survival analysis

Of 91 NSCLC samples, 82 have relevant follow-up data available. Therefore, those samples were included in survival analysis.

Two different approaches were used to determine whether the gene expression profile could predict the prognosis for NSCLC patients. In one approach, samples from patients who died of lung cancer within two years of surgical removal of tumors were assigned to the group of NSCLC with short-time survival. The long-time survival group consisted of samples from patients who survived for longer than 5 years. To avoid unexpected variances introduced by the failure of surgery or postoperative sequelae, those patients who died within six months of surgery were kept out of the analysis. Subsequently, the same analysis was performed conditionally for histological subtypes.

As an alternative way, we developed a step-wise approach based on gene expression profiles to classify NSCLC with respect to prognostic outcome. Firstly, probe sets which were the most likely associated with patient

prognosis were selected among over 11,000 probe sets by their correlation with the defined survival time; A list of candidate probe-sets was created with probe-sets whose univariate p-values, testing the hypothesis that survival time is independent of the expression level for that gene, was smaller than 0.001 by the Wald test in the Cox proportional hazards model [5]. A global test was performed with 1000 permutations to adjust p-values.

In the analysis of the probability that patients would remain free of death, survival time (OS) was defined as the date of surgery to the time of event happened – death, or the date on which data were censored - the last follow-up visit.

The resulting candidate survival probe-sets were subjected to a supervised principal component calculation described in details by Bair et al [6]. The computation of principal components was followed by Cox proportional hazards regression analysis using the computed principal components. As a result, a predictive prognosis model for NSCLC was determined, with regression coefficients derived from the Cox regression described above. With the developed model, a prognostic predictor was calculated for a NSCLC case whose expression profile was provided as the expression levels of selected probe-sets.

The predictive value of the prognosis model was evaluated by performing "leave-one-out" cross-validation", in which a single case was omitted each time and the entire procedure described above was performed to estimate prognosis predictor for the omitted case [6, 7]. This prognosis predictor value was compared and ranked relative to the prognosis predictors of cases included in the cross-validation training cases. Based on the predetermined cut-off percentile rank for defining the risk groups, the omitted case was placed into a risk group. This analysis was repeated until each sample was left out once, resulting in a set of unbiased prognosis prediction for all cross-validated samples.

Having obtained unbiased prognosis predictors and consequent categorizing patients, the difference in the survival outcome between risk groups was estimated by log-rank Mantel-Cox test and plotted by Kaplan-

Meier curve [8]. The analyses were performed with BRB-Array Tools (version 3.8; R.Simon and A.P.Lam, National Cancer Institute, Bethesda, MD).

To evaluate the prognostic value of the prognosis predictor relative to other clinical parameters, we used proportional hazard regression analysis with the defined survival time as dependent variable, death as the occurred event, and the last follow-up visit as the censored. The risk of death studied included age, tumor cell content (%), tumor size (diameter of tumor), smoking year, Forced Expiratory Volume 1, and gender, tumor histology, tumor grade, as well as computed prognosis predictor. The relation between them and the relative hazard ratio was tested with use of the Wald test. The 95% confidence interval for relative hazard ratios, and the p-values are listed in supplementary Table 7. To compare the performance in predicting the overall OS, the proportional hazard regression model was built with either involving a specific parameter or not. The contribution of each parameter to the model was evaluated by chi-square test and P-value was derived from the likelihood ratio test (Table 2) [5].

The correlation between the survival signature and clinical parameters was evaluated using predicted risk as grouping variable and with independent samples t-test for continuous variables, or non-parametric test, Mann-Whitney and maximum possibility Wald-Wolfowitz test, for categorical variables and scalar variables (Supplementary Table 6). Statistical analyses were performed with SPSS 15.0 (SPSS, Chicago, IL). For each tumor from NSCLC validation cohort, we calculated a prognosis predictor by fitting the predetermined predictive model with expression of the 17 probe sets. Patients were predicted with high-risk of death if their prognosis predictor percentile ranking was above the 60^{th}, as determined in the procedure of identifying prognosis signature using training samples.

Comparison with published prognostic signatures

If the original prognostic predictors were provided as gene symbols [9-13], we retrieved gene expression for the Erasmus MC and Duke University cohorts as follows. First, genes were mapped to the Affymetrix U133 plus 2.0 chip, and the corresponding expression data from all relevant probe sets was

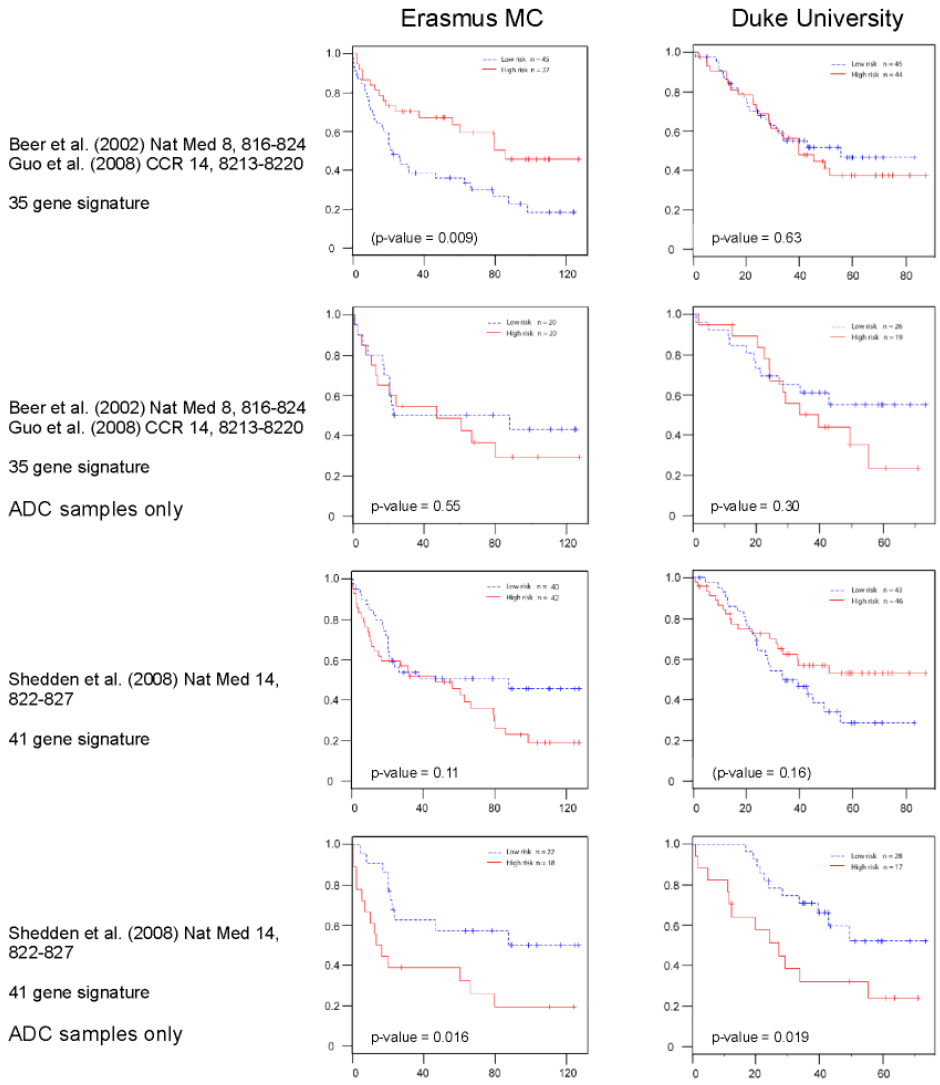
extracted (Supplementary Table 8). Next, probe set level data was converted to gene level data by averaging probe sets targeting the same genes. Due to the variation between platforms, 4 genes from the Roepman et al [12] signature were missing from the Affymetrix U133 plus 2.0 chip, we used the remaining 68 genes.

When the original prognostic predictors were supplied as probe sets [14, 15], either from Affymetrix U133A or U133 plus 2.0 arrays, the data was kept at probe set level. The Affymetrix HuGeneFL chip used by Beer et al / Guo et al [14, 15] deviates too much from the U133 plus 2.0 chip and we therefore used gene symbols to re-map the data to the U133 plus 2.0 chip.
Some studies provide multiple signature sets [10, 14, 15], in which case each signature set was tested. For all re-evaluations, a cut-off at the 50th and 60th percentile was used for dividing the two risk groups. We only show the results for the best stratification obtained (Supplementary Fig. 2).

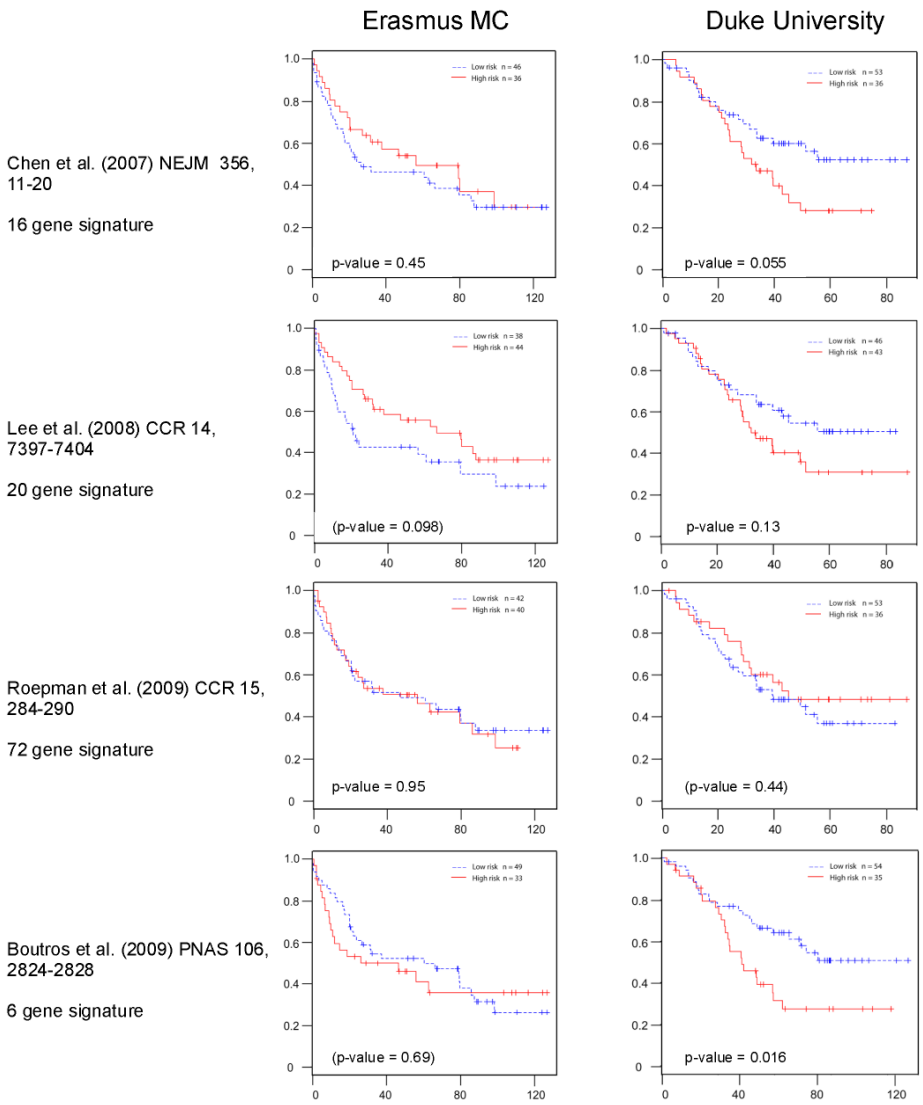
References

1. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 2003: 4(2): 249-264.
2. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* 2001: 98(9): 5116-5121.
3. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America* 2002: 99(10): 6567-6572.
4. Golub T, Slonim D, Tamayo P, Huard C, Gaasenbeek M, Mesirov J, Coller H, Loh M, Downing J, Caligiuri M, Bloomfield C, Lander E. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science (New York, NY* 1999: 286: 531-536.
5. Cox DR. Regression models and life-tables. *J R Stat Soc* 1972: 34: 187-220.
6. Bair E, Tibshirani R. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol* 2004: 2(4): E108.
7. Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute* 2003: 95(1): 14-18.
8. Meier P, Kaplan E. Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 1958: 158: 457-481.

9. Boutros PC, Lau SK, Pintilie M, Liu N, Shepherd FA, Der SD, Tsao M-S, Penn LZ, Jurisica I. Prognostic gene signatures for non-small-cell lung cancer. *Proceedings of the National Academy of Sciences* 2009: 106(8): 2824-2828.
10. Chen HY, Yu SL, Chen CH, Chang GC, Chen CY, Yuan A, Cheng CL, Wang CH, Terng HJ, Kao SF, Chan WK, Li HN, Liu CC, Singh S, Chen WJ, Chen JJ, Yang PC. A five-gene signature and clinical outcome in non-small-cell lung cancer. *N Engl J Med* 2007: 356(1): 11-20.
11. Guo NL, Wan Y-W, Tosun K, Lin H, Msiska Z, Flynn DC, Remick SC, Vallyathan V, Dowlati A, Shi X, Castranova V, Beer DG, Qian Y. Confirmation of Gene Expression-Based Prediction of Survival in Non-Small Cell Lung Cancer. *Clin Cancer Res* 2008: 14(24): 8213-8220.
12. Roepman P, Jassem J, Smit EF, Muley T, Niklinski J, van de Velde T, Witteveen AT, Rzyman W, Floore A, Burgers S, Giaccone G, Meister M, Dienemann H, Skrzypski M, Kozlowski M, Mooi WJ, van Zandwijk N. An immune response enriched 72-gene prognostic profile for early-stage non-small-cell lung cancer. *Clin Cancer Res* 2009: 15(1): 284-290.
13. Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizyness ML, Kuick R, Hayasaka S, Taylor JM, Iannettoni MD, Orringer MB, Hanash S. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med* 2002: 8(8): 816-824.
14. Lee ES, Son DS, Kim SH, Lee J, Jo J, Han J, Kim H, Lee HJ, Choi HY, Jung Y, Park M, Lim YS, Kim K, Shim Y, Kim BC, Lee K, Huh N, Ko C, Park K, Lee JW, Choi YS, Kim J. Prediction of recurrence-free survival in postoperative non-small cell lung cancer patients by using an integrated model of clinical information and gene expression. *Clin Cancer Res* 2008: 14(22): 7397-7404.
15. Shedden K, Taylor JM, Enkemann SA, Tsao MS, Yeatman TJ, Gerald WL, Eschrich S, Jurisica I, Giordano TJ, Misek DE, Chang AC, Zhu CQ, Strumpf D, Hanash S, Shepherd FA, Ding K, Seymour L, Naoki K, Pennell N, Weir B, Verhaak R, Ladd-Acosta C, Golub T, Gruidl M, Sharma A, Szoke J, Zakowski M, Rusch V, Kris M, Viale A, Motoi N, Travis W, Conley B, Seshan VE, Meyerson M, Kuick R, Dobbin KK, Lively T, Jacobson JW, Beer DG. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat Med* 2008: 14(8): 822-827.

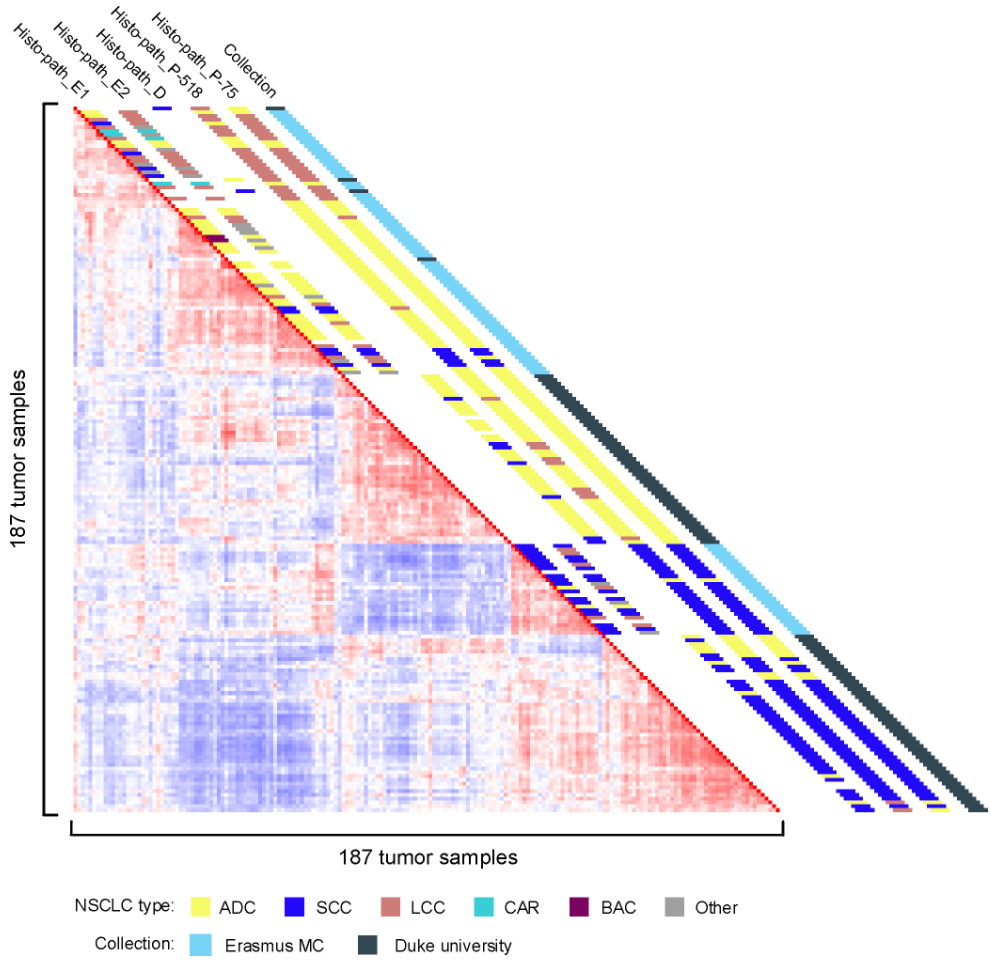


Supplementary Figure 1. Survival prediction by published prognostic signatures Kaplan-Meier curves for the best performing signatures (by P-value) are shown for 82 Erasmus MC patients (left) and 89 Duke University NSCLC patients (right), fitted by their risk assignments. Grey bars indicate patients at last follow-up, still alive. P-values are between brackets if overall survival of the low risk group is actually lower than that of the high risk group. (continued on next page)



Supplementary Figure 1 (continued). Survival prediction by published prognostic signatures

Kaplan-Meier curves for the best performing signatures (by P-value) are shown for 82 Erasmus MC patients (left) and 89 Duke University NSCLC patients (right), fitted by their risk assignments. Grey bars indicate patients at last follow-up, still alive. P-values are between brackets if overall survival of the low risk group is actually lower than that of the high risk group.



Supplementary Figure 2. Correlation view of Erasmus MC and Duke University NSCLC samples

In total 187 tumor samples from the Erasmus MC (n=91) and Duke University (n=96) cohorts are shown. Pairwise correlations between any two samples are displayed, based on 3495 informative probe sets. Histological classification of the samples, and the collection source, are depicted along the diagonal. The key to the color code is shown at the bottom. Histo-path_E1 & Histo-path_E2: initial and second histo-pathological review of Erasmus MC samples. Histo-path_D: histo-pathological review of Duke University samples; Histo-path_P-518 and Histo-path_P-75: predictions by PAM of histological subtypes using the 518 and 75 probe set signatures, respectively (see Supplementary Tables 3 and 4).

Legend to Supplementary Tables

Supplementary Tables 1 and 2

T:N ratio	Ratio of average expression in NSCLC samples / normal lung tissue
T mean	2log transformation of mean expression value in NSCLC samples (average of all NSCLC and normal lung tissue = 0).
N mean	2log transformation of mean expression value in normal lung tissue samples (average of all NSCLC and normal lung tissue = 0).
T SD	Standard deviation of mean expression value in NSCLC samples
N SD	Standard deviation of mean expression value in normal lung tissue samples

Supplementary Tables 3 and 4

ADC:OT ratio	Ratio of average expression in ADC samples / the other NSCLC samples (SCC and LCC)
SCC:OT ratio	Ratio of average expression in SCC samples / the other NSCLC samples (ADC and LCC)
LCC:OT ratio	Ratio of average expression in LCC samples / the other NSCLC samples (ADC and SCC)
ADC mean	2log transformation of mean expression value in ADC samples (average of all Erasmus MC samples = 0).
SCC mean	2log transformation of mean expression value in SCC samples (average of all Erasmus MC samples = 0).
LCC mean	2log transformation of mean expression value in LCC samples (average of all Erasmus MC samples = 0).
ADC SD	Standard deviation of mean expression value in ADC samples
SCC SD	Standard deviation of mean expression value in SCC samples
LCC SD	Standard deviation of mean expression value in LCC samples

Supplementary Tables 5, 6, 7 and 8

Supplementary Tables 5, 6, 7 and 8 are self-explanatory.

SUPPLEMENTARY TABLE 1: NSCLC TUMOR SIGNATURE

Probe set	Gene Symbol	Gene Symbol before 2008	T:N Ratio	T Mean	N Mean	T SD	N SD	q value
213664_at	SLC1A1	SLC1A1	0.13	-1.65	1.92	1.41	0.57	0.01
226625_at	TGFBR3	TGFBR3	0.18	-1.08	1.70	0.98	0.43	0.01
227826_s_at	---	SORBS2	0.14	-1.28	1.90	1.07	0.62	0.01
227874_at	EMCN		0.10	-1.58	2.24	1.16	0.65	0.01
228504_at	---		0.10	-1.77	2.45	1.53	0.44	0.01
229127_at	JAM2	ATP5J	0.16	-1.21	1.71	0.83	0.45	0.01
229308_at	---		0.13	-1.42	2.11	1.13	0.55	0.01
229339_at	---		0.32	-0.67	0.95	0.33	0.44	0.01
229641_at	---		0.15	-1.25	1.76	0.83	0.35	0.01
230711_at	---	EPAS1	0.19	-0.86	1.46	0.55	0.76	0.01
233392_at	---		0.19	-0.92	1.43	0.42	0.71	0.01
235108_at	---		0.18	-0.96	1.66	0.83	0.62	0.01
235570_at	---		0.21	-1.10	1.46	1.03	0.30	0.01
235670_at	STX11	STX11	0.16	-1.31	1.79	1.02	0.44	0.01
235915_at	---		0.15	-1.07	1.36	0.39	1.04	0.01
236065_at	---		0.13	-1.40	1.72	0.88	0.74	0.01
236383_at	---		0.15	-1.13	1.60	0.58	0.77	0.01
236936_at	---		0.13	-1.39	1.79	0.81	0.59	0.01
238178_at	---		0.13	-1.06	1.89	0.68	0.83	0.01
239262_at	---	PRSS23	0.15	-1.15	1.67	0.64	0.55	0.01
239849_at	---		0.10	-1.21	1.99	0.59	1.02	0.01
242009_at	SLC6A4		0.06	-1.66	2.97	1.07	1.52	0.01
242500_at	---		0.17	-0.99	1.67	0.71	0.60	0.01
242868_at	---	EPAS1	0.16	-1.16	1.75	1.14	0.62	0.01
243172_at	---		0.18	-0.94	1.60	0.58	0.56	0.01
243813_at	---		0.15	-1.05	1.75	0.59	0.77	0.01
243818_at	SFTA1P		0.14	-1.35	1.86	0.93	0.39	0.01
204719_at	ABCA8	ABCA8	0.09	-1.83	2.36	1.32	0.39	0.01
223395_at	ABI3BP	ABI3BP	0.16	-1.33	1.76	1.27	0.30	0.01
206069_s_at	ACADL	ACADL	0.11	-1.35	1.87	0.64	0.80	0.01
220677_s_at	ADAMTS8	ADAMTS8	0.12	-1.26	1.95	0.69	0.67	0.01
235649_at	ADAMTS8	ADAMTS8	0.08	-1.73	2.45	1.11	0.58	0.01
203865_s_at	ADARB1	ADARB1	0.15	-1.25	1.70	0.86	0.38	0.01
209612_s_at	ADH1B	ADH1B	0.09	-2.29	2.69	2.20	0.55	0.01

209613_s_at	ADH1B	ADH1B	0.09	-2.32	2.81	2.29	0.46	0.01
209614_at	ADH1B	ADH1B	0.11	-1.49	1.78	0.75	0.61	0.01
229309_at	ADRB1	ADRB1	0.08	-1.90	2.65	1.59	0.72	0.01
206170_at	ADRB2	ADRB2	0.16	-1.34	1.72	1.16	0.53	0.01
210081_at	AGER	AGER	0.06	-2.06	3.05	1.39	0.54	0.01
217046_s_at	AGER	AGER	0.10	-1.41	2.25	0.86	0.51	0.01
202759_s_at	AKAP2 /// PALM2 /// PALM2-AKAP2	PALM2-AKAP2	0.20	-1.13	1.46	0.98	0.43	0.01
226694_at	AKAP2 /// PALM2 /// PALM2-AKAP2	PALM2-AKAP2	0.24	-0.95	1.29	0.86	0.36	0.01
224339_s_at	ANGPTL1	ANGPTL1	0.27	-0.81	0.96	0.36	0.73	0.01
231773_at	ANGPTL1	ANGPTL1	0.23	-0.93	1.15	0.52	0.64	0.01
218418_s_at	KANK2	ANKRD25	0.28	-0.90	1.08	0.71	0.30	0.01
213715_s_at	KANK3	ANKRD47	0.21	-0.86	1.43	0.60	0.59	0.01
222608_s_at	ANLN	ANLN	21.61	1.55	-2.35	1.67	0.54	0.01
204894_s_at	AOC3	AOC3	0.15	-1.38	1.73	1.14	0.36	0.01
206167_s_at	ARHGAP6	ARHGAP6	0.22	-1.07	1.72	1.14	0.43	0.01
206030_at	ASPA	ASPA	0.21	-0.97	1.21	0.43	0.60	0.01
219918_s_at	ASPM	ASPM	16.04	1.39	-2.22	1.52	0.58	0.01
203296_s_at	ATP1A2	ATP1A2	0.20	-0.94	1.40	0.60	0.67	0.01
228434_at	BTNL9	BTNL9	0.11	-1.33	2.29	1.23	0.73	0.01
203571_s_at	C10orf116	C10orf116	0.13	-1.58	1.85	1.32	0.47	0.01
218723_s_at	C13orf15	C13orf15	0.17	-1.22	1.71	0.97	0.30	0.01
235568_at	C19orf59	TRAPPC5	0.10	-1.86	2.45	1.44	0.64	0.01
239349_at	C1QTNF7	C1QTNF7	0.14	-1.24	1.82	0.68	0.74	0.01
213900_at	C9orf61	C9orf61	0.15	-1.26	1.83	0.93	0.38	0.01
206208_at	CA4	CA4	0.17	-1.03	1.68	0.70	0.52	0.01
206209_s_at	CA4	CA4	0.08	-1.60	2.64	1.03	0.55	0.01
203065_s_at	CAV1	CAV1	0.15	-1.37	1.85	1.21	0.43	0.01
212097_at	CAV1	CAV1	0.16	-1.33	1.70	1.11	0.24	0.01
203323_at	CAV2	CAV2	0.14	-1.44	1.88	1.32	0.35	0.01
214710_s_at	CCNB1	CCNB1	14.31	1.33	-2.18	1.32	0.48	0.01
202705_at	CCNB2	CCNB2	14.02	1.37	-2.02	1.45	0.47	0.01
228766_at	CD36	CD36	0.11	-1.78	2.21	1.46	0.72	0.01
202870_s_at	CDC20	CDC20	17.30	1.50	-2.15	1.53	0.45	0.01

204677_at	CDH5	CDH5	0.16	-1.19	1.75	0.98	0.39	0.01
214135_at	CLDN18	CLDN18	0.07	-2.03	2.83	1.75	0.35	0.01
232578_at	CLDN18	CLDN18	0.08	-2.12	2.85	1.53	0.84	0.01
213317_at	CLIC5	CLIC5	0.08	-1.99	2.77	1.79	0.42	0.01
217628_at	CLIC5	CLIC5	0.21	-0.93	1.39	0.59	0.67	0.01
219866_at	CLIC5 ///	CLIC5	0.12	-1.52	2.15	1.18	0.56	0.01
211343_s_at	COL13A1	COL13A1	0.21	-1.01	1.38	0.78	0.44	0.01
230867_at	COL6A6	COL6A6	0.12	-1.58	2.04	1.31	0.54	0.01
227178_at	CUGBP2	CUGBP2	0.21	-1.07	1.36	0.81	0.29	0.01
1555497_a_at	CYP4B1	CYP4B1	0.09	-1.69	2.20	1.26	0.93	0.01
210762_s_at	DLC1	DLC1	0.19	-1.20	1.64	1.19	0.48	0.01
224822_at	DLC1	DLC1	0.16	-1.33	1.79	1.20	0.40	0.01
203764_at	DLGAP5	DLG7	18.09	1.40	-2.27	1.53	0.36	0.01
205003_at	DOCK4	DOCK4	0.27	-0.93	1.19	0.90	0.32	0.01
219597_s_at	DUOX1	DUOX1	0.15	-1.29	1.74	1.08	0.60	0.01
204642_at	S1PR1	EDG1	0.20	-1.00	1.37	0.64	0.57	0.01
204271_s_at	EDNRB	EDNRB	0.11	-1.39	2.02	1.00	0.63	0.01
204273_at	EDNRB	EDNRB	0.08	-1.81	2.49	1.26	0.68	0.01
206701_x_at	EDNRB	EDNRB	0.10	-1.45	2.14	0.88	0.66	0.01
228967_at	EIF1	EIF1	0.27	-0.74	1.23	0.62	0.48	0.01
219436_s_at	EMCN	EMCN	0.15	-1.31	1.74	1.10	0.69	0.01
222885_at	EMCN	EMCN	0.17	-1.20	1.70	1.12	0.52	0.01
203980_at	FABP4	FABP4	0.09	-2.11	2.66	1.72	0.63	0.01
207547_s_at	FAM107A	FAM107A	0.10	-1.28	2.08	0.69	0.69	0.01
209074_s_at	FAM107A	FAM107A	0.08	-1.64	2.67	1.21	0.71	0.01
205866_at	FCN3	FCN3	0.07	-1.83	2.77	1.38	0.90	0.01
201540_at	FHL1	FHL1	0.14	-1.46	1.94	1.22	0.27	0.01
210299_s_at	FHL1	FHL1	0.11	-1.77	2.29	1.40	0.51	0.01
220170_at	FHL5	FHL5	0.16	-1.14	1.63	0.81	0.72	0.01
206742_at	FIGF	FIGF	0.08	-1.77	2.45	1.20	0.67	0.01
1556325_at	FILIP1	FILIP1	0.17	-1.17	1.73	1.00	0.61	0.01
228268_at	FMO2	FMO2	0.12	-1.45	2.17	1.33	0.48	0.01
205935_at	FOXF1	FOXF1	0.17	-1.09	1.62	0.84	0.59	0.01
228568_at	GCOM1 /// GRINL1A	Gcom1	0.15	-1.16	1.72	0.74	0.60	0.01
206159_at	GDF10	GDF10	0.12	-1.26	1.84	0.50	0.67	0.01
206102_at	GINS1	GINS1	15.94	1.47	-2.11	1.46	0.46	0.01

238222_at	GKN2	GKN2	0.08	-1.92	2.58	1.28	0.57	0.01
230360_at	GLDN	GLDN	0.11	-1.38	2.09	0.99	0.80	0.01
209469_at	GPM6A	GPM6A	0.11	-1.58	2.40	1.12	0.48	0.01
209470_s_at	GPM6A	GPM6A	0.09	-1.75	2.62	1.25	0.58	0.01
232267_at	GPR133	GPR133	0.20	-1.25	1.79	1.38	0.30	0.01
204396_s_at	GRK5	GRK5	0.24	-0.87	1.44	0.79	0.35	0.01
238018_at	FAM150B		0.09	-1.40	1.91	0.69	0.96	0.01
213069_at	HEG1	HEG1	0.28	-0.88	1.15	0.78	0.32	0.01
203914_x_at	HPGD	HPGD	0.11	-1.73	2.18	1.64	0.62	0.01
211548_s_at	HPGD	HPGD	0.11	-1.81	2.26	1.56	0.67	0.01
205700_at	HSD17B6	HSD17B6	0.12	-1.62	1.97	1.42	0.59	0.01
37512_at	HSD17B6	HSD17B6	0.12	-1.62	2.08	1.43	0.61	0.01
230670_at	IGSF10	IGSF10	0.18	-1.11	1.53	0.83	0.52	0.01
224061_at	INMT	INMT	0.12	-1.44	2.07	1.24	0.68	0.01
235666_at	---		0.14	-1.41	1.83	1.05	0.37	0.01
219064_at	ITIH5	ITIH5	0.22	-1.12	1.40	0.92	0.38	0.01
219213_at	JAM2	JAM2	0.23	-0.99	1.25	0.70	0.45	0.01
202503_s_at	KIAA0101	KIAA0101	13.20	1.34	-2.16	1.31	0.75	0.01
209408_at	KIF2C	KIF2C	9.03	1.15	-1.70	1.20	0.38	0.01
206481_s_at	LDB2	LDB2	0.18	-1.16	1.63	1.01	0.49	0.01
203766_s_at	LMOD1	LMOD1	0.22	-0.94	1.31	0.71	0.52	0.01
220244_at	LOH3CR2A	LOH3CR2A	0.12	-1.35	1.98	0.94	0.69	0.01
220003_at	LRRC36	LRRC36	0.19	-1.02	1.38	0.46	0.62	0.01
203362_s_at	MAD2L1	MAD2L1	13.64	1.28	-2.05	1.37	0.45	0.01
228885_at	MAMDC2	MAMDC2	0.07	-1.90	2.72	1.38	0.60	0.01
212713_at	MFAP4	MFAP4	0.13	-1.43	1.86	1.08	0.67	0.01
219909_at	MMP28	MMP28	0.23	-1.03	1.26	0.75	0.49	0.01
239272_at	MMP28	MMP28	0.16	-1.43	1.57	0.98	0.49	0.01
239273_s_at	MMP28	MMP28	0.26	-1.05	1.17	0.92	0.44	0.01
219091_s_at	MMRN2	MMRN2	0.22	-0.96	1.36	0.69	0.38	0.01
236262_at	MMRN2	MMRN2	0.19	-1.02	1.57	0.84	0.47	0.01
227417_at	MOSC2	MOSC2	0.21	-1.11	1.38	0.88	0.35	0.01
226856_at	MUSTN1	TMEM110	0.25	-0.86	1.20	0.61	0.65	0.01
212372_at	MYH10	MYH10	0.29	-0.77	1.13	0.70	0.37	0.01
237206_at	MYOCD	MYOCD	0.23	-0.85	1.22	0.52	0.63	0.01
204641_at	NEK2	NEK2	13.66	1.38	-1.92	1.37	0.29	0.01
223381_at	NUF2	NUF2	14.93	1.43	-1.94	1.47	0.30	0.01
218736_s_at	PALMD	PALMD	0.18	-1.21	1.46	0.83	0.56	0.01

227088_at	PDE5A	PDE5A	0.19	-1.15	1.42	0.82	0.47	0.01
209493_at	PDZD2	PDZD2	0.13	-1.70	1.85	1.29	0.44	0.01
208981_at	PECAM1	PECAM1	0.29	-0.84	1.14	0.77	0.30	0.01
206311_s_at	PLA2G1B	PLA2G1B	0.13	-1.46	1.96	1.04	0.73	0.01
227148_at	PLEKHH2	PLEKHH2	0.15	-1.30	1.72	0.95	0.47	0.01
201578_at	PODXL	PODXL	0.29	-0.81	1.05	0.65	0.40	0.01
227006_at	PPP1R14A	PPP1R14A	0.17	-1.12	1.70	1.00	0.55	0.01
226380_at	PTPN21	PTPN21	0.16	-1.23	1.73	0.96	0.52	0.01
205846_at	PTPRB	PTPRB	0.17	-1.13	1.65	0.85	0.63	0.01
230250_at	PTPRB	PTPRB	0.12	-1.46	2.11	1.25	0.59	0.01
205326_at	RAMP3	RAMP3	0.15	-1.08	1.70	0.72	0.69	0.01
210550_s_at	RASGRF1	RASGRF1	0.27	-0.91	1.03	0.68	0.59	0.01
205407_at	RECK	RECK	0.24	-1.01	1.29	0.88	0.42	0.01
243481_at	RHOJ	RHOJ	0.27	-0.79	1.13	0.57	0.61	0.01
226028_at	ROBO4	ROBO4	0.15	-1.12	1.81	0.79	0.50	0.01
235849_at	SCARA5	SCARA5	0.19	-1.10	1.19	0.67	0.86	0.01
236359_at	SCN4B	SCN4B	0.18	-0.97	1.64	0.76	0.77	0.01
222717_at	SDPR	SDPR	0.13	-1.44	2.21	1.36	0.51	0.01
215454_x_at	SFTPC	SFTPC	0.09	-1.85	2.38	1.39	0.51	0.01
226673_at	SH2D3C	SH2D3C	0.26	-0.80	1.18	0.60	0.57	0.01
218087_s_at	SORBS1	SORBS1	0.19	-1.31	1.79	1.26	0.42	0.01
215918_s_at	SPTBN1	SPTBN1	0.16	-1.08	1.75	0.74	0.39	0.01
227480_at	SUSD2	SUSD2	0.11	-1.77	2.33	1.52	0.40	0.01
234310_s_at	SUSD2	SUSD2	0.22	-0.95	1.30	0.62	0.51	0.01
209447_at	SYNE1	SYNE1	0.26	-0.94	1.22	0.92	0.33	0.01
200911_s_at	TACC1	TACC1	0.31	-0.82	1.04	0.77	0.21	0.01
204931_at	TCF21	TCF21	0.08	-1.76	2.21	1.01	0.60	0.01
229529_at	TCF21	TCF21	0.17	-1.08	1.54	0.68	0.69	0.01
206702_at	TEK	TEK	0.13	-1.35	1.97	1.00	0.48	0.01
219230_at	TMEM100	TMEM100	0.07	-2.11	2.84	1.64	0.94	0.01
209904_at	TNNC1	TNNC1	0.12	-1.45	2.35	1.30	0.47	0.01
221747_at	TNS1	TNS1	0.19	-1.05	1.53	0.79	0.38	0.01
221748_s_at	TNS1	TNS1	0.22	-0.98	1.51	0.97	0.33	0.01
206093_x_at	TNXA /// TNXB	TNXB	0.15	-1.17	1.64	0.70	0.49	0.01
213451_x_at	TNXA /// TNXB	TNXB	0.14	-1.19	1.76	0.70	0.51	0.01
216333_x_at	TNXA /// TNXB	TNXB	0.14	-1.19	1.70	0.69	0.49	0.01

201291_s_at	TOP2A	TOP2A	32.13	1.86	-2.84	1.65	0.82	0.01
201292_at	TOP2A	TOP2A	23.60	1.58	-2.47	1.60	0.41	0.01
210052_s_at	TPX2	TPX2	17.03	1.38	-2.15	1.53	0.39	0.01
202954_at	UBE2C	UBE2C	15.20	1.42	-1.96	1.46	0.36	0.01
223229_at	UBE2T	UBE2T	12.18	1.31	-1.91	1.21	0.34	0.01
232122_s_at	VEPH1	VEPH1	0.10	-1.76	2.19	1.19	0.55	0.01
220327_at	VGLL3	VGLL3	0.16	-1.30	1.74	0.92	0.47	0.01
205019_s_at	VIPR1	VIPR1	0.13	-1.08	2.05	1.00	0.94	0.01
222738_at	WWC2	WWC2	0.18	-1.12	1.52	0.73	0.51	0.01
1555800_at	ZNF385B	ZNF533	0.12	-1.70	1.95	1.36	0.83	0.01

SUPPLEMENTARY TABLE 2: NSCLC TUMOR SIGNATURE (SHORT)

Probe set	Gene Symbol	Gene Symbol before 2008	T:N Ratio	T Mean	N Mean	T SD	N SD	q value
210081_at	AGER	AGER	0.06	-2.06	3.05	1.39	0.54	0.01
206209_s_at	CA4	CA4	0.08	-1.6	2.64	1.03	0.55	0.01
209074_s_at	FAM107A	FAM107A	0.08	-1.64	2.67	1.21	0.71	0.01
238222_at	GKN2	GKN2	0.08	-1.92	2.58	1.28	0.57	0.01
201291_s_at	TOP2A	TOP2A	32.13	1.86	-2.84	1.65	0.82	0.01

SUPPLEMENTARY TABLE 3: NSCLC HISTOLOGY SIGNATURE

Probe set	Gene Symbol	Signature	ADC:OT Ratio	SCC:OT Ratio	LCC:OT Ratio	ADC Mean	SCC Mean	LCC Mean	ADC SD	LCC SD	SCC SD
1554246_at	C1orf210	ADC	2.79	0.47	0.49	0.84	-0.49	-0.49	0.73	0.27	0.26
1555950_a_at	CD55	ADC	4.29	0.35	0.31	0.28	-1.20	-1.78	1.30	1.07	0.41
1566766_a_at	7A5	ADC	3.33	0.47	0.35	0.64	-0.61	-1.20	1.15	0.92	0.61
1569208_a_at	---	ADC	2.16	0.60	0.57	0.04	-0.98	-1.20	0.53	0.75	0.34
201042_at	TGM2	ADC	3.19	0.52	0.33	0.32	-0.95	-1.77	1.04	1.34	1.05
201474_s_at	ITGA3	ADC	3.71	0.45	0.30	0.69	-0.77	-1.57	1.10	1.18	0.70
201596_x_at	KRT18	ADC	2.51	0.61	0.42	0.98	-0.18	-1.15	0.43	1.58	0.59
201818_at	LPCAT1	ADC	2.99	0.35	0.59	0.53	-1.38	-1.14	0.84	1.27	0.92
201925_s_at	CD55	ADC	4.03	0.36	0.35	0.23	-1.55	-1.89	1.02	1.15	0.84
201926_s_at	CD55	ADC	4.09	0.36	0.34	0.30	-1.17	-1.68	1.29	1.09	0.48
201941_at	CPD	ADC	2.54	0.52	0.52	0.68	-0.60	-0.81	0.89	1.13	0.76
202071_at	SDC4	ADC	2.76	0.51	0.46	0.34	-0.96	-1.74	0.67	1.61	0.44
202179_at	BLMH	ADC	0.48	1.72	1.21	-0.51	0.76	0.49	0.68	0.35	0.47
202454_s_at	ERBB3	ADC	3.94	0.47	0.24	1.49	-0.21	-1.56	0.65	1.58	0.84
202973_x_at	FAM13A1	ADC	2.42	0.53	0.55	0.38	-0.86	-1.22	0.59	1.21	0.48
203002_at	AMOTL2	ADC	2.14	0.55	0.64	0.22	-0.86	-0.89	0.61	0.92	0.36
203438_at	STC2	ADC	0.33	2.63	1.00	-0.91	0.83	0.15	0.58	0.74	1.10
203908_at	SLC4A4	ADC	5.01	0.30	0.29	0.11	-1.71	-2.14	1.56	1.24	0.85
203913_s_at	HPGD	ADC	7.80	0.23	0.15	-0.17	-2.17	-3.33	1.59	1.44	1.07
203953_s_at	CLDN3	ADC	7.42	0.11	0.33	1.96	-1.56	-0.70	1.05	1.57	0.44
203954_x_at	CLDN3	ADC	4.74	0.21	0.43	1.45	-1.00	-0.42	1.05	1.16	0.38
204160_s_at	ENPP4	ADC	2.27	0.31	0.93	0.20	-1.72	-0.61	0.50	1.07	0.54
204161_s_at	ENPP4	ADC	2.03	0.36	0.96	0.16	-1.49	-0.45	0.39	0.73	0.46

204174_at	ALOX5AP	ADC	3.28	0.48	0.36	0.00	-1.70	-2.33	0.90	1.64	1.32
204351_at	S100P	ADC	7.97	0.27	0.11	2.07	0.00	-2.15	1.91	1.91	1.64
204401_at	KCNN4	ADC	3.39	0.36	0.46	1.30	-0.34	-0.21	1.12	0.84	0.61
204437_s_at	FOLR1	ADC	5.90	0.30	0.20	0.06	-2.23	-3.33	1.52	2.01	1.83
204667_at	FOXA1	ADC	3.07	0.48	0.40	1.33	-0.09	-0.48	1.15	1.13	0.97
204885_s_at	MSLN	ADC	21.61	0.10	0.05	0.86	-1.59	-2.47	2.70	0.40	0.98
204934_s_at	HPN	ADC	3.95	0.37	0.35	0.97	-0.81	-0.98	0.79	0.57	0.28
204988_at	FGB	ADC	119.83	0.01	0.01	3.39	-0.66	-0.98	4.05	0.18	0.21
205076_s_at	MTMR11	ADC	2.69	0.52	0.46	0.61	-0.66	-0.93	0.62	0.68	0.42
205186_at	DNALI1	ADC	3.62	0.37	0.41	0.71	-0.77	-1.01	1.32	1.01	0.51
205190_at	PLS1	ADC	2.29	0.51	0.62	0.81	-0.55	-0.48	0.78	1.12	0.88
205309_at	SMPDL3B	ADC	5.35	0.22	0.34	1.43	-0.96	-0.51	1.07	0.63	0.50
205414_s_at	RICH2	ADC	2.13	0.51	0.69	-0.17	-1.47	-1.20	0.60	0.83	0.79
205455_at	MST1R	ADC	2.98	0.45	0.46	0.77	-0.73	-0.73	0.76	0.56	0.53
205597_at	SLC44A4	ADC	6.24	0.24	0.24	1.19	-1.14	-1.57	1.25	1.12	0.65
205640_at	ALDH3B1	ADC	4.06	0.38	0.32	0.25	-1.45	-1.87	0.86	0.69	0.30
205650_s_at	FGA	ADC	32.25	0.05	0.06	1.89	-1.48	-1.25	3.10	0.73	0.78
205776_at	FMO5	ADC	9.73	0.14	0.18	0.77	-1.62	-1.48	1.97	0.71	0.40
205906_at	FOXJ1	ADC	2.41	0.59	0.48	0.62	-0.29	-0.65	0.85	0.55	0.10
205997_at	ADAM28	ADC	2.10	0.62	0.58	1.14	0.08	-0.24	0.51	1.26	0.64
206100_at	CPM	ADC	4.70	0.44	0.17	0.47	-1.14	-2.74	1.29	1.37	1.10
206595_at	CST6	ADC	11.24	0.16	0.12	0.66	-1.48	-2.06	2.01	0.95	0.74
207836_s_at	RBPMS	ADC	2.10	0.84	0.36	0.04	-0.58	-1.71	0.72	0.80	0.59
207847_s_at	MUC1	ADC	5.36	0.36	0.18	0.85	-1.57	-2.61	1.06	1.58	1.64

208161_s_at	ABCC3	ADC	5.10	0.41	0.16	1.40	-0.30	-1.65	1.49	1.15	1.08
209016_s_at	KRT7	ADC	7.97	0.19	0.19	1.48	-1.70	-2.93	0.72	2.43	1.06
209094_at	DDAH1	ADC	2.76	0.33	0.69	0.73	-1.17	-0.35	0.67	0.75	0.64
209278_s_at	TFPI2	ADC	11.47	0.03	0.27	1.62	-2.01	-1.24	2.69	2.57	1.33
209487_at	RBPMS	ADC	2.53	0.63	0.39	0.14	-0.81	-1.57	0.84	0.95	0.60
209499_x_at	TNFSF12 /// TNFSF12-TNFSF13 /// TNFSF13	ADC	2.17	0.69	0.47	0.08	-0.99	-1.54	0.38	0.95	0.91
209581_at	HRASLS3	ADC	2.63	0.74	0.27	0.55	-0.53	-1.52	0.84	0.59	1.03
209641_s_at	ABCC3	ADC	4.34	0.44	0.21	1.25	-0.24	-1.19	1.45	0.85	0.95
209900_s_at	SLC16A1	ADC	0.32	1.87	1.52	-0.64	1.13	0.95	0.88	0.89	0.72
209925_at	OCLN	ADC	3.62	0.34	0.44	0.64	-1.04	-0.96	1.23	0.95	0.53
210272_at	CYP2B7P1	ADC	5.61	0.26	0.27	0.15	-1.68	-1.84	1.62	0.87	0.53
210325_at	CD1A	ADC	3.61	0.41	0.36	0.72	-0.51	-0.70	1.33	0.36	0.38
210347_s_at	BCL11A	ADC	0.20	2.09	1.80	-0.41	1.78	1.67	0.49	1.05	0.93
210547_x_at	ICA1	ADC	2.10	0.38	0.89	0.64	-0.98	-0.15	0.52	1.03	0.65
210567_s_at	SKP2	ADC	0.38	1.15	2.20	-0.29	0.64	1.08	0.31	0.97	0.71
210830_s_at	PON2	ADC	3.67	0.38	0.39	0.66	-0.92	-1.03	1.23	0.84	0.67
211024_s_at	NKX2-1	ADC	4.61	0.24	0.41	0.74	-1.72	-2.62	1.55	2.20	1.17
211695_x_at	MUC1	ADC	4.43	0.40	0.24	0.54	-1.34	-2.07	1.00	0.89	0.92
212325_at	LIMCH1	ADC	2.44	0.39	0.71	0.10	-1.48	-1.50	0.85	1.80	0.71
212327_at	LIMCH1	ADC	2.97	0.35	0.59	-0.24	-2.30	-2.44	0.51	1.68	1.07
212328_at	LIMCH1	ADC	2.67	0.48	0.51	-0.16	-1.93	-2.46	0.64	1.73	1.61

212444_at	---	ADC	5.28	0.40	0.15	0.39	-1.83	-3.43	0.84	1.59	1.54
213050_at	COBL	ADC	3.26	0.32	0.55	0.17	-1.73	-1.55	0.91	1.34	0.65
213094_at	GPR126	ADC	4.13	0.30	0.39	0.06	-2.00	-1.89	1.16	1.04	0.79
213693_s_at	MUC1	ADC	4.78	0.37	0.23	0.87	-1.21	-2.39	0.87	1.89	1.08
213695_at	PON3	ADC	5.39	0.24	0.31	1.01	-1.63	-1.98	1.11	1.95	1.25
213943_at	TWIST1	ADC	0.13	1.41	3.15	-0.83	1.43	1.57	0.50	1.96	1.10
214033_at	ABCC6	ADC	4.02	0.30	0.42	0.42	-1.70	-1.38	1.00	0.96	0.87
214469_at	HIST1H2AB /// HIST1H2AE	ADC	0.29	0.97	3.04	-0.29	0.91	1.63	0.47	1.29	0.53
214835_s_at	SUCLG2	ADC	2.22	0.64	0.50	0.55	-0.39	-0.71	0.70	0.51	0.51
216238_s_at	FGB	ADC	82.61	0.02	0.02	3.11	-0.74	-0.70	3.59	0.19	0.13
216836_s_at	ERBB2	ADC	2.37	0.68	0.40	0.93	-0.07	-0.97	0.40	1.07	0.50
217047_s_at	FAM13A1	ADC	2.44	0.51	0.56	0.39	-0.88	-1.24	0.61	1.30	0.48
217771_at	GOLM1	ADC	3.68	0.34	0.43	1.87	-0.06	0.13	0.70	0.80	0.57
217989_at	HSD17B11	ADC	2.06	0.56	0.67	0.09	-1.05	-0.84	0.55	0.55	0.58
218211_s_at	MLPH	ADC	5.57	0.22	0.32	0.59	-2.01	-1.85	1.03	1.38	0.91
218322_s_at	ACSL5	ADC	5.54	0.27	0.26	0.76	-1.73	-1.99	0.93	1.54	1.13
218701_at	LACTB2	ADC	4.10	0.29	0.42	1.07	-1.02	-0.68	1.02	0.97	0.73
218931_at	RAB17	ADC	2.20	0.41	0.80	0.35	-1.14	-0.57	0.59	0.96	0.46
218966_at	MYO5C	ADC	2.50	0.58	0.46	0.27	-1.01	-1.41	0.65	0.99	0.92
219105_x_at	ORC6L	ADC	0.47	1.73	1.23	0.27	1.39	1.22	0.52	0.42	0.76
219127_at	ATAD4	ADC	3.67	0.32	0.45	0.96	-1.19	-1.23	0.38	1.36	0.64
219497_s_at	BCL11A	ADC	0.15	2.15	1.99	-0.53	1.73	2.15	0.85	1.02	1.49
219498_s_at	BCL11A	ADC	0.14	2.39	1.85	-0.47	1.83	2.05	0.49	1.21	1.46

219529_at	CLIC3	ADC	3.17	0.43	0.43	-0.23	-1.62	-2.15	1.21	1.28	0.64
219543_at	PBLD	ADC	3.44	0.43	0.37	0.61	-0.64	-0.86	1.13	0.38	0.33
219580_s_at	TMC5	ADC	5.68	0.32	0.20	1.05	-1.25	-2.35	1.11	1.68	1.25
219615_s_at	KCNK5	ADC	4.00	0.28	0.44	1.10	-0.94	-0.54	1.06	0.81	0.52
219795_at	SLC6A14	ADC	11.24	0.17	0.10	0.94	-1.83	-2.99	1.91	1.56	1.32
219902_at	BHMT2	ADC	4.02	0.13	0.67	2.06	-0.80	0.11	1.36	1.80	0.22
220082_at	PPP1R14D	ADC	3.08	0.43	0.45	0.91	-0.30	-0.29	1.14	0.07	0.14
220180_at	CCDC68	ADC	2.72	0.48	0.50	-0.15	-1.51	-1.75	0.70	1.00	0.43
220187_at	STEAP4	ADC	3.24	0.46	0.38	-0.19	-1.42	-1.72	1.18	0.30	0.29
220192_x_at	SPDEF	ADC	3.17	0.43	0.44	0.89	-0.33	-0.38	1.09	0.40	0.19
220393_at	GLULD1	ADC	8.81	0.10	0.27	2.22	-0.68	0.20	2.13	1.47	0.72
221245_s_at	FZD5	ADC	2.29	0.50	0.63	0.07	-1.16	-1.27	0.77	1.11	0.60
222592_s_at	ACSL5	ADC	5.35	0.30	0.25	1.04	-1.14	-1.53	1.06	1.26	1.00
222798_at	PTER	ADC	2.62	0.54	0.46	1.13	-0.02	-0.26	0.84	0.52	0.50
222891_s_at	BCL11A	ADC	0.12	2.28	2.00	-0.63	2.17	2.28	1.00	1.23	1.38
222904_s_at	TMC5	ADC	5.69	0.30	0.21	1.04	-1.17	-1.79	1.09	1.09	0.92
223062_s_at	PSAT1	ADC	0.24	1.72	1.99	0.53	2.87	2.90	1.47	0.95	0.70
223168_at	RHOU	ADC	2.30	0.51	0.62	0.31	-0.95	-0.83	0.77	0.90	0.76
223232_s_at	CGN	ADC	5.80	0.24	0.27	0.99	-1.26	-1.50	1.04	1.25	0.44
223233_s_at	CGN	ADC	4.92	0.28	0.31	0.88	-1.39	-1.54	0.55	1.01	0.34
223315_at	NTN4	ADC	3.01	0.62	0.28	-0.23	-1.48	-2.69	0.90	1.13	1.01
223878_at	INPP4B	ADC	2.11	0.61	0.58	0.65	-0.23	-0.29	0.77	0.16	0.27
225571_at	LIFR	ADC	3.88	0.25	0.51	0.05	-1.81	-1.37	1.44	1.47	0.52
225687_at	FAM83D	ADC	0.33	1.94	1.45	0.28	1.86	1.89	0.72	0.33	0.99
225776_at	---	ADC	2.16	0.69	0.47	-0.18	-1.07	-1.73	0.43	0.88	0.34

225822_at	TMEM125	ADC	4.25	0.33	0.35	0.61	-1.56	-1.96	0.71	1.47	0.90
226213_at	ERBB3	ADC	4.08	0.42	0.27	1.46	-0.30	-1.44	0.70	1.66	0.71
226848_at	---	ADC	3.61	0.27	0.53	0.74	-1.05	-0.31	1.28	0.54	0.47
226992_at	NOSTRIN	ADC	4.37	0.31	0.35	-0.54	-2.04	-2.24	1.32	0.92	0.38
227038_at	SGMS2	ADC	3.94	0.44	0.27	-0.06	-1.76	-2.60	0.93	1.15	0.93
227081_at	DNALI1	ADC	4.44	0.23	0.45	0.15	-1.98	-1.81	1.48	1.34	0.42
227492_at	---	ADC	3.12	0.37	0.52	0.54	-1.21	-0.87	1.01	0.87	0.73
227771_at	LIFR	ADC	2.26	0.35	0.85	-0.09	-1.70	-1.08	0.80	1.36	0.43
227803_at	ENPP5	ADC	4.92	0.26	0.34	1.00	-1.60	-1.29	0.78	1.17	1.11
227808_at	DNAJC15	ADC	2.40	0.43	0.68	0.16	-1.32	-1.23	0.95	1.47	0.67
228806_at	RORC	ADC	4.52	0.32	0.32	0.85	-1.06	-1.27	1.08	0.84	0.58
229030_at	LOC644151	ADC	4.32	0.39	0.26	0.33	-1.18	-1.78	1.17	0.68	0.64
229150_at	---	ADC	7.40	0.21	0.20	0.70	-1.63	-1.79	1.40	1.09	0.98
229372_at	GOLT1A	ADC	3.88	0.28	0.47	0.79	-1.23	-0.77	1.03	0.66	0.24
229616_s_at	GRAMD2	ADC	3.88	0.36	0.37	0.28	-1.45	-1.47	1.03	0.76	0.70
230263_s_at	DOCK5	ADC	3.97	0.41	0.30	0.08	-1.46	-1.98	1.06	0.73	0.50
230349_at	XKRX	ADC	2.68	0.48	0.51	0.89	-0.25	-0.24	0.88	0.21	0.21
230831_at	FRMD5	ADC	2.75	0.38	0.62	1.01	-0.40	0.17	1.04	0.22	0.17
230875_s_at	ATP11A	ADC	4.28	0.24	0.46	0.55	-1.59	-1.01	1.41	1.07	0.80
230935_at	---	ADC	0.39	1.63	1.56	-0.35	0.98	0.84	0.25	0.89	0.49
230951_at	---	ADC	3.03	0.35	0.57	0.16	-1.52	-1.13	1.09	0.96	0.62
231022_at	---	ADC	2.74	0.43	0.55	0.19	-1.20	-1.01	1.00	0.71	0.52

232078_at	PVRL2	ADC	2.37	0.63	0.45	0.64	-0.31	-0.64	0.94	0.27	0.68
232151_at	7A5	ADC	4.17	0.37	0.31	0.56	-1.11	-1.60	1.36	1.16	0.95
232370_at	LOC254057	ADC	2.92	0.52	0.39	0.83	-0.27	-0.62	1.01	0.24	0.42
232381_s_at	DNAH5	ADC	3.27	0.25	0.64	0.56	-1.48	-0.72	1.07	1.08	0.48
233375_at	EFCAB2	ADC	2.09	0.50	0.73	0.26	-0.93	-0.61	0.62	0.72	0.35
234297_at	LOC731488 /// RGS8	ADC	3.59	0.33	0.46	0.68	-0.83	-0.55	1.28	0.60	0.31
235019_at	CPM	ADC	3.38	0.48	0.33	0.40	-1.02	-2.07	1.17	1.56	1.02
235046_at	---	ADC	3.15	0.38	0.50	0.37	-1.18	-1.06	1.09	0.98	0.54
235296_at	EIF5A2	ADC	0.33	2.55	1.06	-0.48	1.39	0.55	0.71	1.03	0.89
235299_at	---	ADC	2.88	0.32	0.66	0.85	-1.07	-0.81	0.57	1.62	0.42
235515_at	ALKBH6 /// C19orf46	ADC	2.99	0.49	0.41	1.22	-0.47	-0.53	0.65	0.70	1.00
236579_at	---	ADC	2.26	0.50	0.64	0.69	-0.49	-0.47	0.72	0.85	0.33
236795_at	---	ADC	2.05	0.47	0.79	0.16	-0.99	-0.66	0.76	0.85	0.31
236979_at	BCL2L15	ADC	2.17	0.69	0.47	0.92	0.10	-0.31	0.77	0.36	0.56
238862_at	MFSD4	ADC	2.53	0.54	0.50	0.59	-0.43	-0.55	0.92	0.13	0.21
240304_s_at	TMC5	ADC	4.90	0.32	0.27	1.17	-1.09	-2.19	1.09	2.25	1.15
241459_at	---	ADC	2.60	0.53	0.48	0.18	-1.19	-1.43	0.63	0.95	0.69
241716_at	HSPD1	ADC	0.36	1.70	1.58	-0.12	1.34	1.09	0.58	1.15	0.73
242271_at	SLC26A9	ADC	3.66	0.38	0.38	-0.01	-1.81	-1.82	0.66	0.39	0.50
242372_s_at	MFSD4	ADC	3.74	0.37	0.38	0.99	-0.48	-0.50	1.26	0.22	0.36
242722_at	LMO7	ADC	2.65	0.61	0.38	0.07	-0.84	-1.64	0.94	0.87	0.31
242931_at	LONRF3	ADC	3.10	0.25	0.70	-0.09	-2.39	-1.26	0.86	1.00	0.82
244056_at	SFTA2	ADC	8.12	0.24	0.13	0.63	-2.12	-4.28	1.98	2.33	2.28

35148_at	TJP3	ADC	2.81	0.40	0.57	0.62	-0.89	-0.56	0.95	0.85	0.71
1555867_at	GNG4	LCC	0.20	0.12	12.78	-0.19	-0.54	2.60	1.15	2.45	0.23
1557078_at	SLFN5	LCC	1.37	1.42	0.33	0.27	0.42	-1.21	0.80	0.62	0.64
1557430_at	LOC147670	LCC	0.53	0.44	3.46	-0.05	-0.36	1.05	0.20	1.28	0.10
200660_at	S100A11	LCC	1.22	1.65	0.30	0.43	0.75	-1.22	0.48	0.73	0.45
200783_s_at	STMN1	LCC	0.44	0.59	3.36	0.12	0.40	2.04	0.72	0.55	0.65
200862_at	DHCR24	LCC	1.30	1.34	0.43	-0.09	-0.04	-1.60	0.48	1.15	0.60
200872_at	S100A10	LCC	1.80	1.03	0.35	0.03	-0.10	-1.52	0.98	0.72	0.16
201292_at	TOP2A	LCC	0.51	0.85	2.29	1.29	1.76	2.90	0.98	0.46	0.91
201310_s_at	C5orf13	LCC	0.35	0.43	4.82	-0.21	-0.03	1.56	0.75	1.38	0.56
201650_at	JUP /// KRT19	LCC	0.68	3.38	0.18	0.79	2.22	-2.55	0.84	2.97	0.65
201774_s_at	NCAPD2	LCC	0.46	0.88	2.40	0.27	0.86	1.75	0.49	0.68	0.51
201798_s_at	FER1L3	LCC	2.02	0.92	0.33	0.27	-0.12	-1.50	0.83	0.86	0.45
202286_s_at	TACSTD2	LCC	1.51	1.65	0.13	0.99	1.23	-2.73	0.76	1.85	0.52
202357_s_at	C2 /// CFB	LCC	2.81	0.75	0.21	1.50	0.54	-1.40	1.06	1.75	1.01
202517_at	CRMP1	LCC	0.45	0.85	2.53	-0.47	0.00	0.83	0.23	1.07	0.68
202954_at	UBE2C	LCC	0.51	0.77	2.49	1.10	1.41	2.64	0.85	0.65	0.90
203228_at	PAFAH1B3	LCC	0.64	0.63	2.35	0.74	0.61	1.74	0.47	0.92	0.56
203234_at	UPP1	LCC	1.75	1.09	0.32	0.39	0.15	-1.36	0.86	0.89	0.68
203407_at	PPL	LCC	1.11	2.05	0.20	-0.05	0.65	-2.10	1.00	1.16	0.82
203440_at	CDH2	LCC	0.43	0.33	4.73	-0.28	-0.20	2.44	1.42	1.61	1.18
203441_s_at	CDH2	LCC	0.71	0.56	2.31	-0.04	-0.10	1.12	0.80	0.87	0.32
203702_s_at	TTLL4	LCC	0.66	0.57	2.44	0.13	-0.14	1.10	0.41	0.93	0.51
203755_at	BUB1B	LCC	0.36	0.98	2.61	0.82	1.80	2.62	0.71	0.86	0.70

203849_s_at	KIF1A	LCC	0.24	0.25	8.32	-0.13	-0.23	2.77	0.47	0.85	0.45
204026_s_at	ZWINT	LCC	0.59	0.84	2.04	0.86	1.20	2.01	0.54	0.57	0.39
204092_s_at	AURKA	LCC	0.52	0.90	2.14	0.63	1.11	1.99	0.56	0.42	0.59
204146_at	RAD51AP1	LCC	0.53	0.64	2.75	0.86	1.08	2.62	1.05	0.57	0.91
204260_at	CHGB	LCC	0.08	0.12	23.05	-0.61	-0.39	1.86	0.51	2.79	0.47
204388_s_at	MAOA	LCC	0.83	2.71	0.21	-0.66	0.46	-2.83	1.09	1.38	0.60
204389_at	MAOA	LCC	0.71	2.90	0.27	-0.74	0.46	-2.10	0.86	0.91	0.71
204519_s_at	PLLP	LCC	1.35	1.41	0.35	-0.52	-0.31	-2.23	0.82	1.18	0.55
204641_at	NEK2	LCC	0.57	0.53	2.92	1.13	1.34	3.03	1.20	0.22	0.37
204867_at	GCHFR	LCC	1.41	1.27	0.40	-0.04	-0.08	-1.46	0.50	0.76	0.58
204913_s_at	SOX11	LCC	0.18	0.09	14.75	0.03	-0.51	2.23	1.23	2.75	0.16
204914_s_at	SOX11	LCC	0.18	0.10	14.11	-0.07	-0.40	2.39	1.23	2.62	0.30
204993_at	GNAZ	LCC	0.49	0.58	3.11	-0.17	-0.06	1.43	0.59	0.84	0.58
205122_at	TMEFF1	LCC	0.23	0.49	5.76	-0.30	0.26	2.35	0.68	1.14	0.98
205123_s_at	TMEFF1	LCC	0.38	0.55	3.88	-0.21	0.03	1.27	0.26	1.31	0.45
205184_at	GNG4	LCC	0.27	0.29	7.15	-0.26	-0.32	1.85	0.66	1.85	0.51
205413_at	MPPED2	LCC	0.07	0.10	27.62	-0.60	-0.35	2.36	0.24	2.93	0.47
205445_at	PRL	LCC	0.14	0.16	14.87	-0.09	-0.10	2.00	0.21	2.58	0.22
205472_s_at	DACH1	LCC	0.18	0.25	10.07	-0.68	-0.52	1.52	0.32	2.30	0.50
205709_s_at	CDS1	LCC	1.68	1.06	0.38	0.55	0.16	-1.33	0.51	1.42	0.80
205741_s_at	DTNA	LCC	0.50	0.42	3.72	-0.38	-0.57	1.03	0.58	1.14	0.18
205751_at	SH3GL2	LCC	0.65	0.35	3.22	-0.20	-0.54	1.28	0.99	1.26	0.14
205938_at	PPM1E	LCC	0.27	0.29	7.12	-0.14	-0.25	1.53	0.32	1.72	0.23
206051_at	ELAVL4	LCC	0.54	0.59	2.85	-0.21	-0.22	0.93	0.18	1.21	0.22
206135_at	ST18	LCC	0.07	0.06	34.29	-0.12	-0.34	2.42	0.53	2.87	0.15

206290_s_at	RGS7	LCC	0.52	0.50	3.28	-0.20	-0.30	0.99	0.40	1.40	0.12
206502_s_at	INSM1	LCC	0.12	0.07	22.48	-0.06	-0.68	2.65	0.97	2.87	0.25
207120_at	ZNF667	LCC	0.43	0.45	3.99	-0.23	-0.27	1.23	0.39	1.48	0.19
207401_at	PROX1	LCC	0.32	0.24	6.84	-0.04	-0.37	1.70	0.78	1.86	0.28
207625_s_at	CBFA2T2	LCC	0.58	0.57	2.71	-0.12	-0.21	1.29	0.42	0.34	0.36
207828_s_at	CENPF	LCC	0.45	0.78	2.71	0.92	1.62	2.85	1.03	0.48	0.63
207981_s_at	ESRRG	LCC	0.14	0.12	16.24	-0.32	-0.52	2.97	0.91	1.95	0.32
208079_s_at	AURKA	LCC	0.53	0.92	2.07	0.74	1.30	2.08	0.74	0.52	0.53
208510_s_at	PPARG	LCC	2.11	0.90	0.31	-0.12	-0.69	-1.86	0.81	0.50	0.71
209172_s_at	CENPF	LCC	0.49	0.57	3.12	0.43	0.72	2.36	0.98	0.65	0.71
209173_at	AGR2	LCC	4.43	0.57	0.08	1.47	0.02	-3.11	2.07	2.23	1.83
209211_at	KLF5	LCC	0.81	2.92	0.17	0.25	1.45	-1.70	0.95	1.00	0.67
209270_at	LAMB3	LCC	1.64	1.46	0.16	0.57	0.68	-2.16	0.98	1.04	0.69
209373_at	MALL	LCC	1.78	1.17	0.26	0.22	0.13	-2.23	1.25	1.41	1.00
209550_at	NDN	LCC	0.48	0.87	2.37	-1.21	-0.66	0.31	0.79	0.75	0.87
209642_at	BUB1	LCC	0.46	0.71	2.83	0.94	1.42	2.79	1.03	0.50	0.70
209679_s_at	LOC57228	LCC	1.31	1.51	0.32	0.48	0.75	-1.00	0.70	0.45	0.40
209966_x_at	ESRRG	LCC	0.14	0.13	16.13	-0.08	-0.22	2.71	0.71	2.23	0.35
210036_s_at	KCNH2	LCC	0.43	0.31	4.81	-0.23	-0.42	1.68	0.99	1.31	0.27
210052_s_at	TPX2	LCC	0.45	0.79	2.68	1.06	1.61	2.72	0.78	0.77	0.63
210367_s_at	PTGES	LCC	1.68	1.22	0.27	0.99	0.75	-0.89	0.81	0.74	0.98
210683_at	NRTN	LCC	0.49	0.63	2.99	-0.25	-0.02	1.18	0.44	1.00	0.23
211026_s_at	MGLL	LCC	1.93	1.11	0.24	-0.23	-0.57	-2.53	0.85	0.96	0.76
211080_s_at	NEK2	LCC	0.57	0.80	2.17	0.16	0.41	1.29	0.46	0.75	0.53
212268_at	SERPINB1	LCC	2.21	0.96	0.23	0.25	-0.18	-2.29	0.91	1.45	0.54

212531_at	LCN2	LCC	2.37	1.05	0.11	1.50	0.77	-1.78	1.32	1.35	1.67
212607_at	AKT3	LCC	0.53	0.53	3.09	-1.01	-1.03	0.60	1.10	1.05	0.77
212741_at	MAOA	LCC	1.44	1.47	0.26	-0.62	-0.45	-3.46	0.94	1.70	0.66
213056_at	FRMD4B	LCC	1.69	1.03	0.40	-0.08	-0.38	-1.60	0.65	0.90	0.51
213245_at	ADCY1	LCC	0.47	0.33	4.37	-0.11	-0.48	1.50	0.69	1.36	0.21
213479_at	NPTX2	LCC	0.12	0.10	19.66	-0.46	-0.41	3.00	1.18	2.63	0.60
213492_at	COL2A1	LCC	0.19	0.24	9.72	-0.28	-0.20	2.09	0.21	2.04	0.21
213506_at	F2RL1	LCC	1.30	1.90	0.13	0.87	1.47	-1.52	1.19	0.51	0.81
213712_at	ELOVL2	LCC	0.35	0.41	4.95	-0.22	-0.11	1.89	0.61	1.23	0.50
214097_at	RPS21	LCC	0.50	0.76	2.55	-0.04	0.34	1.38	0.68	0.87	0.66
214162_at	LOC284244	LCC	0.59	0.66	2.45	-0.25	-0.19	0.76	0.22	1.00	0.18
214933_at	CACNA1A	LCC	0.28	0.48	5.21	-0.35	0.02	1.54	0.39	1.73	0.71
215143_at	DPY19L2P2	LCC	0.28	0.24	7.49	-0.17	-0.51	1.68	0.34	1.91	0.11
215767_at	ZNF804A	LCC	0.62	0.61	2.47	-0.24	-0.35	0.68	0.23	1.04	0.21
217404_s_at	COL2A1	LCC	0.42	0.48	3.94	-0.19	-0.16	1.26	0.18	1.43	0.20
217728_at	S100A6	LCC	1.91	0.99	0.32	0.26	-0.10	-1.78	0.67	1.26	0.33
218186_at	RAB25	LCC	1.84	0.99	0.36	1.01	0.65	-2.01	0.54	2.50	0.22
218397_at	FANCL	LCC	0.61	0.61	2.51	0.36	0.23	1.52	0.65	0.94	0.77
218677_at	S100A14	LCC	1.99	1.19	0.16	0.76	0.26	-2.25	1.02	1.48	1.47
218829_s_at	CHD7	LCC	0.64	0.47	2.85	0.12	-0.30	1.45	0.61	0.65	0.59
218888_s_at	NETO2	LCC	0.61	0.72	2.22	0.23	0.45	1.70	1.02	0.41	0.76
218959_at	HOXC10	LCC	0.25	0.22	8.36	0.24	0.20	2.80	1.29	1.98	1.03
219014_at	PLAC8	LCC	1.13	2.17	0.14	0.18	0.49	-2.66	1.17	1.54	1.56
219170_at	FSD1	LCC	0.68	0.70	2.08	-0.02	-0.05	0.86	0.39	0.83	0.34
219476_at	C1orf116	LCC	4.29	0.46	0.19	0.15	-1.02	-3.16	1.41	1.73	0.62

219537_x_at	DLL3	LCC	0.60	0.48	2.99	0.13	-0.17	1.26	0.47	1.24	0.26
219740_at	VASH2	LCC	0.46	0.47	3.73	0.05	-0.11	1.69	0.45	1.01	0.61
220167_s_at	LOC729355 /// TP53TG3	LCC	0.45	0.46	3.88	-0.21	-0.25	1.35	0.47	1.22	0.24
221591_s_at	FAM64A	LCC	0.66	0.59	2.40	0.52	0.40	1.75	0.62	0.51	0.38
221959_at	FAM110B	LCC	0.53	0.76	2.41	-0.54	-0.23	0.66	0.32	0.91	0.33
222549_at	CLDN1	LCC	0.70	3.62	0.11	0.37	1.51	-1.96	1.16	0.96	1.42
222771_s_at	MYEF2	LCC	0.74	0.36	2.88	0.06	-0.72	1.36	1.03	0.91	0.76
222797_at	DPYSL5	LCC	0.40	0.51	3.94	-0.21	-0.04	1.45	0.29	1.08	0.16
223307_at	CDCA3	LCC	0.45	0.87	2.50	0.73	1.51	2.48	0.88	0.52	0.44
223374_s_at	B3GALNT1	LCC	0.42	0.51	3.77	-0.87	-0.63	1.02	0.88	1.14	0.65
223381_at	NUF2	LCC	0.55	0.70	2.49	1.15	1.43	2.83	1.03	0.45	0.95
223523_at	TMEM108	LCC	0.42	0.25	5.41	-0.93	-1.38	0.72	1.00	1.80	0.30
223540_at	PVRL4	LCC	1.39	1.37	0.34	0.93	1.08	-0.51	0.72	0.56	0.48
223591_at	RNF135	LCC	1.52	1.07	0.47	0.33	0.12	-0.94	0.55	0.76	0.39
223614_at	C8orf57	LCC	0.49	0.53	3.29	-0.20	-0.08	1.31	0.76	1.17	0.46
223627_at	MEX3B	LCC	0.44	0.55	3.54	-0.11	0.00	1.23	0.15	1.24	0.28
224414_s_at	CARD6	LCC	1.38	1.22	0.44	-0.24	-0.31	-1.51	0.55	0.78	0.64
224521_s_at	CCDC77	LCC	0.57	0.67	2.51	-0.05	0.09	1.31	0.51	0.56	0.34
224650_at	MAL2	LCC	1.74	1.16	0.28	0.68	0.47	-2.30	0.74	2.70	0.60
225482_at	KIF1A	LCC	0.35	0.33	5.62	-0.19	-0.40	2.25	0.61	0.58	0.53
225613_at	LOC100128443 /// MAST4	LCC	1.02	1.97	0.31	-0.37	0.10	-1.85	0.57	0.65	0.87
225745_at	LRP6	LCC	0.69	0.52	2.49	-0.27	-0.68	0.99	0.82	0.57	0.82

225834_at	FAM72A /// FAM72B /// GCUD2	LCC	0.46	0.78	2.68	0.60	0.98	2.24	0.80	0.88	0.93
226147_s_at	PIGR	LCC	2.02	0.85	0.38	0.32	-0.11	-3.39	1.26	3.19	0.83
226269_at	GDAP1	LCC	0.42	0.43	4.22	-0.29	-0.22	1.53	0.95	1.31	0.42
226285_at	CAPRIN1	LCC	0.52	0.90	2.14	-0.27	0.24	0.93	0.43	0.78	0.38
226346_at	MEX3A	LCC	0.49	0.40	3.85	0.34	0.11	1.87	0.71	1.18	0.21
226560_at	---	LCC	2.24	0.91	0.25	0.82	0.21	-1.77	0.73	1.58	0.67
226610_at	PRR6	LCC	0.32	0.66	3.91	-0.19	0.47	2.36	1.05	0.69	1.18
226612_at	FLJ25076	LCC	0.16	0.26	10.60	-0.66	-0.24	1.70	0.48	2.05	0.56
226618_at	FLJ25076	LCC	0.51	0.67	2.75	-0.36	-0.15	0.73	0.19	1.14	0.21
226809_at	FLJ30428	LCC	0.16	0.16	13.14	0.13	0.10	3.09	1.17	2.33	0.73
226933_s_at	ID4	LCC	0.49	0.72	2.66	-0.83	-0.25	0.83	0.92	0.88	0.46
226960_at	CXCL17	LCC	3.14	0.62	0.24	0.63	-0.25	-3.37	1.61	2.29	0.82
227134_at	SYTL1	LCC	1.22	1.65	0.30	0.28	0.75	-1.43	0.94	1.23	0.51
227188_at	C21orf63	LCC	1.45	1.48	0.24	-0.09	0.17	-2.05	1.04	0.93	0.53
227230_s_at	KIAA1211	LCC	0.35	0.24	6.47	-0.03	-0.52	1.89	0.83	1.61	0.43
227512_at	MEX3A	LCC	0.60	0.44	3.10	0.46	0.07	1.83	0.61	0.84	0.31
227525_at	GLCCI1	LCC	0.86	0.48	2.13	-0.44	-0.93	0.51	0.78	0.61	0.32
227545_at	---	LCC	0.57	0.43	3.30	0.56	0.37	2.21	1.13	1.06	0.63
227593_at	FLJ37453	LCC	0.39	0.70	3.26	-0.41	0.01	1.28	0.59	1.13	0.88
227612_at	ELAVL3	LCC	0.08	0.11	23.62	-0.35	-0.23	2.50	0.19	2.92	0.12
227647_at	KCNE3	LCC	1.19	1.62	0.34	0.22	0.43	-1.28	0.58	0.85	0.80
227746_at	ELAVL1	LCC	0.40	0.73	3.12	-0.52	-0.16	1.17	0.29	0.83	0.83

227998_at	S100A16	LCC	1.36	1.64	0.22	0.49	0.92	-1.60	0.84	0.85	0.23
228107_at	LOC100127983	LCC	0.39	0.49	4.10	-0.59	-0.47	0.97	0.22	1.43	0.18
228241_at	AGR3	LCC	2.72	0.86	0.15	0.15	-1.06	-4.30	1.43	2.34	2.48
228245_s_at	LOC100132881 /// LOC728715 /// OVOS /// OVOS2	LCC	0.33	0.41	5.19	-0.35	-0.31	1.97	1.16	1.25	1.34
228291_s_at	C20orf19	LCC	0.56	0.66	2.57	-0.38	-0.37	0.77	0.31	0.98	0.60
228547_at	NRXN1	LCC	0.25	0.30	7.52	-0.41	-0.37	1.28	0.13	1.88	0.15
228708_at	RAB27B	LCC	3.36	0.59	0.23	0.46	-0.65	-2.40	1.02	1.78	0.84
228882_at	TUB	LCC	0.32	0.73	3.65	-0.81	-0.13	1.22	0.45	1.01	0.79
228906_at	TET1	LCC	0.39	0.58	3.67	0.00	0.32	1.64	0.74	1.36	0.79
229057_at	SCN2A	LCC	0.33	0.34	5.68	-0.14	-0.19	1.45	0.49	1.84	0.14
229349_at	LIN28B	LCC	0.16	0.18	12.92	-0.28	-0.31	2.01	0.21	2.30	0.12
229442_at	C18orf54	LCC	0.40	0.34	4.94	0.01	-0.21	1.89	0.62	1.32	0.25
229610_at	CKAP2L	LCC	0.53	0.73	2.49	0.47	0.61	1.84	0.53	0.69	0.79
229921_at	---	LCC	0.64	0.48	2.82	0.02	-0.32	1.14	0.48	1.03	0.19
230193_at	WDR66	LCC	0.96	2.84	0.06	1.05	2.83	-1.53	2.15	0.60	0.85
230407_at	---	LCC	0.61	0.56	2.65	-0.25	-0.42	1.14	0.55	0.43	0.61
230664_at	H2BFM /// H2BFXP	LCC	0.38	0.41	4.68	-0.28	-0.30	1.38	0.36	1.55	0.34
230861_at	DKFZP434L187	LCC	0.42	0.53	3.72	-0.17	-0.04	1.10	0.10	1.46	0.16
231936_at	HOXC9	LCC	0.24	0.21	9.06	0.16	0.04	2.85	1.00	1.87	0.68
232038_at	C6orf170	LCC	0.53	0.73	2.51	-0.15	0.02	1.03	0.40	0.96	0.67
232282_at	WNK3	LCC	0.40	0.23	5.93	0.17	-0.33	1.87	0.99	1.86	0.27

233300_at	---	LCC	0.59	0.63	2.51	0.05	0.10	1.36	0.68	0.69	0.51
233536_at	ASXL3	LCC	0.65	0.61	2.37	-0.28	-0.41	0.65	0.29	1.05	0.18
234192_s_at	GKAP1	LCC	0.67	0.37	3.06	-0.17	-0.78	1.11	0.74	0.94	0.31
235343_at	VASH2	LCC	0.36	0.23	6.37	0.23	-0.39	1.97	0.67	1.68	0.43
235352_at	---	LCC	1.69	1.04	0.39	0.07	-0.28	-1.64	0.56	1.28	0.62
235759_at	---	LCC	0.28	0.34	6.38	-0.92	-0.82	0.91	0.35	1.95	0.22
235762_at	TAS2R14	LCC	0.52	0.54	3.10	-0.20	-0.35	1.04	0.34	1.02	0.60
236236_at	---	LCC	0.62	0.30	3.62	0.22	-0.43	1.62	0.99	1.31	0.37
236302_at	PPM1E	LCC	0.34	0.05	9.51	0.31	-0.77	2.69	1.92	2.42	0.25
236635_at	ZNF667	LCC	0.44	0.37	4.36	-0.24	-0.48	1.33	0.62	1.54	0.30
236641_at	KIF14	LCC	0.50	0.62	2.95	0.65	0.90	2.50	0.96	0.46	0.76
237116_at	LOC646903	LCC	0.54	0.53	3.04	-0.06	-0.15	1.20	0.31	1.08	0.15
237248_at	PDE11A	LCC	0.35	0.44	4.78	-0.10	0.02	1.52	0.14	1.63	0.24
237305_at	---	LCC	0.43	0.47	3.89	-0.11	-0.11	1.35	0.26	1.37	0.18
238073_at	ELAVL4	LCC	0.37	0.34	5.32	-0.10	-0.36	1.41	0.16	1.72	0.13
238763_at	RBM20	LCC	0.41	0.51	3.87	-0.35	-0.22	1.11	0.11	1.40	0.14
238850_at	LOC645323	LCC	0.44	0.53	3.61	-0.18	-0.08	1.05	0.14	1.43	0.22
239466_at	LOC344595	LCC	0.41	0.52	3.79	-0.39	-0.27	0.97	0.48	1.43	0.52
240084_at	CBX2	LCC	0.61	0.76	2.15	0.02	0.17	1.16	0.52	0.60	0.58
242301_at	CBLN2	LCC	0.32	0.44	5.03	-0.20	-0.07	1.53	0.17	1.65	0.50
242822_at	MGC39584	LCC	0.32	0.35	5.77	-0.26	-0.28	1.68	0.20	1.55	0.17
243932_at	---	LCC	0.54	0.61	2.77	-0.14	-0.10	0.92	0.11	1.24	0.12
244660_at	ELAVL1	LCC	0.33	0.43	5.08	-0.38	-0.23	1.29	0.24	1.42	0.32
244780_at	SGPP2	LCC	1.71	1.25	0.24	1.15	1.09	-1.35	1.02	1.79	0.75
39248_at	AQP3	LCC	2.34	0.99	0.16	0.24	-0.39	-3.23	1.29	1.97	1.50

40016_g_at	LOC100128443 /// MAST4	LCC	1.05	1.83	0.35	-0.23	0.28	-1.61	0.41	0.69	0.42
1552291_at	PIGX	SCC	0.42	2.90	0.70	0.02	1.42	0.22	0.51	1.13	1.29
1552477_a_at	IRF6	SCC	0.35	4.88	0.34	0.10	2.24	-0.35	0.60	1.09	0.76
1553605_a_at	ABCA13	SCC	0.16	13.57	0.09	-0.26	2.17	-0.86	1.01	0.19	2.40
1554018_at	GPNMB	SCC	0.46	3.86	0.35	-0.35	1.79	-0.40	1.67	0.73	0.87
1554556_a_at	ATP11B	SCC	0.31	4.40	0.51	-0.02	1.84	0.35	0.26	0.44	1.18
1554667_s_at	METTL8	SCC	0.41	2.63	0.83	0.06	1.55	0.75	0.51	0.47	0.96
1555007_s_at	WDR66	SCC	0.41	3.98	0.40	-0.28	1.08	-0.46	0.54	0.43	1.59
1556793_a_at	FAM83C	SCC	0.13	12.03	0.20	-0.50	2.75	-0.23	0.09	0.14	1.19
1559606_at	GBP6	SCC	0.09	19.12	0.12	-0.51	3.45	-0.56	0.11	0.09	0.97
1559607_s_at	GBP6	SCC	0.05	38.23	0.06	-0.74	4.19	-0.84	0.39	0.15	1.09
1563111_a_at	PIGX	SCC	0.38	3.54	0.57	-0.04	1.85	0.13	0.60	0.87	0.78
1564064_a_at	ATP11B	SCC	0.33	4.34	0.48	-0.10	1.99	0.23	0.69	0.57	0.96
1568932_at	---	SCC	0.32	4.98	0.37	-0.36	1.32	-0.42	0.12	0.13	1.29
201249_at	SLC2A1	SCC	0.41	3.31	0.57	-0.25	1.45	-0.14	0.34	0.67	0.68
201250_s_at	SLC2A1	SCC	0.30	4.43	0.52	0.39	3.10	0.30	1.13	1.86	0.42
202219_at	SLC6A8	SCC	0.25	6.88	0.28	-0.31	3.08	0.24	1.39	0.94	0.97
202504_at	TRIM29	SCC	0.17	13.20	0.10	-0.65	3.82	-0.85	1.90	1.37	0.79
202597_at	IRF6	SCC	0.48	2.99	0.56	0.04	1.64	-0.19	0.40	1.15	0.38
202804_at	ABCC1	SCC	0.49	2.42	0.76	-0.03	1.42	0.14	0.56	0.91	0.45
202912_at	ADM	SCC	0.36	3.42	0.63	-0.91	1.43	-0.69	1.11	1.86	0.55
203074_at	ANXA8 /// ANXA8L1 /// ANXA8L2	SCC	0.26	7.92	0.18	-1.61	1.52	-1.82	1.31	0.77	1.12

203797_at	VSNL1	SCC	0.13	7.01	0.51	-1.11	2.79	0.10	1.06	1.73	0.70	
204060_s_at	PRKX /// PRKY	SCC	0.49	2.35	0.88	-0.09	0.65	0.26	0.66	0.97	1.00	
204061_at	PRKX	SCC	0.35	2.40	1.07	-0.23	1.37	0.50	0.63	1.20	1.02	
204136_at	COL7A1	SCC	0.32	5.03	0.36	-0.36	1.62	-0.48	0.49	0.66	1.09	
204203_at	CEBPG	SCC	0.41	2.98	0.69	-0.11	1.50	0.20	0.43	0.77	0.75	
204268_at	S100A2	SCC	0.19	14.45	0.03	0.16	4.92	-1.46	1.92	1.42	0.77	
204455_at	DST	SCC	0.07	30.69	0.06	-0.45	4.17	-1.29	1.07	1.23	1.49	
204469_at	PTPRZ1	SCC	0.09	26.09	0.04	-0.53	4.44	-1.00	1.48	0.69	1.06	
204614_at	SERPINB2	SCC	0.15	10.02	0.26	-0.94	1.74	-0.53	0.18	0.28	1.78	
204653_at	TFAP2A	SCC	0.33	3.77	1.98	0.43	2.17	1.27	1.39	2.44	0.87	
204734_at	KRT15	SCC	0.09	23.51	0.06	-0.19	4.63	-0.98	1.63	1.44	1.06	
204952_at	LYPD3	SCC	0.26	7.36	0.21	0.05	2.48	-0.32	1.03	0.87	1.54	
204971_at	CSTA	SCC	0.18	11.45	0.13	-0.76	2.62	-2.40	0.98	2.53	0.83	
205014_at	FGFBP1	SCC	0.09	24.37	0.06	-0.86	3.48	-1.41	1.06	0.38	1.31	
205064_at	SPRR1B	SCC	0.06	42.61	0.02	-0.52	4.65	-1.07	1.55	0.20	1.99	
205157_s_at	KRT17	SCC	0.06	25.58	0.11	-0.91	4.04	-0.74	1.19	1.52	1.17	
205379_at	CBR3	SCC	0.43	4.29	0.31	-0.02	1.92	-0.67	0.78	0.90	0.71	
205623_at	ALDH3A1	SCC	0.06	28.91	0.10	-0.78	4.13	-0.79	0.88	1.38	1.36	
205724_at	PKP1	SCC	0.37	4.14	0.43	-0.10	1.82	-0.09	0.46	0.39	0.81	
205816_at	ITGB8	SCC	0.47	3.41	0.45	0.01	1.34	-0.01	0.80	0.53	1.26	
206122_at	SOX15	SCC	0.51	3.01	0.50	-0.13	1.44	-0.09	0.88	0.59	0.91	
206156_at	GJB5	SCC	0.31	5.67	0.29	-0.15	2.14	-0.44	0.48	0.41	0.79	
206164_at	CLCA2	SCC	0.07	26.84	0.09	-0.94	3.30	-0.96	0.15	0.16	1.52	
206165_s_at	CLCA2	SCC	0.02	92.80	0.03	-1.38	4.79	-1.23	0.15	0.46	1.58	
206166_s_at	CLCA2	SCC	0.05	37.97	0.07	-0.97	3.87	-0.82	0.12	0.20	1.72	

206307_s_at	FOXD1	SCC	0.64	2.74	0.81	0.33	1.67	0.79	1.57	1.51	1.14
206912_at	FOXE1	SCC	0.04	38.68	0.08	-0.58	4.20	-0.35	0.43	1.03	1.85
207469_s_at	PIR	SCC	0.31	2.79	0.99	-0.53	1.59	-0.34	1.02	2.09	0.84
207602_at	TMPRSS11D	SCC	0.09	19.64	0.11	-0.26	3.48	-0.29	0.40	0.16	1.61
207675_x_at	ARTN	SCC	0.26	4.99	0.50	-0.13	2.17	0.08	0.45	1.10	1.17
207935_s_at	KRT13	SCC	0.01	152.16	0.02	-1.12	5.17	-1.14	0.31	0.16	2.03
208539_x_at	SPRR2B	SCC	0.06	29.75	0.08	-0.53	3.59	-0.56	0.27	0.47	1.89
208836_at	ATP1B3	SCC	0.40	3.40	0.56	-0.62	1.28	-0.41	0.67	0.63	0.49
209125_at	KRT6A	SCC	0.04	64.32	0.01	-0.96	6.83	-1.49	2.13	0.78	0.68
209126_x_at	KRT6B	SCC	0.09	25.31	0.05	-0.50	4.23	-1.07	1.00	0.28	0.61
209203_s_at	BICD2	SCC	0.33	3.87	0.57	-0.58	1.33	-0.17	0.30	0.53	0.92
209260_at	SFN	SCC	0.43	4.97	0.20	0.30	2.39	-1.26	0.85	1.75	0.67
209351_at	KRT14	SCC	0.01	220.32	0.01	-1.07	5.24	-0.91	0.50	0.38	2.50
209380_s_at	ABCC5	SCC	0.24	6.74	0.30	-0.23	2.56	-0.05	0.80	0.22	0.72
209652_s_at	PGF	SCC	0.51	3.01	0.50	-0.14	1.26	-0.26	0.48	0.33	0.82
209719_x_at	SERPINB3	SCC	0.15	10.97	0.21	-0.87	2.83	-0.60	1.45	1.42	1.74
209800_at	KRT16	SCC	0.06	31.35	0.07	-0.51	3.50	-0.70	0.66	0.79	1.95
210020_x_at	CALML3	SCC	0.08	20.73	0.12	-0.49	4.37	-0.19	1.23	1.35	1.22
210180_s_at	SFRS10	SCC	0.42	2.73	0.77	-0.56	1.14	0.04	0.70	0.61	0.70
210237_at	ARTN	SCC	0.34	3.43	0.68	0.00	1.89	0.45	0.61	0.96	0.99
210505_at	ADH7	SCC	0.07	25.28	0.09	-0.72	3.50	-0.75	0.15	0.17	1.27
210854_x_at	SLC6A8	SCC	0.19	6.26	0.47	-0.19	2.94	0.78	0.92	0.97	0.92
211002_s_at	TRIM29	SCC	0.36	5.32	0.25	-0.17	2.13	-0.54	0.98	0.44	0.94
211194_s_at	TP63	SCC	0.13	13.33	0.16	-0.48	2.74	-0.57	0.16	0.45	1.31
211361_s_at	SERPINB13	SCC	0.05	35.59	0.07	-0.68	3.26	-0.58	0.09	0.13	2.30

212236_x_at	KRT17	SCC	0.08	18.33	0.15	-0.71	3.88	-0.55	1.22	1.51	0.97
212702_s_at	BICD2	SCC	0.35	3.77	0.57	-0.59	1.31	-0.14	0.43	0.16	0.90
212810_s_at	SLC1A4	SCC	0.45	2.58	0.77	-0.03	1.59	0.33	0.82	1.28	0.65
213110_s_at	COL4A5	SCC	0.24	4.36	0.67	-1.37	1.12	-1.03	0.72	1.76	1.06
213139_at	SNAI2	SCC	0.39	3.61	0.54	-0.91	1.38	-0.68	1.27	1.33	0.62
213154_s_at	BICD2	SCC	0.35	3.74	0.56	-0.62	1.20	-0.21	0.41	0.21	0.92
213680_at	KRT6B	SCC	0.01	124.00	0.03	-1.21	5.99	-0.82	0.85	1.36	1.15
213707_s_at	DLX5	SCC	0.21	3.38	1.01	-0.46	1.67	0.60	0.20	1.34	1.27
213796_at	SPRR1A	SCC	0.01	232.64	0.01	-0.91	5.46	-1.12	0.54	0.17	2.53
213843_x_at	SLC6A8	SCC	0.17	6.91	0.42	-0.31	3.06	0.67	1.03	1.02	0.87
213992_at	COL4A6	SCC	0.20	8.33	0.25	-0.91	1.41	-0.90	0.23	0.36	1.68
214549_x_at	SPRR1A	SCC	0.11	17.60	0.10	-0.27	2.97	-0.56	0.66	0.23	1.84
214580_x_at	KRT6A /// KRT6B /// KRT6C	SCC	0.07	35.21	0.03	-0.80	4.95	-1.17	1.40	0.33	0.31
215812_s_at	LOC653562 /// SLC6A10P /// SLC6A8	SCC	0.23	5.76	0.44	-0.21	2.62	0.44	0.87	0.95	0.91
216918_s_at	DST	SCC	0.23	7.85	0.23	-0.35	2.12	-0.64	0.36	0.32	1.33
217272_s_at	SERPINB13	SCC	0.02	91.29	0.03	-1.00	4.74	-0.80	0.14	0.19	1.83
217312_s_at	COL7A1	SCC	0.46	3.35	0.48	-0.27	1.28	-0.38	0.46	0.40	0.79
217528_at	CLCA2	SCC	0.02	86.63	0.03	-1.36	4.47	-1.27	0.12	0.30	2.05
217744_s_at	PERP	SCC	0.41	4.61	0.29	0.44	2.47	-0.94	0.47	2.24	0.33
218847_at	IGF2BP2	SCC	0.57	3.21	0.35	-0.04	1.45	-1.00	0.78	1.24	0.71
218990_s_at	SPRR3	SCC	0.03	88.82	0.01	-0.35	5.02	-1.01	1.32	0.15	2.67
219412_at	RAB38	SCC	0.64	3.46	0.59	0.19	1.07	-0.58	1.30	1.56	0.85

219522_at	FJX1	SCC	0.54	2.68	1.07	-0.41	0.95	0.08	0.63	0.96	0.76
219936_s_at	GPR87	SCC	0.81	3.31	0.07	0.41	2.93	-1.35	2.16	0.38	0.77
219995_s_at	ZNF750	SCC	0.31	7.69	0.10	0.04	2.53	-0.98	1.49	0.45	1.60
220230_s_at	CYB5R2	SCC	0.23	6.20	0.37	-0.57	2.13	-0.41	0.75	1.04	1.02
220658_s_at	ARNTL2	SCC	0.46	3.30	0.49	0.41	2.83	0.53	1.73	1.53	0.69
221203_s_at	YEATS2	SCC	0.41	3.00	0.68	-0.20	1.38	0.35	0.79	0.54	1.14
221291_at	ULBP2	SCC	0.66	2.23	0.58	-0.16	0.92	-0.41	0.37	0.43	0.38
221795_at	NTRK2	SCC	0.03	59.38	0.05	-1.89	3.94	-1.36	0.53	0.73	1.18
221796_at	NTRK2	SCC	0.03	53.35	0.05	-1.76	3.59	-1.62	0.55	0.62	1.38
221854_at	PKP1	SCC	0.06	21.34	0.15	-0.70	4.80	-0.01	1.34	1.79	0.66
222392_x_at	PERP	SCC	0.57	2.96	0.42	0.30	1.77	-0.82	0.62	1.96	0.38
222457_s_at	LIMA1	SCC	0.59	2.58	0.53	-0.18	0.93	-0.41	0.50	0.51	0.99
222634_s_at	TBL1XR1	SCC	0.47	2.74	0.67	0.08	1.77	0.50	0.87	0.50	0.60
222892_s_at	TMEM40	SCC	0.98	2.87	0.33	0.01	1.07	-0.49	1.32	0.53	0.55
223278_at	GJB2	SCC	0.12	13.34	0.19	0.23	3.84	-0.10	1.17	2.10	1.60
223586_at	ARNTL2	SCC	0.70	2.62	0.65	0.40	1.50	0.21	1.34	1.40	0.65
223704_s_at	DMRT2	SCC	0.26	6.73	0.28	-0.17	2.31	-0.24	0.76	0.61	1.27
223832_s_at	CAPNS2	SCC	0.20	9.54	0.17	-0.34	2.27	-0.74	0.49	0.13	1.27
224204_x_at	ARNTL2	SCC	0.53	2.68	0.59	0.43	2.35	0.54	1.39	1.34	0.61
224458_at	C9orf125	SCC	0.33	4.26	0.49	-0.83	1.44	-0.64	0.93	0.98	0.81
224950_at	PTGFRN	SCC	0.56	2.72	0.52	0.29	1.57	-0.06	0.31	0.77	0.60
225464_at	FRMD6	SCC	0.34	4.49	0.43	-0.71	1.62	-0.62	0.76	0.74	0.34
225481_at	FRMD6	SCC	0.38	4.15	0.43	-0.66	1.50	-0.61	0.99	0.73	0.62
226189_at	ITGB8	SCC	0.59	2.76	0.43	0.13	1.12	-0.66	1.43	2.01	1.30
226363_at	ABCC5	SCC	0.29	5.53	0.35	0.06	2.46	0.13	0.80	0.73	0.95

226464_at	C3orf58	SCC	0.41	3.29	0.58	-0.84	1.04	-0.46	0.84	0.43	0.62
226499_at	NRARP	SCC	0.35	3.52	0.62	-0.50	1.39	0.07	0.70	0.63	1.05
226755_at	LOC642587	SCC	0.40	3.61	0.51	-0.46	1.34	-0.34	0.34	0.34	0.70
226817_at	DSC2	SCC	0.56	2.85	0.71	-0.03	0.93	-0.35	0.55	1.41	1.12
226885_at	---	SCC	0.43	2.90	0.67	-0.60	0.93	-0.22	0.79	0.76	1.05
226907_at	PPP1R14C	SCC	0.49	3.77	0.33	-0.83	1.03	-1.40	1.19	1.17	1.04
227174_at	WDR72	SCC	0.72	2.37	0.29	0.31	2.87	0.71	3.44	1.43	1.16
227249_at	NDE1	SCC	0.59	2.19	1.00	0.15	1.24	0.68	0.79	0.71	0.80
227875_at	KLHL13	SCC	0.39	5.09	0.24	-0.50	2.69	-0.40	1.66	1.19	1.09
228286_at	GEN1	SCC	0.51	2.27	1.21	0.07	0.90	0.71	0.54	0.94	1.22
228575_at	IL20RB	SCC	0.32	6.23	0.22	-0.31	1.99	-0.46	1.07	0.24	1.62
229290_at	DAPL1	SCC	0.02	78.79	0.04	-1.18	4.63	-0.81	0.48	0.72	2.02
229296_at	LOC100128501	SCC	0.19	2.22	4.62	-0.37	0.81	0.77	0.27	2.35	0.69
229761_at	LOC440173	SCC	0.43	3.12	0.59	-0.31	1.33	-0.10	0.50	0.56	0.77
229764_at	TPRG1	SCC	0.15	14.97	0.09	-0.89	2.11	-1.78	0.90	0.47	1.80
229900_at	CD109	SCC	0.58	2.82	0.46	0.02	1.33	-0.28	0.81	0.65	0.99
230229_at	DLG1	SCC	0.48	2.52	1.14	-0.14	1.22	0.55	0.98	1.14	0.61
230464_at	S1PR5	SCC	0.27	4.76	0.50	-0.66	1.76	-0.30	0.28	0.71	0.49
230769_at	DENND2C	SCC	0.30	5.30	0.36	-0.77	1.53	-0.78	0.21	0.27	0.63
231183_s_at	JAG1	SCC	0.17	3.49	1.09	-1.43	1.43	-0.05	0.81	1.69	0.95
231331_at	---	SCC	0.04	40.80	0.07	-0.68	4.45	-0.31	0.58	0.93	1.53
231771_at	GJB6	SCC	0.02	102.57	0.02	-0.93	5.12	-1.20	0.65	0.29	1.53
231867_at	ODZ2	SCC	0.08	12.01	0.33	-0.90	2.67	-0.07	0.15	1.42	2.05
231928_at	HES2	SCC	0.54	2.06	1.19	-0.31	0.59	0.20	0.55	0.94	0.42
232082_x_at	SPRR3	SCC	0.02	119.71	0.02	-0.77	5.07	-0.95	0.44	0.37	2.24

232116_at	GRHL3	SCC	0.32	5.04	0.37	-0.12	2.05	-0.16	0.43	0.49	0.87
232202_at	---	SCC	0.13	11.82	0.22	-0.35	3.73	-0.10	1.24	1.60	1.03
232406_at	---	SCC	0.33	3.37	0.72	-0.59	1.17	-0.22	0.46	1.18	1.03
233479_at	---	SCC	0.30	3.49	0.74	-0.35	1.49	0.37	0.32	0.78	1.16
233869_x_at	---	SCC	0.45	2.83	1.25	-0.12	1.03	0.41	0.95	1.42	1.11
235373_at	---	SCC	0.76	2.26	0.65	0.16	0.87	-0.11	0.62	0.70	0.58
235852_at	STON2	SCC	0.45	2.68	0.72	-0.29	1.17	0.09	0.28	0.34	0.71
236009_at	---	SCC	0.40	4.71	0.29	-0.17	1.78	-0.57	0.99	0.61	1.22
236937_at	VPS8	SCC	0.51	2.35	0.76	-0.29	0.80	0.05	0.32	0.30	1.04
238542_at	ULBP2	SCC	0.19	8.95	0.24	-0.58	2.66	-0.52	0.77	0.77	0.72
238827_at	---	SCC	0.48	3.75	0.35	-0.42	1.13	-0.90	0.52	0.20	0.87
238967_at	---	SCC	0.77	2.97	0.19	0.01	1.32	-1.34	1.39	0.76	1.19
239370_at	---	SCC	0.09	24.41	0.07	-0.22	3.66	-0.65	0.93	0.66	2.24
239430_at	IGFL1	SCC	0.69	2.09	0.68	-0.14	0.60	-0.23	0.19	0.16	0.55
240353_s_at	C12orf54	SCC	0.32	5.10	0.35	-0.26	1.32	-0.34	0.23	0.13	1.50
240991_at	---	SCC	0.52	2.69	0.60	-0.52	1.38	-0.30	1.08	1.24	0.65
241418_at	LOC344887	SCC	0.05	42.39	0.05	-0.60	4.23	-0.70	1.13	0.88	2.25
242873_at	---	SCC	0.23	4.14	1.42	0.10	1.30	0.51	0.75	2.15	1.60
242940_x_at	DLX6	SCC	0.31	3.40	0.81	-0.38	1.10	0.33	0.22	0.86	1.44
243018_at	---	SCC	0.38	3.04	0.72	-0.22	1.81	0.31	0.89	1.10	0.65
243252_at	---	SCC	0.50	3.17	0.47	-0.02	1.28	-0.34	0.66	0.77	1.26
244107_at	---	SCC	0.15	11.04	0.22	-0.52	2.36	-0.30	0.13	0.22	1.53
244665_at	---	SCC	0.38	2.02	0.90	-0.34	0.57	0.29	0.41	0.89	1.14
57703_at	SENP5	SCC	0.43	2.54	0.82	-0.16	1.40	0.37	0.52	0.69	0.63

SUPPLEMENTARY TABLE 4: NSCLC HISTOLOGY SIGNATURE (SHORT)

Probe set	Gene Symbol	Signature	ADC:OT Ratio	SCC:OT Ratio	LCC:OT Ratio	ADC Mean	SCC Mean	LCC Mean	ADC SD	LCC SD	SCC SD
203953_s_at	CLDN3	ADC	7.42	0.11	0.33	1.96	-1.56	-0.7	1.05	1.57	0.44
204934_s_at	HPN	ADC	3.95	0.37	0.35	0.97	-0.81	-0.98	0.79	0.57	0.28
205640_at	ALDH3B1	ADC	4.06	0.38	0.32	0.25	-1.45	-1.87	0.86	0.69	0.3
223233_s_at	CGN	ADC	4.92	0.28	0.31	0.88	-1.39	-1.54	0.55	1.01	0.34
242271_at	SLC26A9	ADC	3.66	0.38	0.38	-0.01	-1.81	-1.82	0.66	0.39	0.5
200660_at	S100A11	LCC	1.22	1.65	0.3	0.43	0.75	-1.22	0.48	0.73	0.45
202286_s_at	TACSTD2	LCC	1.51	1.65	0.13	0.99	1.23	-2.73	0.76	1.85	0.52
203849_s_at	KIF1A	LCC	0.24	0.25	8.32	-0.13	-0.23	2.77	0.47	0.85	0.45
207625_s_at	CBFA2T2	LCC	0.58	0.57	2.71	-0.12	-0.21	1.29	0.42	0.34	0.36
225482_at	KIF1A	LCC	0.35	0.33	5.62	-0.19	-0.4	2.25	0.61	0.58	0.53
227998_at	S100A16	LCC	1.36	1.64	0.22	0.49	0.92	-1.6	0.84	0.85	0.23
40016_g_at	LOC100128443 /// MAST4	LCC	1.05	1.83	0.35	-0.23	0.28	-1.61	0.41	0.69	0.42
1556793_a_at	FAM83C	SCC	0.13	12.03	0.2	-0.5	2.75	-0.23	0.09	0.14	1.19
1559606_at	GBP6	SCC	0.09	19.12	0.12	-0.51	3.45	-0.56	0.11	0.09	0.97
1559607_s_at	GBP6	SCC	0.05	38.23	0.06	-0.74	4.19	-0.84	0.39	0.15	1.09
201249_at	SLC2A1	SCC	0.41	3.31	0.57	-0.25	1.45	-0.14	0.34	0.67	0.68
202504_at	TRIM29	SCC	0.17	13.2	0.1	-0.65	3.82	-0.85	1.9	1.37	0.79
203797_at	VSNL1	SCC	0.13	7.01	0.51	-1.11	2.79	0.1	1.06	1.73	0.7
204268_at	S100A2	SCC	0.19	14.45	0.03	0.16	4.92	-1.46	1.92	1.42	0.77
204455_at	DST	SCC	0.07	30.69	0.06	-0.45	4.17	-1.29	1.07	1.23	1.49

204469_at	PTPRZ1	SCC	0.09	26.09	0.04	-0.53	4.44	-1	1.48	0.69	1.06
204734_at	KRT15	SCC	0.09	23.51	0.06	-0.19	4.63	-0.98	1.63	1.44	1.06
205014_at	FGFBP1	SCC	0.09	24.37	0.06	-0.86	3.48	-1.41	1.06	0.38	1.31
205064_at	SPRR1B	SCC	0.06	42.61	0.02	-0.52	4.65	-1.07	1.55	0.2	1.99
205157_s_at	KRT17	SCC	0.06	25.58	0.11	-0.91	4.04	-0.74	1.19	1.52	1.17
205623_at	ALDH3A1	SCC	0.06	28.91	0.1	-0.78	4.13	-0.79	0.88	1.38	1.36
205724_at	PKP1	SCC	0.37	4.14	0.43	-0.1	1.82	-0.09	0.46	0.39	0.81
206156_at	GJB5	SCC	0.31	5.67	0.29	-0.15	2.14	-0.44	0.48	0.41	0.79
206164_at	CLCA2	SCC	0.07	26.84	0.09	-0.94	3.3	-0.96	0.15	0.16	1.52
206165_s_at	CLCA2	SCC	0.02	92.8	0.03	-1.38	4.79	-1.23	0.15	0.46	1.58
206166_s_at	CLCA2	SCC	0.05	37.97	0.07	-0.97	3.87	-0.82	0.12	0.2	1.72
206912_at	FOXE1	SCC	0.04	38.68	0.08	-0.58	4.2	-0.35	0.43	1.03	1.85
207602_at	TMPRSS11D	SCC	0.09	19.64	0.11	-0.26	3.48	-0.29	0.4	0.16	1.61
207935_s_at	KRT13	SCC	0.01	152.16	0.02	-1.12	5.17	-1.14	0.31	0.16	2.03
208539_x_at	SPRR2B	SCC	0.06	29.75	0.08	-0.53	3.59	-0.56	0.27	0.47	1.89
208836_at	ATP1B3	SCC	0.4	3.4	0.56	-0.62	1.28	-0.41	0.67	0.63	0.49
209125_at	KRT6A	SCC	0.04	64.32	0.01	-0.96	6.83	-1.49	2.13	0.78	0.68
209126_x_at	KRT6B	SCC	0.09	25.31	0.05	-0.5	4.23	-1.07	1	0.28	0.61
209351_at	KRT14	SCC	0.01	220.32	0.01	-1.07	5.24	-0.91	0.5	0.38	2.5
209380_s_at	ABCC5	SCC	0.24	6.74	0.3	-0.23	2.56	-0.05	0.8	0.22	0.72
209800_at	KRT16	SCC	0.06	31.35	0.07	-0.51	3.5	-0.7	0.66	0.79	1.95
210020_x_at	CALML3	SCC	0.08	20.73	0.12	-0.49	4.37	-0.19	1.23	1.35	1.22
210505_at	ADH7	SCC	0.07	25.28	0.09	-0.72	3.5	-0.75	0.15	0.17	1.27
210854_x_at	SLC6A8	SCC	0.19	6.26	0.47	-0.19	2.94	0.78	0.92	0.97	0.92
211194_s_at	TP63	SCC	0.13	13.33	0.16	-0.48	2.74	-0.57	0.16	0.45	1.31

211361_s_at	SERPINB13	SCC	0.05	35.59	0.07	-0.68	3.26	-0.58	0.09	0.13	2.3
212236_x_at	KRT17	SCC	0.08	18.33	0.15	-0.71	3.88	-0.55	1.22	1.51	0.97
212702_s_at	BICD2	SCC	0.35	3.77	0.57	-0.59	1.31	-0.14	0.43	0.16	0.9
213680_at	KRT6B	SCC	0.01	124	0.03	-1.21	5.99	-0.82	0.85	1.36	1.15
213796_at	SPRR1A	SCC	0.01	232.64	0.01	-0.91	5.46	-1.12	0.54	0.17	2.53
213843_x_at	SLC6A8	SCC	0.17	6.91	0.42	-0.31	3.06	0.67	1.03	1.02	0.87
214549_x_at	SPRR1A	SCC	0.11	17.6	0.1	-0.27	2.97	-0.56	0.66	0.23	1.84
214580_x_at	KRT6A /// KRT6B /// KRT6C	SCC	0.07	35.21	0.03	-0.8	4.95	-1.17	1.4	0.33	0.31
216918_s_at	DST	SCC	0.23	7.85	0.23	-0.35	2.12	-0.64	0.36	0.32	1.33
217272_s_at	SERPINB13	SCC	0.02	91.29	0.03	-1	4.74	-0.8	0.14	0.19	1.83
217528_at	CLCA2	SCC	0.02	86.63	0.03	-1.36	4.47	-1.27	0.12	0.3	2.05
218990_s_at	SPRR3	SCC	0.03	88.82	0.01	-0.35	5.02	-1.01	1.32	0.15	2.67
221291_at	ULBP2	SCC	0.66	2.23	0.58	-0.16	0.92	-0.41	0.37	0.43	0.38
221795_at	NTRK2	SCC	0.03	59.38	0.05	-1.89	3.94	-1.36	0.53	0.73	1.18
221796_at	NTRK2	SCC	0.03	53.35	0.05	-1.76	3.59	-1.62	0.55	0.62	1.38
221854_at	PKP1	SCC	0.06	21.34	0.15	-0.7	4.8	-0.01	1.34	1.79	0.66
223832_s_at	CAPNS2	SCC	0.2	9.54	0.17	-0.34	2.27	-0.74	0.49	0.13	1.27
225464_at	FRMD6	SCC	0.34	4.49	0.43	-0.71	1.62	-0.62	0.76	0.74	0.34
226755_at	LOC642587	SCC	0.4	3.61	0.51	-0.46	1.34	-0.34	0.34	0.34	0.7
229290_at	DAPL1	SCC	0.02	78.79	0.04	-1.18	4.63	-0.81	0.48	0.72	2.02
230464_at	S1PR5	SCC	0.27	4.76	0.5	-0.66	1.76	-0.3	0.28	0.71	0.49
230769_at	DENND2C	SCC	0.3	5.3	0.36	-0.77	1.53	-0.78	0.21	0.27	0.63
231331_at	---	SCC	0.04	40.8	0.07	-0.68	4.45	-0.31	0.58	0.93	1.53

231771_at	GJB6	SCC	0.02	102.57	0.02	-0.93	5.12	-1.2	0.65	0.29	1.53
232082_x_at	SPRR3	SCC	0.02	119.71	0.02	-0.77	5.07	-0.95	0.44	0.37	2.24
232116_at	GRHL3	SCC	0.32	5.04	0.37	-0.12	2.05	-0.16	0.43	0.49	0.87
232202_at	---	SCC	0.13	11.82	0.22	-0.35	3.73	-0.1	1.24	1.6	1.03
238542_at	ULBP2	SCC	0.19	8.95	0.24	-0.58	2.66	-0.52	0.77	0.77	0.72
241418_at	LOC344887	SCC	0.05	42.39	0.05	-0.6	4.23	-0.7	1.13	0.88	2.25
244107_at	---	SCC	0.15	11.04	0.22	-0.52	2.36	-0.3	0.13	0.22	1.53

SUPPLEMENTARY TABLE 5: NSCLC PATIENT SURVIVAL SIGNATURE

Probe set	Gene symbol	correlation p value	Cytoband
1553300_a_at	DGKH	0.0009322	13q14.11
1557638_at	--	1.08E-05	
201123_s_at	EIF5A	0.0009121	17p13-p12
203634_s_at	CPT1A	0.0004619	11q13.1-q13.2
206262_at	ADH1C	0.0001821	4q21-q23
206581_at	BNC1	0.0007771	15q25.2
206985_at	HSD17B3	7.74E-05	9q22
208459_s_at	XPO7	0.0005945	8p21
210839_s_at	ENPP2	0.0009995	8q24.1
227115_at	--	0.0003025	
231487_at	COX8C	0.0007767	14q32.13
231916_at	EXOSC6	0.0005624	16q22.1
232120_at	EGFR	0.0007851	7p12
233044_at	EGFR	6.40E-06	7p12
233488_at	RNASE7	0.0001726	14q11.2
236646_at	C12orf59	0.0003665	12p13.2
237510_at	MYNN	7.12E-05	3q26.2

Supplementary Table 6. Association between the prognostic predictor and clinical parameters

		Predicted_LOW	Predicted_HIGH	p-value
		(n = 50)	(n = 32)	
Age	Mean	60.49	65.96	0.023
	SD	10.56	10.18	
Tumor cell %	Mean	63.22	68.30	0.242
	SD	21.00	15.44	
Stage_I		34	17	0.197
		68.0%	53.1%	
Stage_II		11	9	0.366
		22.0%	28.1%	
Stage_III+IV		4	6	0.789
		8.0%	18.8%	
Smoking years	Mean	33.75	39.88	0.014
	SD	10.98	10.50	
Gender				
Female		19	4	0.012
		38.0%	12.5%	
Male		31	28	
		62.0%	87.5%	
Forced Expiratory Volume 1	Mean	88.09	78.10	0.009
	SD	17.49	14.45	
Tumor size	Mean	37.07	38.56	0.622
	SD	13.79	12.74	
Histology				
ADC		16	8	0.991
		32.0%	25.0%	
SCC		5	11	0.942
		10.0%	34.4%	
LCC		17	7	0.967
		34.0%	21.9%	
OTHER		12	6	0.975
		24.0%	18.8%	

Supplementary Table 7. Relation between variables and the relative hazard ratio

	Wald test	p-value
Age	1.68	0.20
Tumor cell %	0.95	0.33
Stage	3.80	0.05
Gender	0.00	0.99
Smoking years	0.08	0.78
Forced Expiratory Volume 1	0.54	0.46
Tumor size	0.02	0.90
Histology	2.43	0.12
Prognostic predictor	21.68	0.000003

Supplementary Table 8. Prognostic signatures for NSCLC

Publication	NSCLC subtypes	platform	# of genes in signature	genes (probe sets) present on Affy U133 plus 2.0
Chen et al. (2007) NEJM 356, 11-20	ADC, SCC, other	9.6k cDNA home made	16	16 (55)
			5	5 (18)
Beer et al. (2002) Nat Med 8, 816-824 **Guo et al. (2008) CCR 14, 8213-8220**	ADC	Affy HuGeneFL	35	35 (134)
Shedden et al. (2008) Nat Med 14, 822-827	ADC	Affy U133A	A_>1000	not tested
			B_52	50 (52)
			C_26	23 (26)
			D-42	36 (42)
			E_1	1 (1)
			F_41	37 (41)
			G_38	36 (38)
			H_313	249 (313)
Lee et al (2008) CCR 14, 7397-7404	ADC, SCC	Affy U133 plus 2.0	20	20 (20)
			6	6 (6)
Roepman et al. (2009) CCR 15, 284-290	ADC, SCC, other	Agilent 44k	72	68 (224)
Boutros et al. (2009) PNAS 106, 2824-2828	ADC, SCC	QRT-PCR	6	6 (8)

Chapter 5

Expression profiling-based prediction of the putative response of NSCLC patients to Pemetrexed therapy

Abstract

Pemetrexed effectiveness has been related to the expression of its target Thymidylate Synthase (TYMS). The more frequent resistance to Pemetrexed in lung squamous cell carcinoma (SCC) patients is ascribed to high level of TYMS. In this study, the gene expression level of TYMS and other targets of Pemetrexed was profiled in 91 non-small cell lung cancer (NSCLC) subjects using Affymetrix expression microarrays. A novel subgroup of putative resistant NSCLC cases to Pemetexed was identified and its distinct molecular attributes were bioinformatically studied.

Introduction

Presently, Pemetrexed is one of the most effective drugs for the treatment of non-small cell lung cancer (NSCLC). Pemetrexed is an anti-folate metabolite and targets multiple molecules essential in nucleotide biosynthesis (Chapter 1. Fig. 2). It was established that it has possibly superior activity compared to commonly used agents in adenocarimoma (ADC) and large cell carcinoma (LCC) but is less effective in squamous cell carcinoma (SCC). In previous published studies, the effectiveness was related to the expression level of Thymidylate Synthase (TYMS). It was demonstrated that high expression of TYMS is associated with resistance to Pemetrexed in NSCLC ($p = 0.006$); furthermore a higher expression of TYMS is more often seen in SCC than ADC and LCC [1-3]. Based on those preclinical observations, Pemetrexed is approved as the first-line treatment in combination with other agents for advanced non-SCC NSCLC patients [2]. Its efficacy when used as a single agent for first-line treatment is under investigation.

Among ADC and LCC patients, the response rate to Pemetrexed varies between 16 to 45%. Intriguingly, a significant number of ADC and LCC cases with high level of TYMS expression were observed, as determined by immunohistochemistry. Therefore, it indicates that to reach a higher CR/RR of Pemetrexed, it is vital to predict the response for individual patients by actually recruiting a substantial criterion, such as TYMS expression level, while histology-based stratification is proved insufficient in clinical practice.

Genome-wide expression studies have revealed that NSCLC subgroups may be classified beyond classical histo-pathological criteria [4-7]. The clinical potential of such refined phenotyping of NSCLC tumors, however, has as yet not been fully explored.

In this study, we show that ADC and LCC can be partitioned into novel subgroups based on global gene expression profiles. A subset of ADC and LCC were clustered in a novel group. Analysis of the expression level of TYMS and relevant genes indicates that tumors in this group are highly likely to be resistant to Pemetrexed therapy. The identification of this distinct subgroup of NSCLC suggests that biological characteristics assessed by gene expression profiling may aid in reliably stratifying patients in respect of Pemetrexed first-line therapy. Thus, we propose a clinically feasible approach to evaluate expression levels of TYMS and associated genes, in order to molecularly predict Pemetrexed efficacy in NSCLC patients.

Materials and Methods

NSCLC tumor smaples

Ninety-one resected tumor samples from NSCLC patients were collected at the Erasmus MC between 1992 and 2004. Tissues was collected and studied under an anonymous tissue protocol approved by the local medical ethical committee of the institution.

Histopathological analysis

All tumor samples were independently reviewed by two pathologists. The dominant molecular characteristics of tumors were also verified by histology gene signatures established in Chapter 4. According to the molecular level classification, the cohort included 45 ADC, 27 SCC, and 19 LCC, including 3 CAR classified as LCC and 1 as ADC. Patient and tumor characteristics are listed in Supplementary Table 1.

NSCLC cell lines

Nine NSCLC cell lines within NCI-60 drug screen panel were transcriptionally profiled by Affymetrix U133 Plus 2.0 array (GSE8332). The expression of

relevant probe-sets of interest was directly retrieved from Gene Expression Omnibus (GEO) database (NCBI) with using a script written in MATLAB. The sensitivity of these NSCLC cell lines to Pemetrexed was tested in vitro using a standard MTT colorimetric assay via quantifying the amount of viable cancer cells [8].

Microarray data analysis

RNA from frozen tumor tissues was isolated and processed according to the standard protocol for Affymetrix U133 Plus 2.0 arrays. The details about microarray data processing and normalization are described in Chapter 4 Supplementary Method.

Scoring formula using Internal Reference Genes (IRG)

The detailed predictive scheme methodology is described in Supplementary Materials and Methods. The scheme predicted tumor response utilizing the expression of TYMS, the major target of Pemetrexed, alone firstly, and then the expression of 3 targets, TYMS, DHFR, and GART.

In the IRG scheme, the average expression level of 11 probe sets was used as the internal reference to determine the relative expression of Pemetrexed target genes. NSCLC patients were stratified in such a way that around 60% of cases were supposed to respond to Pemetrexed, while the remaining 40% of cases (~ 40%) with higher expression levels of the signature genes were deemed to be non-responders.

Supervised analysis to identify resistance related genes

An optimized gene signature characterizing high TYMS expression, and hence presumably resistance to Pemetrexed treatment, was obtained and validated with previously described bioinformatics approaches [9-11].

Tissue Microarray Analysis (TMA)/Immunohistochemistry

Tissue microarrays composed of 91 paraffin-embedded primary lung tumors from the same patients used for microarray analysis were constructed. Cores of 0.6 mm were punched from selected tumor areas from the donor block and placed in triplicate in an acceptor block. For immunohistochemistry, 5-µm sections were mounted onto slides and stored at room temperature until used.

Immunohistochemistry was performed using standard methods.

Results

Six novel NSCLC subgroups

Unsupervised clustering on expression profiles revealed six subclasses within 91 NSCLC cases. The initial clustering was based on the similarity in global gene expression between the NSCLC samples and performed with 11,515 probe sets. The distinct subgroups were also recognized when the number of probe sets was reduced to 4791, with the resulting six NSCLC subclasses presenting with the strongest similarities in gene-expression profiles within each subclass but dissimilar with other subclasses (Fig.1).

The majority of tumor samples (89 out of 91) were clustered into 6 groups. Two of these groups correlated well with classical histopathology classes: Group3 displayed a dominant SCC contribution, while the CAR samples (n = 4) were exclusively assigned to Group5. In contrast to these two groups, other groups did not show such a strong association with classical histology. They were to varying degrees composed of mixed histopathological NSCLC. ADC accounted for a major part of each of these groups, and most of LCC - large cell neuroendocrine carcinoma (LCNEC) were mingled with ADC in Group4 and Group6. Compared to Group1 and Group2, ADC in Group4 and Group6 displayed gene expression patterns suggestive of neuroendocrine features. Regardless of histological consistency between Group1 and Group2, the NSCLCs in these two groups were distinguished by a low degree of cell differentiation and the expression of a large number of immune-related genes, respectively.

Gene expression-based prediction for response to Pemetrexed

The potential sensitivity of tumors to Pemetrexed treatment is thought to be positively correlated with the expression levels of enzymes in the nucleotide metabolic pathway [3, 12]. Expression of the relevant genes was extracted from the microarray data, and these were subsequently utilized as the basis of predictive schemes for Pemetrexed responsiveness.

1. **Prediction in primary NSCLC**

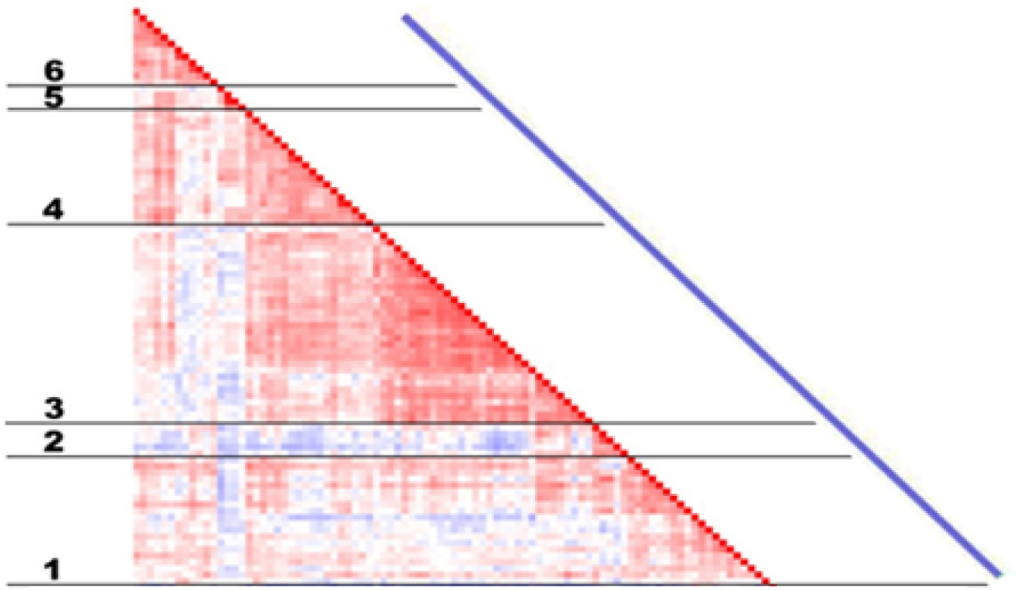


Fig. 1. Correlation view of 91 samples from patients with NSCLC.
Pairwise correlations between any two samples are displayed. The colors of the cells represent Pearson's correlation coefficient values, with deeper red indicating higher positive and deeper blue lower negative correlations. The red diagonal line displays the self-to-self comparison of each sample.

The expression of TYMS and/or DHFR, GART was compared to the average expression of internal reference genes. For each patient, if more than half of target gene representing probe sets had expression above the 60th percentile of that gene's expression in the cohort, that patient was supposed to be resistant to Pemetrexed targeted therapy (NR); if none of probe sets had higher than the 60th percentile expression, then responder to Pemetrexed was assigned (R); if just around half of probe sets had expression above the 60th, the medium sensitivity (M) was assigned to leave a margin between NR and R. According to the internal reference gene scheme, out of 91 NSCLC patients 35.2% were predicted as non-responders and 51.5% were predicted to be responders.

The average relative expression value of TYMS from predicted non-responders is 177.1 (95% CI: 143.2~210.9), 8.2-fold more compared to that in normal lungs; while predicted responders displayed a 2.2 fold change in relative expression value of TYMS compared to that in normal lungs.

2. **Predicted resistance profile in relation to histology**

The predicted tumor resistance profile was studied in relation to the three major NSCLC subtypes, ADC, SCC and LCC. The histology was assigned using histology gene signatures identified in Chapter 4 which captured the most predominant molecular features of NSCLC.

Within these three major groups, LCC contained the highest relative expression of TYMS (192.0; 95% CI: 125.6~258.4), followed by SCC (86.6; 95% CI: 73.0~100.1) and ADC (76.1; 95% CI: 58.5~93.7) (Fig. 2A and Supplementary Fig. 1). When expression of TYMS was compared to normal lung tissues, a significant difference was observed between each subtype of NSCLC and non-cancerous tissues, with 8.85-, 4-, and 3.5-fold differences in LCC, SCC, and ADC, respectively. Among subtypes of NSCLC, the difference in TYMS expression was statistically significant between LCC and other two subtypes, ADC and SCC; but not between ADC and SCC (Table 1).

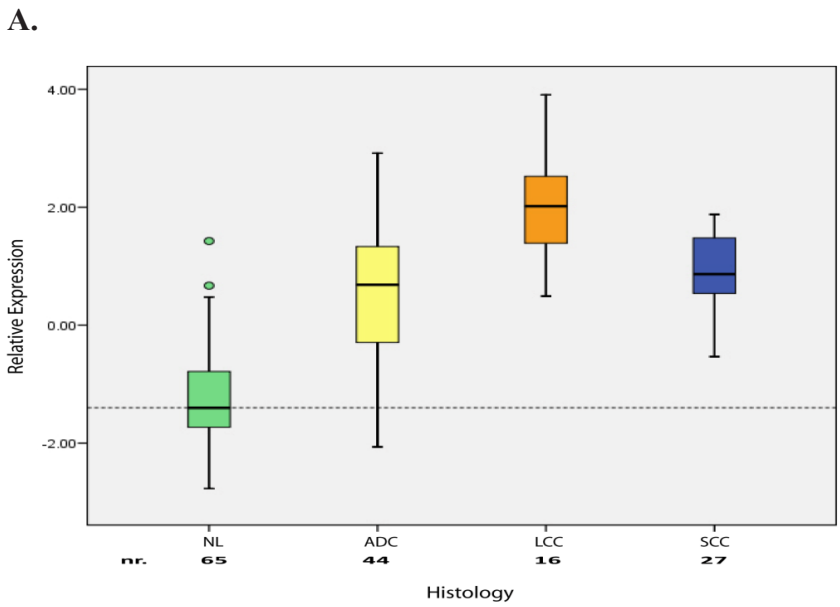
	ADC	**SCC**	**LCC**	**NL**
ADC	-	0.398	0.002*	0 *
SCC	-	-	0.004*	0 *
LCC	-	-	-	0 *

Table 1. The differences in TYMS expression between NSCLC and normal lung, or between any two NSCLC subtypes. Values in table are t-test P-values.

The predicted resistance to Pemetrexed was correlated to 25% of ADC, 33% of SCC, and 63% of LCC (Fig. 2B). By contrast, all non-cancerous tissues from same patients except for one were stratified as being sensitive to Pemetrexed, in accordance with the observation that the expression of TYMS is significantly higher in LCC, SCC, and ADC than in normal lung tissues.

3. **A novel NSCLC group is associated with predicted Pemetrexed resistance**

Since the number of predicted resistant cases (NR) in each histological subtype was not deviated enough to reliably make that subtype of NSCLC being contradictive to Pemetrexed therapy. We tried to correlate predicted



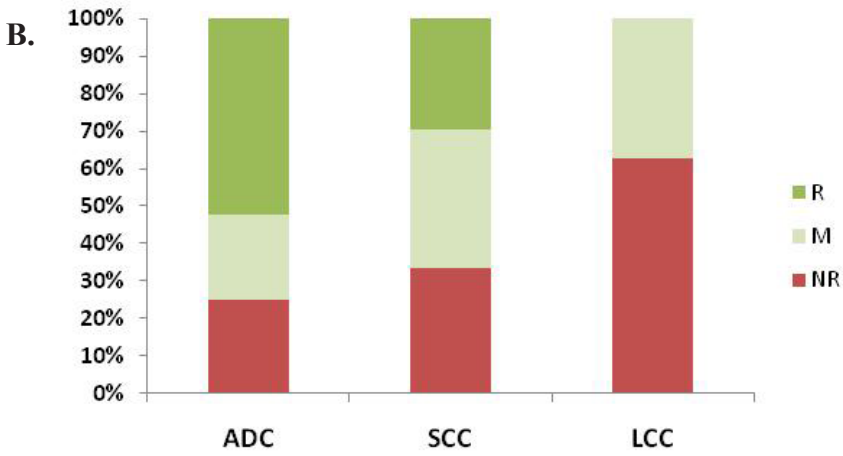


Fig. 2. Expression of TYMS and predicted non-responder in relation to histology
A. Expression of TYMS and distribution of high TYMS NSCLC in adenocarcinoma (ADC), squamous cell carcinoma (SCC), and large cell carcinoma (LCC) is compared to TYMS expression in normal lung (NL). **B.** the composition of internal reference gene scheme defined resistant NSCLC (NR), sensitive NSCLC (R), and medium sensitive NSCLC (M) in each NSCLC subtype.

Pemetrexed resistance to six novel NSCLC subgroups described in Fig. 1. It is noticed that a distinct resistance pattern of NSCLC to Pemetrexed was

recognizable by novel grouping of NSCLC

In four out of the six groups, no more than 25% of the cases were defined as predicted non-responders, with percentages of 25%, 0%, 25% and 8.3% in Group1, Group2, Group5 and Group6 respectively. About 39% cases from Group3, which were characterized by SCC, were predicted as non-responders.

In Group4, comprising large cell neuroendocrine carcinoma (LCNEU) and ADC with neuroendocrine features, a remarkable proportion of predicted non-responders was observed, 88.9% (16 out of 18) of cases in this group were gene expression-defined NR to Pemetrexed therapy (Fig. 3).

Prediction based on the expression of all three targets

Next, the expression profiles of DHFR or GART were included to predict Pemetrexed resistant and sensitive NSCLCs. This resulted in a similar "non-responder" distribution according to either classical histology or the novel groups as we observed with TYMS-only stratification. Similarly, Group4 was recognized by a high proportion of NR, with percentages ~60% and ~82% of cases predicted being resistant to Pemetrexed.

When all three Pemetrexed target-encoding genes were utilized to equip the predictive scoring formulae, the resulting resistance pattern favored Group4 again (Supplementary Fig.2).

We conclude that the gene expression profile-based prediction shows that NSCLC cases falling into Group4 and around 40% of SCC are most likely to be resistant to Pemetrexed.

Distinct molecular characters of Pemetrexed resistant NSCLC-Group4

Three out of six distinct groups of NSCLC, Group4, Group5, and Group6 comprised neuroendocrine tumors, including LCENC, CAR, as well as ordinary NSCLC with neuroendocrine differentiation, mainly ADC. Among them, Group4 and Group6 were histologically similar but varied in expression profiles.

1. **Neuroendocrine NSCLC and distinct features of Group4**

Several neuroendocrine markers, including ASCL1, DDC, and MAST4 were expressed by tumors presenting evidence of neuroendocrine differentiation,

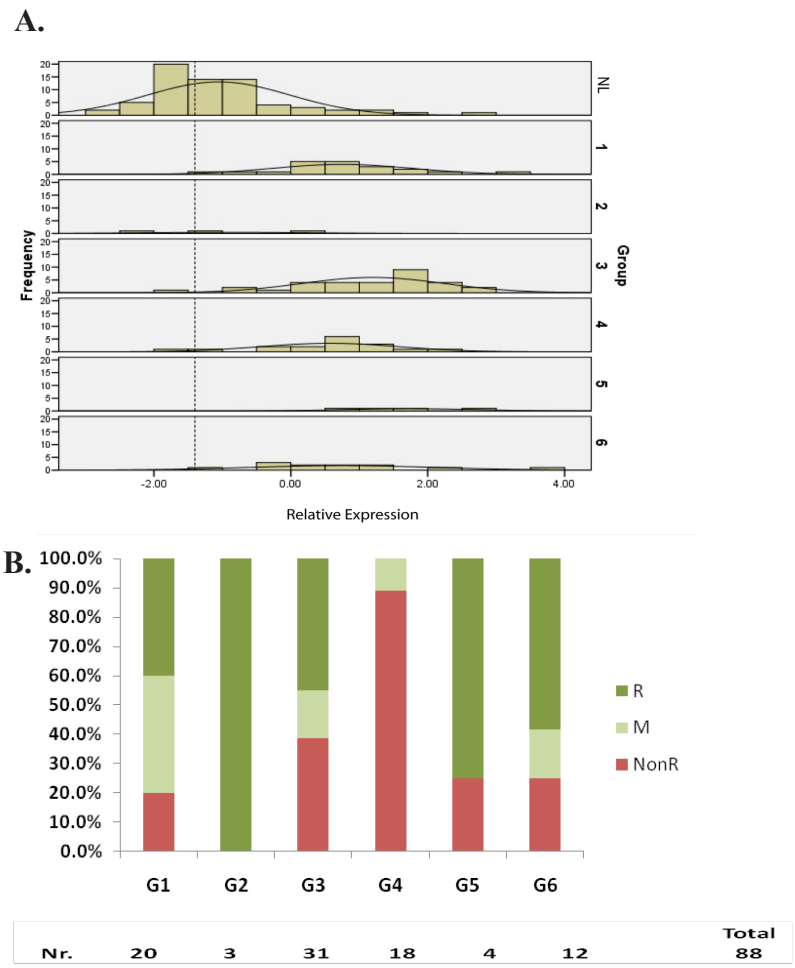


Fig. 3. Expression of TYMS and predicted Pemetrexed response in relation to NSCLC novel groups. **A.** relative expression of TYMS and distribution of high TYMS NSCLC in each group compared to TYMS expression in normal lung tissue (NL). **B.** the composition of internal reference gene scheme defined resistant NSCLC (NR), sensitive NSCLC (R), and medium sensitive NSCLC (M) in each NSCLC novel subgroup.

such as LCNEC and a subset of ADC [6]. These genes were over-expressed by Group4 and Group6 NSCLC as well, indicating the presence of common neuroendocrine features among these groups regardless of their molecularly defined different histological characters. However, there was a 2- to 4-fold difference in expression of these markers between Group4 and Group6.

2. **Predicted Pemetrexed resistance in Group4 NSCLC**

Pemetrexed is transported into and out cells by a class of membrane proteins, such as FOLR1, SLC19A1, and ATP-binding cassette family members. Moreover, Pemetrexed is metabolized by folylpolyglutamate synthetase in vivo. Besides high expression of TYMS, DHFR and GART, the aberrant expression of such related molecules may contribute to Pemetrexed resistance as well. Interstingly, we observed lower expression of FOLR1 in Group4 NSCLC compared to other Groups.

We conclude that the novel Group4 of NSCLC has distinct molecular characteristics associated with predicted drug sensitivity.

Proposed Pemetrexed resistance signature

1. **Predicted NSCLC resistance was characterized by 346 genes**

A set of 346/426 genes/probe sets characterizing predicted Pemetrexed resistance was identified using supervised analysis as described in Chapter 4. The bioinformatics analysis revealed that cell cycle and cell proliferation genes, pyrimidine and purine metabolism, as well as folate biosynthesis were enriched in resistance signature (Supplementary Table 2).

SOX2 and SOX4 are among the up-regulated genes, and surfactant genes are down-regulated in predicted non-responders. Interestingly, SOX7, a marker for squamous cell differentiation, was also down-regulated among patients in non-responders. Two solute carriers were also down-regulated genes, SLC16A4 and SLC46A3. Hereinto, SLC46A3 is from the same family as SLC46A1 which is responsible for the folate transporter in vivo at low pH level. By contrast, DNA damage repair associated genes, attributing to multi-drug resistance, were found over-expressed in NRs, including TOP2A, PRIM1, and RFCs. Among resistance signature genes, a large number of cell cycle regulatory genes were found, including cyclin A2, B1, E1 and E2; CDC2, CDC6, CDC7, CDC20, and CDC25; checkpoint kinase CHEK1, and related PLK1, PLK4; proliferation or mitosis related genes like E2Fs, GTSE1, KIFs, MCMs, and IGFBP2; cell growth and invasion related genes MMP19; as well as known oncogenes and suppressor genes MYB, NBL1, RAS genes.

A subset of genes, represented by 25 probe sets, optimally performed in predicting Pemetrexed resistance [10, 11].

2. Validation of the putative Pemetrexed resistance signature by NSCLC cell lines

The expression of resistance associated genes identified with primary NSCLC cohort was measured in transcriptionally profiled NSCLC cell lines (GSE8332). The performance of this signature was evaluated by comparing the prediction for cell line reaction to Pemetrexed to actual observations from drug sensitivity assays [8] (Fig. 4). The minimized prediction signature was achieved with a correct prediction of sensitivity to Pemetrexed in 94% (17 out of 18) of the cell lines. We note that resistant cell lines were all correctly predicted; the sensitivity of predicting resistance is 100%, with a corresponding specificity 91.7%.

Comparison of non-responder classifiers to other signatures

The putative non-responder biomarker genes were compared to onco-pathway signatures to further explore the role of these genes in oncogenesis (Supplementary Table 3) [13]. We noticed that the E2F3/G1S regulatory pathway was over-represented by the biomarker genes, followed by the SRC, P53, and RAS pathways.

The comparisons were also performed between putative non-responder biomarkers and other gene signatures developed in Chapter 4. The biggest overlap was found with novel Group4 signature (Supplementary Table 3).

Validation of TYMS expression in NSCLC

To validate expression of TYMS and other relevant genes measured on microarray, quantitative polymerase chain reaction (qPCR) was performed. Pearson's correlation coefficient was used to measure the consistency in relative expression of tested genes between two assays (Supplementary Table 4).

TYMS protein staining was graded from 0 to 3 corresponding to no staining, plus 1 to 3 staining. The mRNA expression level measured on microarray for each staining category was shown in Fig. 5. Since only one sample had staining scaled with 3 plus, it was combined with other plus 2 staining NSCLCs.

The average expression of TYMS of Grade2 was 3.73-fold and 2.52-fold higher than Grade0 and Grade1 respectively (P_value = 1.11E-07). The

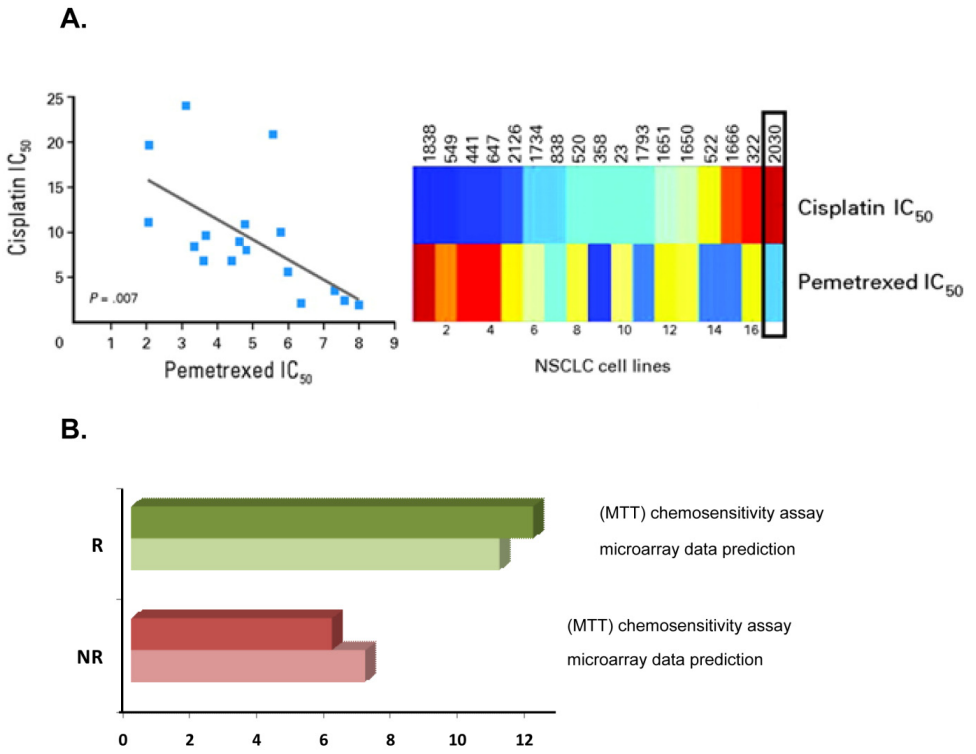


Fig. 4. A. The correlation of cisplatin response and Pemetrexed response [8]. **B.** Predicted sensitivity to Pemetrexed for NSCLC cell lines in comparison to experimentally established sensitivity.

correlation between staining intensity and predicted Pemetrexed resistance is shown in Figure 5 and Supplementary Figure 4. Over 85% NSCLC with Grade2 staining were predicted resistant to Pemetrexed based on mRNA expression of TYMS. Conversely, less than 30% of NSCLC with Grade1 staining were predicted non-responder. The expression difference of TYMS in non-responder between staining Grade0 and Grade1 was not statistically significant (P-value = 0.993), however, TYMS presented a significant higher expression in Grade2 compared to either Grade0 or Grade1 with p-values of 0.067 and 0.048.

Discussion

Chemotherapy acts as a two-edged sword in cancer therapy. It provides an alternative to surgical removal to kill cancer cells and prevents disease progression on one side. On the other side, chemotherapy introduces a range

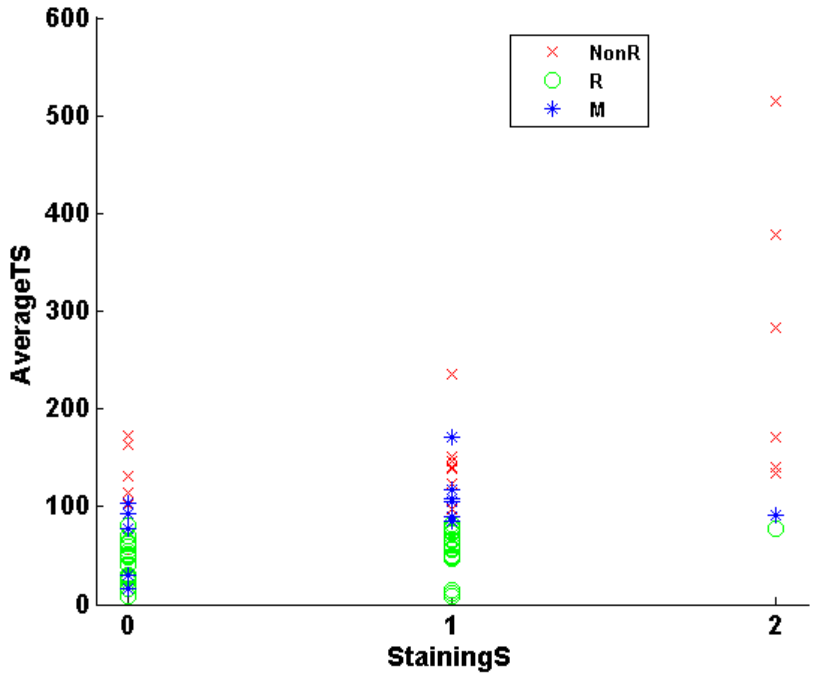


Fig. 5. The correlation of TYMS protein staining and mRNA expression.
Protein staining on TMA was graded from 0 to 2 (StainingS), mRNA expression measured on expression microarray was represented as mean of two probe sets for TYMS (AverageTS).The predicted response to Pemetrexed was illustrated by non-responder (NR), responder (R), and medium sensitivity (M).

of side-effects affecting normal cells in the body, leading to clinical complications, such as nausea, anemia, multiple organ toxicity, myelo-suppression, and secondary neoplasms. Contradictive roles of chemotherapy require tailored regimens for individual patients, obtaining maximal effect while drug-related toxicity remaining at a tolerable grade. Currently, the application of chemotherapy is based on histology, resulting in no more of a 16% of response for common chemotherapeutic agents and 40% for combined regimens, at the same time a 5-year survival kept low [14].

The fact that cancer patients with similar pathological features response dramatically different to the same therapeutic agent indicates that histology alone is insufficient to determine tumor response to chemotherapy. Thus, there is need to improve the stratification of cancer patients and predict clinical

outcome following a particular treatment.

Pemetrexed is a promising anti-metabolite agent for the treatment of NSCLC, which produces comparable anticancer-activity to other conventional agents, but drug-related toxicity is much milder [1]. Currently, the administration of Pemetrexed is limited to ADC and LCC cases, since preclinical studies demonstrated no apparent benefits for SCC patients.

In this study, however, we find that the TYMS expression-based prediction of NSCLC response to Pemetrexed does not correlate well with the classical histological subtypes. Using the genome-wide expression data, we identified a novel subgroup of NSCLC with both neuroendocrine features and altered tyrosine metabolism that is predicted to be resistant to Pemetrexed treatment (Supplementary Fig. 3). In contrast, other NSCLCs, including ADC and LCC in Group1 and Group6 in particular, were identified as candidates for Pemetrexed treatment as they were predicted to respond favourably to the treatment. Remarkably, around of 60% SCC in our cohort were predicted to be sensitive to Pemetrexed therapy. This suggests that Pemetrexed may be an effective chemotherapeutic agent for this subset of patients. Similarly, our analysis suggests that CAR in Group5 is another subset of tumours that might benefit from Pemetrexed therapy. Although the activity of this agent in patients with this type of tumor needs to be evaluated further.

In summary, NSCLC cases predicted to be resistant to Pemetrexed include:

- ADCs, with evidence of neuroendocrine differentiation and in association with aberrant tyrosine metabolism, in Group4;
- LCNEC classified into Group4;
- A subset of SCC, around 40%;

Validation of the predicted Pemetrexed non-responder gene signature

The role of TYMS expression in the efficacy of Pemetrexed therapy has been established by several studies [3, 8, 12, 15]. We stratified NSCLC patients into different response groups using TYMS expression levels extracted from microarray data, and extended these observations to develop a Pemetrexed resistance gene signature. The performance of this resistance signature

was evaluated with NSCLC cell lines whose sensitivity to Pemetrexed was previously determined [8].

This non-responder gene signature accurately predicted Pemetrexed sensitivity of the NSCLC cell lines. Unfortunately, it is currently not feasible to use patient samples for validation of the performance of the non-responder gene signature, due to several factors.

1.) Lack of availability of tumor material before treatment to identify determinant genes for Pemetrexed response. The postoperation survival rate of NSCLC remains low, even among stage I patients. Most of patients deceased within 2 years of the operation, due to the progression of NSCLC or accompanying systematic complications. It is extremely rare to have the collection of primary and the second surgery-resections from the same patients, not mention to patients received Pemetrexed treatment.
2.) Combination with cisplatin or following other agents makes it difficult to assess the role of Pemetrexed alone. Pemetrexed is approved as second-line treatment of NSCLC, or first line treatment of NSCLC in combination with cisplatin. According to previous studies [16], chemotherapy can induce a global change of gene expression. Therefore the independent effect introduced by Pemetrexed is hard to monitored using clinical patient materials. Although few clinical trials where Pemetrexed was administrated as the single agent of first line therapy for NSCLC patients are ongoing, the availability of corresponding tumor samples is still quite limited.

Evaluating TYMS expression level by IHC

TYMS staining was performed at two different antibody titers, 1:10 and 1:50. As shown in Supplementary Fig. 4, strong staining/TMA high grade NSCLCs are predominantly associated with non-responder predicted on the basis of mRNA expression levels. In weak staining/low grade NSCLCs, however, resistant cases accounted for 21 to 29% of whole cases.

The similar scale of staining between Grade0/1 non-responder and responder was postulated to result from unspecific staining of TYMS. This

assumption is supported by staining with a lower antibody titer (1:50). Supplementary Figure 4 shows TYMS staining with more diluted TYMS antibody, and scaled identically with low titer TMA. Predicted non-responders are still predominant in TMA Grade2 population of similar size as the previous one. By contrast with TMA at titer 1:10, only one non-responder is found in low titer Grade1 NSCLCs (1 out of 8, 12.5%), with unspecific staining eliminated to great extent. Higher diluted antibody on the other hand failed to detect relatively low expression of TYMS, with 80% of the cases showing no TYMS protein expression and scaled as Grade0. The results illustrate the technical difficulties in determining patients' TYMS levels through routine immunohistochemistry. Herewith, we provide candidate molecules, such as DEPDC1, as surrogate markers whose expression levels in NSCLC should reflect TYMS expression levels, while allowing faithful measurement in routine clinical practice.

Gene expression profiles may guide the choice of chemotherapy regimens

We suggest that molecular profiles of individual NSCLC tumors may be used to predict patient response to Pemetrexed chemotherapy (Fig 6). Our observations indicate that a subset of NSCLC, independent of histology, should be contra-indicated for Pemetrexed as therapeutic agent. Since a reverse correlation between Pemetrexed and Cisplatin sensitivity that has been observed in ovarian cancer patients [8], Cisplatin could be used in those cases. Furthermore, the efficacy of Pemetrexed in patients with low-TYMS SCC deserves further exploration. Our data suggest that a significant number of SCC patients may benefit from Pemetrexed treatment.

To implement in clinical practice, we propose a set of biomarkers for IHC which overcome the poor reliability of IHC staining to evaluate TYMS expression levels. We show that IHC results obtained with DEPDC1 or other antibodies correlates much better with the TYMS expression levels determined by microarray and Q-PCR. Thus, use of these novel markers may aid in deciding to select the treatment regime for individual NSCLC patients.

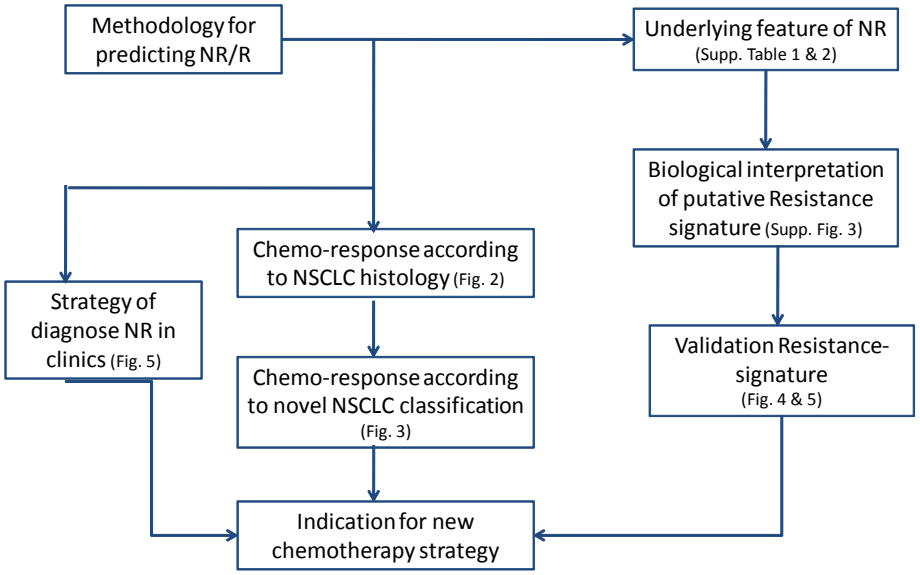


Fig. 6. Study scheme and hypothesized clinical application of biomarker-based prediction for Pemetrexed non-responders

References

1. Esteban, E., M. Casillas, and A. Cassinello, *Pemetrexed in first-line treatment of non-small cell lung cancer.* Cancer Treat Rev, 2009. **35**(4): p. 364-73.
2. Longo-Sorbello, G.S., et al., *Role of pemetrexed in non-small cell lung cancer.* Cancer Invest, 2007. **25**(1): p. 59-66.
3. Travis, W.D., et al., *Reproducibility of neuroendocrine lung tumor classification.* Hum Pathol, 1998. **29**(3): p. 272-9.
4. Takeuchi, T., et al., *Expression profile-defined classification of lung adenocarcinoma shows close relationship with underlying major genetic changes and clinicopathologic behaviors.* J Clin Oncol, 2006. **24**(11): p. 1679-88.
5. Bhattacharjee, A., et al., *Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses.* Proc Natl Acad Sci U S A, 2001. **98**(24): p. 13790-5.
6. Anbazhagan, R., et al., *Classification of small cell lung cancer and pulmonary carcinoid by gene expression profiles.* Cancer Res, 1999. **59**(20): p. 5119-22.
7. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-

21.
8. Tibshirani, R., et al., *Diagnosis of multiple cancer types by shrunken centroids of gene expression.* Proc Natl Acad Sci U S A, 2002. **99**(10): p. 6567-72.
9. Golub, T., et al., *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science, 1999. **286**: p. 531-536.
10. Hsu, D.S., et al., *Pharmacogenomic strategies provide a rational approach to the treatment of cisplatin-resistant patients with advanced cancer.* J Clin Oncol, 2007. **25**(28): p. 4350-7.
11. Meier, P. and E. Kaplan, *Nonparametric estimation from incomplete observations.* J Am Stat Assoc, 1958. **158**: p. 457-481.
12. Bepler, G., et al., *Clinical efficacy and predictive molecular markers of neoadjuvant gemcitabine and pemetrexed in resectable non-small cell lung cancer.* J Thorac Oncol, 2008. **3**(10): p. 1112-8.
13. Giovannetti, E., et al., *Cellular and pharmacogenetics foundation of synergistic interaction of pemetrexed and gemcitabine in human non-small-cell lung cancer cells.* Mol Pharmacol, 2005. **68**(1): p. 110-8.
14. Bild, A.H., et al., *Oncogenic pathway signatures in human cancers as a guide to targeted therapies.* Nature, 2006. **439**(7074): p. 353-7.
15. Rosell, R., et al., *The biology of non-small-cell lung cancer: identifying new targets for rational therapy.* Lung Cancer, 2004. **46**(2): p. 135-48.
16. Hanauske, A.R., et al., *In vitro chemosensitivity of freshly explanted tumor cells to pemetrexed is correlated with target gene expression.* Invest New Drugs, 2007. **25**(5): p. 417-23.
17. Buchholz, T.A., et al., *Global gene expression changes during neoadjuvant chemotherapy for human breast cancer.* Cancer J, 2002. **8**(6): p. 461-8.

Supplementary Material and Method

Microarray data processing and normalization

RMA normalization Microarray data was normalized by RMA algorithm. RMA (Robust Multi-Array average) is an integrated algorithm comprising background adjustment, quantile normalization, and expression summarization by median polish [17]. The intensities of mismatch probes were entirely ignored due to their spurious estimation of non-specific binding. The intensities were background-corrected in such a way that all corrected values must be positive. The RMA algorithm utilized quantile normalization in which the signal value of individual probes was substituted by the average of all probes with the same rank of intensity on each chip/array. Finally Tukey's median polish algorithm was used to obtain the estimates of expression for normalized probe intensities. Intensities of probe sets lower than 30 were reset to 30.

Probe sets filtering Probe sets were involved in further analysis only if their expression levels deviated from the overall mean in at least one array by a minimum factor of 2.5, because the remaining data were unlikely to be informative. The result was that 43,160 probe sets were eliminated, and 11,515 probe sets remained for further analysis.

Unsupervised clustering and Novel grouping of NSCLC

Omniviz software (Omniviz, Maynard, MA) was used to measure the similarities in expression profiles among samples based on 11,515 selected probe sets (Chapter 4 Supplementary material). The samples were ordered so that those sharing strong similarities were arranged together into clusters. The clusters and the individual samples within the clusters were sorted in such a manner that the more similar subjects were more closely positioned in the visualization matrix. Six distinct NSCLC clusters were identified by gene expression profiles, as described in Chapter 4.

Scoring formula elaborating

The expression of genes encoding Pemetrexed targets measured by microarray was employed to classify NSCLC to different response groups. All schemes predicted tumor response utilizing the expression of TYMS, the major target

of Pemetrexed, alone firstly, and then the expression of all 3 targets, TYMS, DHFR, and GART.

1. **Internal Reference Genes (IRG)**

To be less prone to cohort-inherent and any analysis-derived variability, the methodology was adjusted to be individually determinant, the expression level of TYMS genes being scaled in comparison with other genes from the same tumor instead of with the expression of same genes across all tumors in the cohort.

To define internal reference probes/genes, we firstly selected top 100 probe sets which showed a constant expression under various conditions. The constant expression of those probe sets was confirmed by an independent data set [18], which contained a similar number of NSCLC samples (n = 96). To be applicable in future for different platforms or different generations of the same platform, we checked the presence of these probe sets on U133 set of Affymetrix chips. The average expression of 11 probe-sets was used as the internal reference to determine the relative expression of genes encoding Pemetrexed targets.

2. **The percentile rank-based definition**

Responder: none of counted probe-sets/genes showed an expression above the 60^{th} percentile of that population; in case all 3 targets were used, no more than 3 counted probe-sets had expression intensity above the 60^{th} percentile of the population studied.

Non-Responder: at least 2 out of 3 probe-sets/genes presented an expression higher than 60% of studied population; Or 6 out of 14 in cases where all 3 targets were counted.

Supervised analysis to identified Pemetrexed resistance associated genes

Gene profiling with respect to predicted sensitivity to Pemetrexed was performed by using Significance Analysis of Microarray (SAM). SAM discovered differentially expressed genes between two classes [9], predicted non-responders and responders.

The obtained signatures were subjected to identify subgroups of genes that maintain the capacity of the complete signatures in distinguishing dif-

ferent groups maximally [10]. The performance of optimized signatures was validated by “leave-one-out” cross validation [11].

References:

1. Esteban, E., M. Casillas, and A. Cassinello, *Pemetrexed in first-line treatment of non-small cell lung cancer.* Cancer Treat Rev, 2009. **35**(4): p. 364-73.
2. Longo-Sorbello, G.S., et al., *Role of pemetrexed in non-small cell lung cancer.* Cancer Invest, 2007. **25**(1): p. 59-66.
3. Bepler, G., et al., *Clinical efficacy and predictive molecular markers of neoadjuvant gemcitabine and pemetrexed in resectable non-small cell lung cancer.* J Thorac Oncol, 2008. **3**(10): p. 1112-8.
4. Travis, W.D., et al., *Reproducibility of neuroendocrine lung tumor classification.* Hum Pathol, 1998. **29**(3): p. 272-9.
5. Takeuchi, T., et al., *Expression profile-defined classification of lung adenocarcinoma shows close relationship with underlying major genetic changes and clinicopathologic behaviors.* J Clin Oncol, 2006. **24**(11): p. 1679-88.
6. Bhattacharjee, A., et al., *Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses.* Proc Natl Acad Sci U S A, 2001. **98**(24): p. 13790-5.
7. Anbazhagan, R., et al., *Classification of small cell lung cancer and pulmonary carcinoid by gene expression profiles.* Cancer Res, 1999. **59**(20): p. 5119-22.
8. Hsu, D.S., et al., *Pharmacogenomic strategies provide a rational approach to the treatment of cisplatin-resistant patients with advanced cancer.* J Clin Oncol, 2007. **25**(28): p. 4350-7.
9. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-21.
10. Tibshirani, R., et al., *Diagnosis of multiple cancer types by shrunken centroids of gene expression.* Proc Natl Acad Sci U S A, 2002. **99**(10): p. 6567-72.
11. Golub, T., et al., *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science, 1999. **286**: p. 531-536.
12. Giovannetti, E., et al., *Cellular and pharmacogenetics foundation of synergistic interaction of pemetrexed and gemcitabine in human non-small-cell lung cancer cells.* Mol Pharmacol, 2005. **68**(1): p. 110-8.
13. Bild, A.H., et al., *Oncogenic pathway signatures in human cancers as a guide to targeted therapies.* Nature, 2006. **439**(7074): p. 353-7.
14. Rosell, R., et al., *The biology of non-small-cell lung cancer: identifying new targets for rational therapy.* Lung Cancer, 2004. **46**(2): p. 135-48.
15. Hanauske, A.R., et al., *In vitro chemosensitivity of freshly explanted tumor cells to pemetrexed is correlated with target gene expression.* Invest New Drugs, 2007. **25**(5): p. 417-23.
16. Buchholz, T.A., et al., *Global gene expression changes during neoadjuvant*

chemotherapy for human breast cancer. Cancer J, 2002. **8**(6): p. 461-8.
17. Irizarry, R.A., et al., *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* Biostatistics, 2003. **4**(2): p. 249-64.
18. Potti, A., et al., *A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer.* N Engl J Med, 2006. **355**(6): p. 570-80.

Supplementary Table 1. Clinical characteristics of NSCLC patient cohort

Table 1. Characteristics of Patients		
		Tumor (N = 91)
Age-yr		
	Mean	62.84±10.73
Sex-%		
	Female	30
	Male	70
Race-%		
	Caucasian	90
	other	4
	unknown	6
Tobacco history-%		
	None	-
	≤ 30 yr	22
	31-49 yr	19
	≥ 50 yr	18
	unknown	40
Tumor type-%		
	ADC	54
	SCC	33
	LCC	20
	CAR	5
Stage-%		
	I	61
	II	27
	III	9
	IV	4
Status		
	Alive	32
	Deceased	62
	unknown	6
Cause of death		
	Lung cancer	30
	other	18
	unknown	52

Supplementary Table 2. Bioinformatics analysis revealed over-represented biological functions by non-responder classifiers

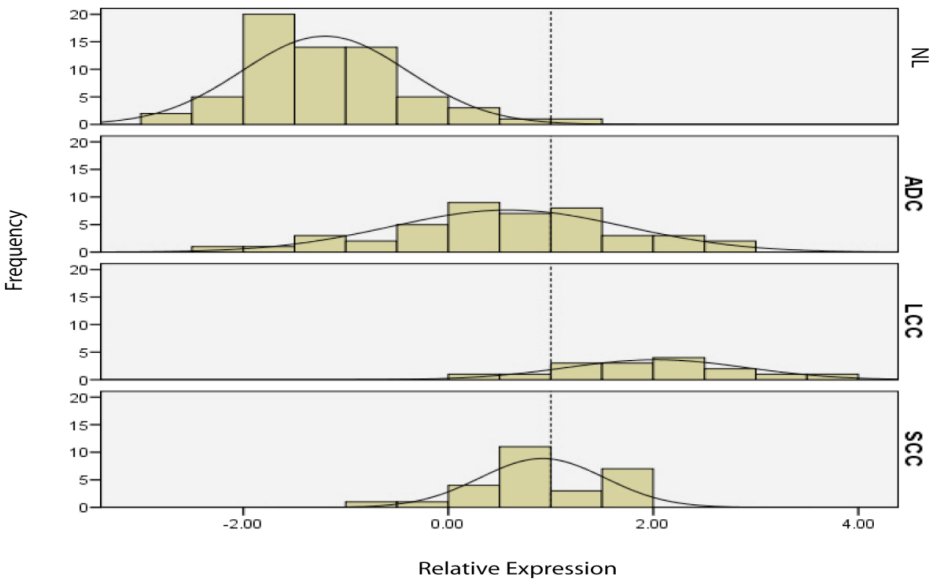
KEGG-pathways	Biocarta-pathways	Ingenuity-canonical pathways
Cell cycle	Role of Ran in mitotic spindle regulation	Cell cycle: G2/M regulation
Pyrimidine metabolism	Cyclins and Cell cycle regulation	Role of BRCA1 in DNA damage response
P53 signaling pathway	Cell cycle: G2/M & G1/S check point	Pyrimidine Metabolism
DNA polymerase	CDK regulation of DNA replication	Cell Cycle: G1/S regulation
Purine metabolism	AKAP95 role in mitosis and chromosome dynamics	Purine metabolism
Folate biosynthesis	Regulation of p27 phosphorylation during cell cycle progression	Folate biosynthesis

Supplementary Table 3. Comparisons of resistance classifiers to other signatures. Numbers in the table are overlapped probesets between two signatures.

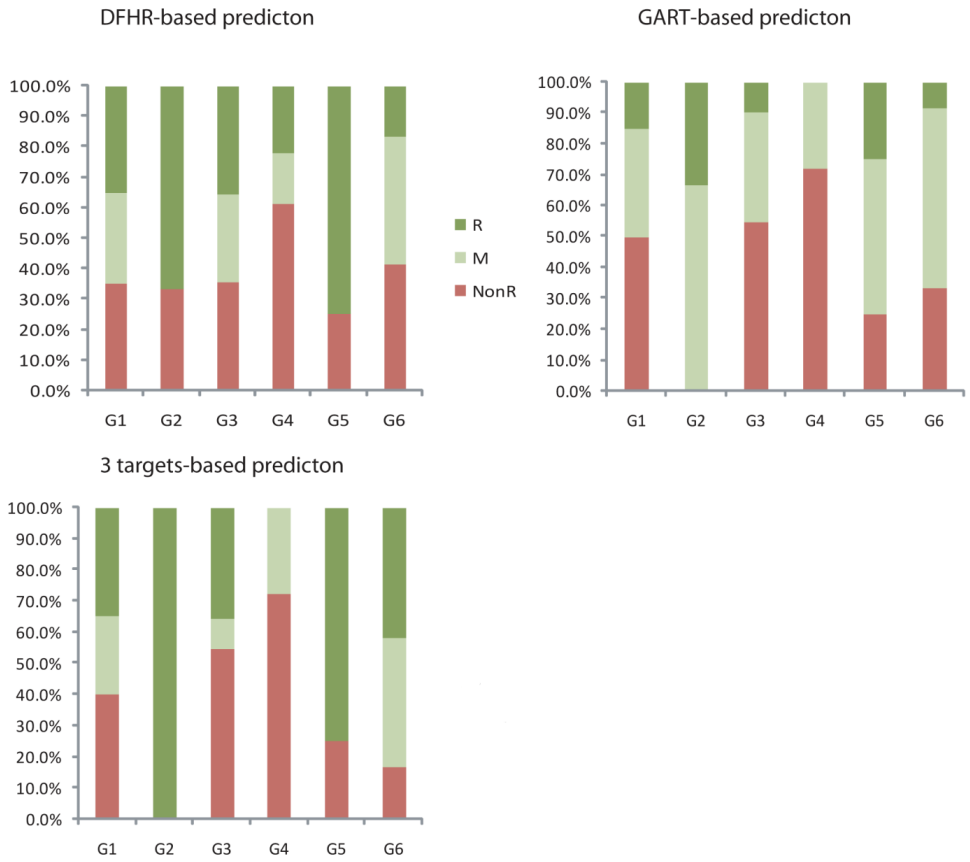
	Resistance Signature_524			
Histology Signature_518	42	(8 : ADCSCC; 34 : LCC)		8.1%
LCC Signature_189	34			18%
Histology Signature_75	0			
Tumor Signature_187	45			24.1%
Tumor Signature_5	2			40.0%
Group Signature_964	51	(44 : G4)		5.3%
Group4 Signature_139	44			31.7%
Survival Signature_17	0			
Duke_Ras signature_209	1			0.5%
Duke_Src signature_31	3			9.7%
Duke_Myc signature_133	8			6.0%
Duke_E2F3 signature_164	25			15.2%
Duke_bCatenin signature_38	0			
KEGG_EGFR pathway_29	0			
KEGG_RAS pathway_143	10			7.0%
KEGG_P53 pathway_79	7			8.9%
KEGG_G1S pathway_30	9			30.0%

Supplementary Table 4. The consistency of gene expression between microarray and qPCR assay in form of Pearson Correlation coefficient (CC)

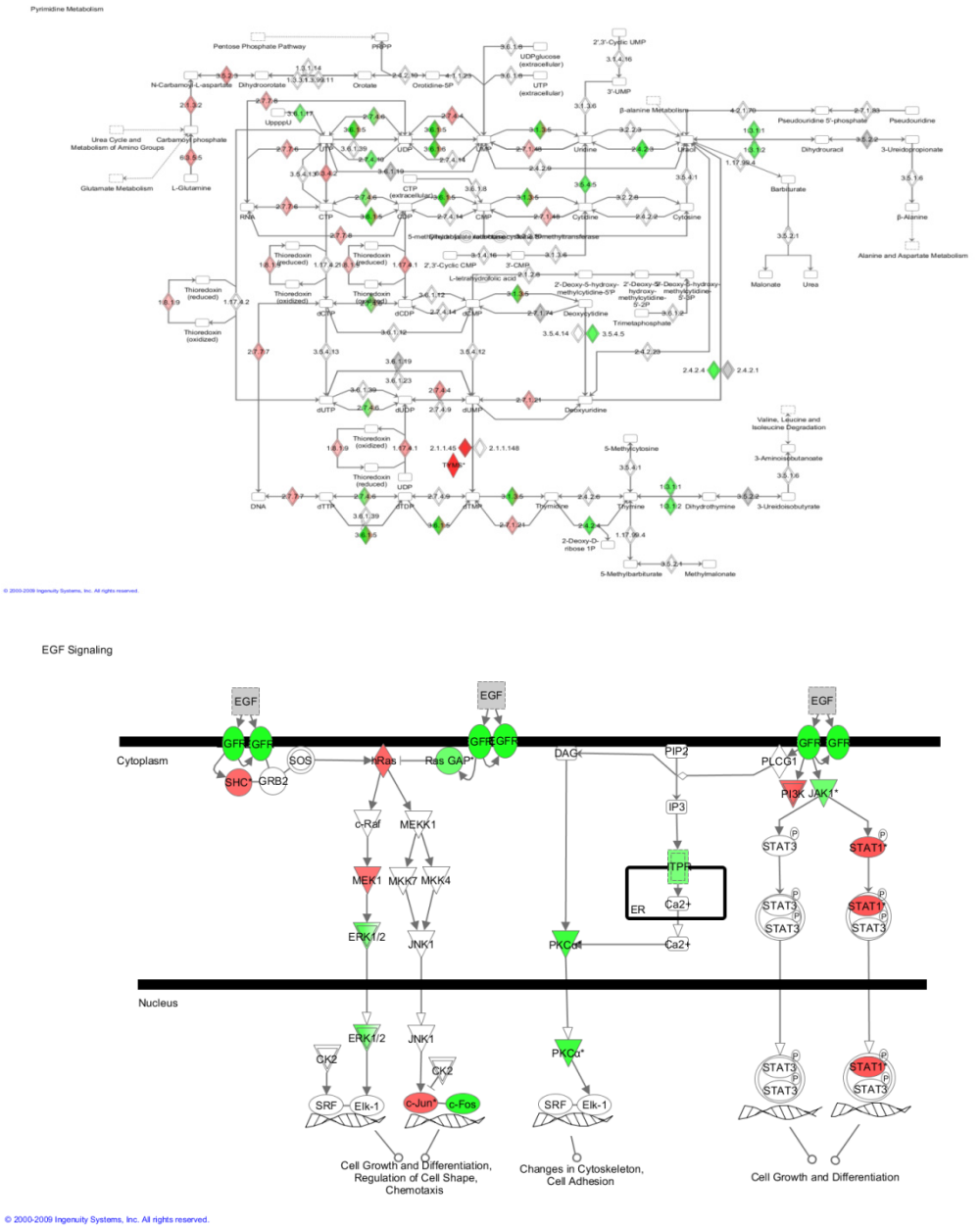
qPCR	microarray	CC
TYMS	1554696_s_at	0.69
	202589_at	0.70
DHFR	202532_s_at	0.87
	202534_x_at	0.84
GART	217445_s_at	0.16
	230097_at	0.04
	average of 2 Probe-sets	
TYMS	TYMS	0.72
DHFR	DHFR	0.85
GART	GART	0.05



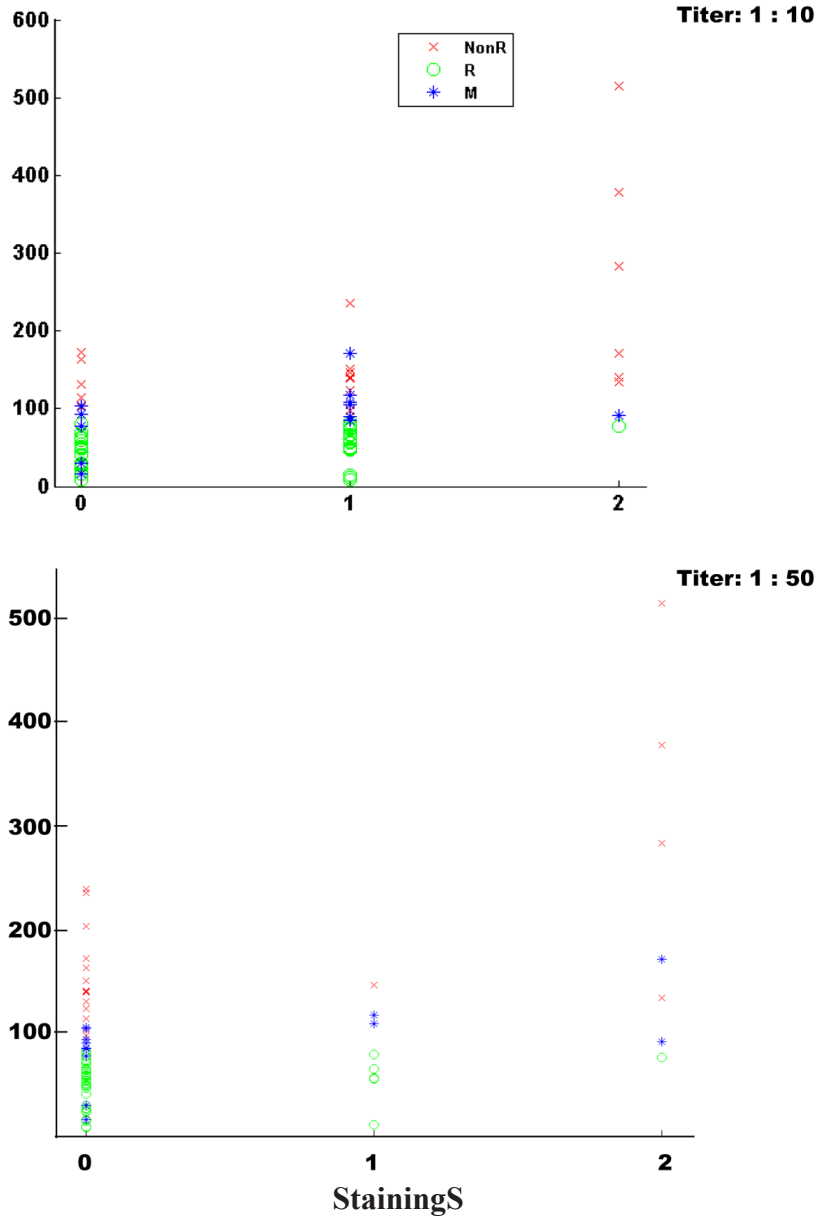
Supplementary Fig. 1. Relative expression of TYMS and distribution of high TYMS tumor in subtypes of NSCLC, adenocarcinoma (ADC), squamous cell carcinoma (SCC), and large cell carcinoma (LCC).



Supplementary Fig. 2. Prediction of NSCLC sensitivity to Pemetrexed using expression of DHFR, GART, or all three targets of Pemetrexed, in correlation to NSCLC novel subgroups. **NR,** resistant NSCLC; **R,** sensitive NSCLC; **M,** medium sensitive NSCLC.



Supplementary Fig. 3. Pyrimidine metabolic pathway (upper) and EGFR signaling pathway (lower) deregulated in a subgroup of NSCLC, which is characterized by putative resistance to Pemetrexed therapy.



Supplementary Fig. 4. TYMS protein staining on TMA was graded from 0 to 2 (StaingS). The TMA staining scores at two different titers (**A,** 1:10; **B,** 1:50) are shown in relation to TYMS mRNA expression on microarray.

Chapter 6

Discussion

Discussion

1) Pharmacogenomics: toward personalized chemotherapy

In **Chapter 5**, we discriminated a subset of NSCLC with respect to putative Pemetrexed resistance from the rest of NSCLC. Instead of common histological characters, this subset of NSCLC is distinct in altered folate-related pathway and tyrosine metabolism. The unique expression pattern of enzymes directly targeted by Pemetrexed and other relevant genes, such as transporter molecules, not only provided a molecular rationale to predict the outcome of Pemetrexed therapy, but also postulated additional targets for other therapeutic agents that might act in combination with Pemetrexed to further improve RR and prognosis for specific subsets of NSCLC patients.

Furthermore, we propose a scheme for the design of personalized therapy on the basis of molecular profiling of individual tumors. The expression measurement of specific genes, such as TYMS, combined with integrated global expression profiling classified NSCLCs into novel groups beyond conventional histopathological characteristics. Tumors of different groups were aggregated by their similar molecular behaviors rather than morphological hallmarks. Accordingly, different therapeutic regimens were proposed for different subgroups of NSCLC. Thus, Group1, Group2 and Group5 identified in this study, regardless of their histology, may receive Pemetrexed-based first-line treatment. For Group4 NSCLC, with strong contra-indication for Pemetrexed therapy and EGFR-TKI/antibody, standard platinum-based regimens might be the best option. SCC is a NSCLC group distinct in pathogenesis and also carcinogenesis. The stratification of SCC with respect of chemoresistance by TYMS and ABCC1 expression in this study suggests that the use of Pemetrexed to treat SCC patients should be re-evaluated.

As suggested by our study, new methodologies should be developed that enable division of NSCLC into molecularly characterized subgroups. These subgroups are anticipated to guide the clinical algorithms used to treat NSCLC patients in the future.

2) Experimental design an interpretation of analysis results

Different experimental set-ups, such as cell type composition of patient samples, attribute to inconsistent results among studies addressing the same questions. For instance, if we see that outcome signatures make predictions beyond histology subtypes in the original data set, it is conceivable that genes associated with prognosis are still histology-related. When these signatures are applied to other sample sets with different histology composition, they do not necessarily reflect hazard of recurrence or chance of survival.

In Chapter 4, we validated published prognostic signatures with our NSCLC Cohort and the Duke NSCLC Cohort. It is noticeable that a signature developed with ADC samples remained histology-dependent, and failed to stratify mixed NSCLC samples from both cohorts. It performed satisfactorily, however, on the ADC samples from these cohorts which were assigned by the histology gene signature developed also in this study. Clearly, it is essential to test the current prognostic classifiers further with more independent datasets containing a broad histo-pathological spectrum of NSCLC cases.

3) Gene signatures versus conventional factors

Expression profiling can provide improved tumor classification and outcome prediction compared to traditional clinical and histo-pathological assessments. However, gene expression profiles to a great extent reproduce histopathologcal characteristics. Considering this, it is likely that the predictions derived from gene signatures should also take conventional factors into account (smoking history, tumor stage, grade, size, and cell type).

It remains challenging to determine on which aspects expression profiling should focus. For questions which could be answered already with high accuracy by conventional examinations, with gene-based classifiers or predictors merely providing the same information, there is no need to perform relatively expensive microarray analysis. Expression profiling is better suited to answer questions that challenge conventional histo-pathological methods, such as identifying markers that provide prediction values beyond conventional factors or molecular predictors that explain variable clinical outcomes among

patients with similar histo-pathological features.

4) Pathway analysis may overcome some of the inherent shortfalls of gene expression analysis

Like many other methodologies, gene expression profiling is a snapshot of a tissue at a certain time point and within a certain *in vivo* environment. Current bioinformatics algorithms favor genes with high expression or highly differential expression. As a result, genes associated with obvious phenotypes or highly expressed when the tissue is isolated may overwhelm those genes with low-level expression or associated with subtle expression differences. Such genes might be relevant to specific clinical questions.

It easily explained why cell cycle- and cell-type associated genes are predominantly found in phenotypic comparisons, while metastasis-associated genes are rarely picked out. Furthermore, metastasis genes are more likely recruited as determinants for cell differentiation since fast growth and early metastasis are hallmarks of poorly differentiated tumors. Therefore, such genes usually appear in histologicy signatures and are not easily associated with metastasis.

Knowledge-based analyses, such as gene set enrichment analysis, make it possible to gain further insight into genes that do not show dramatic changes under different conditions. This type of approach is based on prior biological knowledge, and focuses on predefined gene sets, such as genes involved in a given biochemical pathway, so that the analysis is performed under the guidance of biological relevance. By applying this methodology, profiling analyses are less prone to magnitude-derived bias.

5) Lack of overlap between different tumor/survival signatures

The lack of overlap between different signatures can be explained by the differences between patient populations with respect to ethnic background, tumor heterogeneity, and environmental influence.

Other possible explanations include the different bioinformatics approaches utilized by different studies which may also attribute to this diver-

gence. For instance, validation of our prognostic signature with the Duke Cohort failed when the microarray data was processed and normalized together with those from the EMC cohort. In contrast, the signature was able to stratify NSCLC patient outcome when the Duke Cohort was processed independent of the EMC Cohort.

In another effort in this study to validate the prognostic signature, the raw microarray files (.CEL) of a Korean Cohort were not accessible. The normalized microarray data deposited in the GEO database was downloaded and directly fitted to our prognostic model. The normalization utilized for the data provided was GCRMA - a modified RMA with MM signals included in the calculation. This leads to much more information retained at probe-set level compared to RMA normalized data used in our study. Thus, such variations in data processing obscure relevant information and obstruct comparative analyses.

There are other technical issues responsible for inconsistent results between different studies, such as microarray platforms used and sample-processing techniques. To make microarray results comparable across independent studies, it is therefore still needed to further develop standardized experimental protocols and more robust data analysis pipelines.

6) Microarray studies are prone to different bioinformatics approaches

Except for underling biological variations among microarray experiment sets, microarray data sets might differ from each other as a result of data preprocessing, normalization, and statistic analysis.

Although a large number of publicly accessible analytic tools for microarray data exist to date, a ‘gold standard’ for dealing with microarray data is still lacking. Most microarray experiments follow a step-wise analysis, including data processing, class comparison, class prediction, clustering, and pathway analysis. Each single step, however, can be accomplished by diverse algorithms, software, and statistical analyses. Even for datasets undergoing comparable processing and analytical steps, the different adjustments on statistical significance and selection criteria will contribute to different outcome

gene lists.

There are no definitive answers to the best procedures that should be followed to address these issues. Instead, depending on the specific questions addressed in a study, some algorithms or analyses will be superior to others.

In this study, the sample dataset was subjected to SAM analyses integrated in different software to answer the same questions. The best overlap between two generated gene lists was 70%. Another example, in a predictive classifying analysis the optimized gene set generated by PAM is only half of the gene set produced by correlation analysis followed by a regression model. The reason is that the PAM algorithm excludes genes negatively correlated to a phenotype in building a predictive expression matrix.

7) Meta-analysis of expression microarray data

Meta-analysis (analysis of the analyses) has superiority in increasing sample size and thus statistical power, enabling generalization of the conclusions drawn. However, cross-study analysis of microarray data is complicated by variations in the types of array platform, the format of raw data, the number of genes studied, the design and nomenclature of representative probes, and the analytical methodologies. It is not uncommon that the predictor or signature genes identified in original studies performed poorly when applied to datasets independently collected at other institutes. Sun et al. identified two sets of prognostic genes for ADC and SCC respectively by using gene expression profiling [1]. However, they did not perform satisfactorily in two independent patient cohorts. Some studies reselected outcome classifiers from published microarray analyses. The classifier genes were validated by RT-PCR on independent samples, but their predictive abilities were not great [2-4]. Such discrepancies cannot be totally removed by simple data re-processing and re-normalization.

Many algorithms and statistical methods have been developed to solve this issue. However, none of these performs sufficiently well to be used as a standard for the inter-group/-platform studies aimed at integrating multiple array datasets. Sohal et al. showed that when unsupervised clustering was performed

using multiple microarray datasets, samples were grouped primarily based on the data sources or microarray platforms instead of biological status. However, the gene expression signature defining a certain phenotype could distinguish different biological tissues even in the presence of study-specific bias, as described in **Chapter 4** (Supplementary Fig.1), indicating that the differential biological processes represented by selected genes was still retained despite the systematic inter-study variability [5].

8) Systematically studying NSCLC carcinogenesis

Gene expression profiling produces a snapshot of global expression at a certain time point, however, it is unable to decipher the regulatory mechanisms and regulatory networks as underlying gene expression determinants. To answer such questions, it is necessary to integrate high-throughput data from different levels, including genomic profiling such as single nucleotide polymorphisms (SNP), transcriptomics, and DNA methylation by high-throughput sequencing, and proteomics.

Technically similar to expression microarray, Affymetrix SNP arrays probes up to 906,600 SNPs and 946,000 copy number variation (CNVs). It is able to get insight into the genotype of cancer cells, revealing common or rare SNPs, and chromosomal abnormalities such as gains or losses. A study published recently linked a small number of CNV and SNPs, including well-known 3q and SNPs located in STK39, with overall survival of early stage NSCLC, and the association was retained in two validation cohorts [6]. By integrating SNP array data with gene expression profiles, the effect of DNA changes on mRNA transcription can be assessed and it might provide more reliable predictors for cancer patient outcome. Under this hypothesis, Broet and his colleagues developed a methodology to integrate CGH data and expression profiles. The application of this methodology identified around 100 genes located within 4 chromosomal regions. The predictive performance of those integrated markers for the risk of early relapse was more robust than the performance of either genomic or transcriptomic predictors alone in NSCLC [7].

A direct assay to study the interaction of regulatory factors and mRNA transcription is more and more widely used, that is Chromatin immunoprecipitation (ChIP)-based high-throughput technique. This technique makes use of the interaction of protein-DNA to assess the binding (sites) of particular transcription factors to DNA. ChIP technique is enhanced by combining with microarray technology, termed as ChIP-on-chip and allowing the identification of genome-wide binding sites for a certain protein. Alternative application of ChIP-on-chip is to study the role of promoter methylation in carcinogenesis. In this case, DNA regions containing methylated cytosine residues are recognized and precipitated by antibodies specific to meC. The resulting DNA fragments are poured over an array, which is spotted with a large number of probes tiling promoter regions – 10 kb adjacent to transcription start sites taking Affymetrix chip as an example. By comparing to the output fragments from normal tissues, the cancer specific methylation within promoter CpG islands can be portrayed. A further development of ChIP-based technique is termed as ChIP-seq, the combination of ChIP and high-throughput sequencing. A prominent improvement of ChIP-seq compared to ChIP-on-chip is the unbiased whole-genome identification of protein-DNA interactions or methylated CpG islands. By contrast with ChIP-on-chip, sequencing requires no hybridization arrays and therefore is not restricted by the probes spotted on the array. All precipitated DNA fragments are sequenced followed by alignment to the whole-genome sequence. Using ChIP-seq analyses in conjunction with whole-genome expression profiling is a promising methodology to get a closer insight into transcription factor networks and epigenetic modifications underlying biological processes and diseases; to understand how epigenetic abnormalities trigger and promote carcinogenesis through regulating gene expression; and to discover the most reliable biomarkers for early detection and prognosis prediction of cancer.

Other levels of high-throughput analyses potentially to be integrated with expression profiles include exon or whole-transcript arrays targeting all alternative splicing variants of a gene; mciroRNA profiling to reveal mechanisms of mRNA processing and translation regulation; and mass-spectrometry to

assess active complexes at protein level in cancer cells.

The most challenging task currently is how to combine massive experimental data from different levels and to single out biologically relevant information in order to interpret cancer-associated molecular profiles at systems level. To solve these problems, intensive knowledge of biology, oncology, bioinformatics, database, mathematics, statistics, are required and need to be used in combined approaches.

References:

1. Sun, Z., D.A. Wigle, and P. Yang, *Non-overlapping and non-cell-type-specific gene expression signatures predict lung cancer survival.* J Clin Oncol, 2008. **26**(6): p. 877-83.
2. Lau, S.K., et al., *Three-gene prognostic classifier for early-stage non small-cell lung cancer.* J Clin Oncol, 2007. **25**(35): p. 5562-9.
3. Raz, D.J., et al., *A multigene assay is prognostic of survival in patients with early-stage lung adenocarcinoma.* Clin Cancer Res, 2008. **14**(17): p. 5565-70.
4. Skrzypski, M., et al., *Three-gene expression signature predicts survival in early-stage squamous cell carcinoma of the lung.* Clin Cancer Res, 2008. **14**(15): p. 4794-9.
5. Sohal, D., et al., *Meta-analysis of microarray studies reveals a novel hematopoietic progenitor cell signature and demonstrates feasibility of inter-platform data integration.* PLoS ONE, 2008. **3**(8): p. e2965.
6. Huang, Y.-T., et al., *Genome-Wide Analysis of Survival in Early-Stage Non-Small-Cell Lung Cancer.* J Clin Oncol, 2009. **27**(16): p. 2660-2667.
7. Broet, P., et al., *Prediction of clinical outcome in multiple lung cancer cohorts by integrative genomics: implications for chemotherapy selection.* Cancer Res, 2009. **69**(3): p. 1055-62.

Summary

Lung cancer is the most common cause of cancer death worldwide. The difficulty to detect lung cancer at early stages with current techniques is believed to be a major reason for this high mortality rate. Another important factor is that patients with advanced NSCLC respond highly variable to the same treatment. To improve NSCLC-related clinical outcome, it is important to identify highly specific markers for early diagnosis, to classify tumors with different biological behaviors, and to develop accurate predictive models for disease progression. It is also highly desirable to understand the molecular mechanisms underlying chemotherapy resistance.

We aimed to address some of these issues by gene expression profiling of tumors and adjacent non-cancerous lung tissues.

In Chapter 1, a general introduction to NSCLC is given. Relevant information includes lung cancer epidemiology, tumor histology, and clinical pathology. Subsequently, prognosis and present therapeutic regimens are discussed. From this chapter, the reader will have an overview of diagnosis and classification of lung cancer in clinical practice; the expected outcome of NSCLC, and which possible treatments the patients could receive.

Chapter 2 discusses the variety of factors contributing to lung carcinogenesis. The acquired abnormalities drive respiratory epithelia to deviate from normal cell growth and differentiation fate, leading to initiation of cancer cell transformation. Critical alterations discussed in this chapter include genetic instability at Chr. 3p and Chr. 9p, and aberrant expression of oncogenes (e.g. RAS, EGFR) and tumor suppressor (e.g. TP53, RB).

The technological and analytic methods employed in this study are reviewed in Chapter 3. Firstly, the general principles of DNA microarray technology are briefly described, taking the Affymetrix GeneChips as an example. This is followed by a description of microarray data analysis at different levels and for different purposes. The broad spectrum of applications of gene expression profiling in oncology is discussed.

Chapter 4 reports the identification of new diagnostic, classification, and prognostic gene signatures for NSCLC. We show that NSCLC tumors can

be molecularly distinguished from normal lung tissue by the expression of only five genes. The heterogeneity of NSCLC is investigated by comparing the expression profiles of the three main subtypes – ADC, SCC, and LCC. The dominant properties of these three main histological categories are summarized molecularly by the expression of seventy-five genes, including members of the keratin- and kinesin families. We also find that the overall survival of NSCLC patients, regardless of histological subtypes, can be predicted by the expression of seventeen genes. The established association between the gene expression profiles described in this study and NSCLC properties is retained when tested in an independent NSCLC cohort from Duke University. We propose that these newly identified NSCLC expression profiles could be used as biomarkers to distinguish NSCLC from non-cancerous cases at early stages, to recognize the predominant molecular attributes of individual NSCLC cases, and to predict overall survival of the patients.

In Chapter 5, I report on a study aiming to predict NSCLC sensitivity to Pemetrexed therapy using gene expression profiles. A novel subgroup of NSCLC is correlated to predicted Pemetrexed resistance based on the expression level of TYMS, a major target of Pemetrexed. This novel subgroup shows no association with routine histopathology. Distinctive characteristics are the presence of neuroendocrine tumor components and a deregulated EGFR signaling pathway. We propose that biological characteristics assessed by gene expression profiling may aid in reliably stratifying patients with respect to the efficacy of Pemetrexed therapy. We also suggest an approach applicable in clinical practice that is based on genes whose expression correlates well with that of TYMS. Such surrogate markers are potentially useful since it is notoriously difficult to assess TYMS expression levels by routine histopathological techniques.

Chapter 6 is a general discussion about the studies carried out in this thesis and focuses on current issues that need to be tackled to improve future applications of bioinformatics analyses in oncology.

Samenvatting

Longkanker is wereldwijd de meest voorkomende doodsoorzaak ten gevolge van kanker. Men denkt dat deze hoge sterftecijfers mede worden veroorzaakt doordat vroege detectie van longkanker niet goed mogelijk is met de huidige technieken. Een andere belangrijke factor is dat patiënten met gevorderde longkanker een zeer variabele response vertonen op dezelfde behandeling. Om dit te verbeteren is het belangrijk dat specifieke merkers geïdentificeerd worden die de vroege diagnostiek verbeteren. Daarnaast is het wenselijk om tumoren te classificeren zodat hun biologisch gedrag voorspeld kan worden. De ontwikkeling van accurate modellen om de voortgang van longkanker te voorspellen staat nog in de kinderschoenen. Om de kans op slagen van de behandeling te verhogen is het nodig om de moleculaire mechanismen die resistentie tegen chemotherapie veroorzaken beter te begrijpen.

In het werk beschreven in dit proefschrift hebben wij een aantal van deze onderwerpen benaderd door genexpressieprofielen van tumoren en daarnaast gelegen gezond longweefsel te bepalen.

Hoofdstuk 1 is een algemene inleiding over longkanker. De epidemiologie van longkanker, de histologische karakteristieken van de tumoren, klinische pathologie, en de huidige prognoses en therapeutische benaderingen passeren de revue. Er wordt een overzicht gegeven van de diagnose en classificatie van longkanker in de klinische praktijk, de vooruitzichten voor de patiënten, en welke mogelijkheden voor behandeling er zijn.

In Hoofdstuk 2 wordt de diversiteit aan factoren die een bijdrage kunnen leveren aan carcinogenese van de long besproken. De verworven afwijkingen zorgen ervoor dat het longepitheel van het normale pad van celgroei en differentiatie afdwaalt, wat kan leiden tot de initiatie van de transformatie naar tumorcellen. Onder de kritische veranderingen die hier beschreven worden bevinden zich instabiliteit van chromosoom 3p en 9p, en afwijkende expressie van oncogenen (bijvoorbeeld RAS en EGFR) en tumorsuppressorgenen (bijvoorbeeld TP53 en RB).

De technische en analytische methodes die bij onze studies zijn gebruikt zijn het onderwerp van Hoofdstuk 3. De algemene principes van DNA micro-array technologie worden kort beschreven, waarbij de Affymetrix GeneChips als voorbeeld worden gebruikt. Dit wordt gevolgd door een beschrijving van de analyse van micro-array data, op verschillende niveaus en voor verschillende doeleinden. Het brede spectrum van toepassingen van genexpressieprofielen in de oncologie wordt belicht.

In Hoofdstuk 4 beschrijven we de identificatie van nieuwe diagnostische, classificatie en prognostische genexpressiesignaturen voor longkanker. We laten zien dat de expressie van slechts 5 genen voldoende is om longtumoren

moleculair te onderscheiden van normaal longweefsel. De heterogeniteit van longkanker werd onderzocht door de genexpressieprofielen van de drie belangrijkste subtypes niet-kleincellige longkanker met elkaar te vergelijken. De dominante eigenschappen van deze drie belangrijkste histologische categorieën kunnen moleculair worden samengevat door de expressieprofielen van vijfenzeventig genen, waaronder leden van de keratine- en kinesine genfamilies. We vinden ook dat de overlevingstijd van de patiënten voorspeld kan worden aan de hand van de expressie van zeventien genen. De associaties tussen deze genexpressieprofielen en de eigenschappen van de longtumoren hielden stand als ze getest werden op een onafhankelijk cohort longtumoren verzameld door de Duke University in de Verenigde Staten. We stellen voor dat deze nieuw geïdentificeerde genexpressieprofielen gebruikt kunnen worden als biologische merkers om al in vroege stadia longkanker van gezond weefsel te onderscheiden, om de dominante histologische kenmerken van individuele longtumoren eenduidig vast te stellen, en om de overlevingstijd van longkankerpatiënten te voorspellen.

In Hoofdstuk 5 beschrijf ik een studie gericht op het gebruiken van genexpressieprofielen voor het voorspellen van de gevoeligheid van longtumoren voor behandeling met Pemetrexed. We identificeren een nieuwe subgroep tumoren die volgens de voorspelling resistent zijn tegen Pemetrexed. Dit is gebaseerd op de expressie van het enzym thymidylaat synthase (TYMS), een belangrijke target van Pemetrexed. Deze nieuwe subgroep kan niet geclassificeerd worden met behulp van routine histopathologie. Onderscheidende karakteristieken zijn de aanwezigheid van neuro-endocriene componenten, en deregulatie van signaaltransductie via de epidermale groeifactor receptor (EGFR). We stellen voor dat dergelijke biologische karakteristieken, gereflecteerd in specifieke genexpressieprofielen, gebruikt kunnen worden om patiënten die in aanmerking komen voor Pemetrexed therapy te selecteren. We doen een suggestie om voor de diagnostiek de expressie van genen te gebruiken die nauw gecorreleerd zijn met TYMS expressie. Dergelijke surrogaat markers zijn potentieel heel nuttig, omdat het vrijwel onmogelijk is om TYMS expressieniveaus betrouwbaar te bepalen met routinematige histopathologische technieken.

Hoofdstuk 6 is een algemene discussie over de studies die beschreven worden in dit proefschrift. De meeste aandacht gaat uit naar de huidige beperkingen van genoom-brede analyses, en de uitdaging om daar oplossingen voor te vinden die in de toekomst toegepast kunnen bij bioinformatica analyses in de oncology.

Curriculum Vitae

Personal information:

Name: Jun Hou
Born: 27 Nov, 1977, Lanzhou, China

Education:

2002 – 2004	M.Sc., Bioinformatics, Wageningen University, the Netherlands
2000 – 2002	M.Sc. (*combined with Residency*), Department of Gynecology & Obstetrics, LanZhou University Medical School, China
1995 – 2000	M.D., Department of Clinical Medicine, LanZhou University Medical School, China

Research and professional experience

2004 – present	Ph.D. student at the Department of Cell Biology, Erasmus MC Rotterdam, the Netherlands
2003 – 2004	MSc research project at the Center for Human and Clinical Genetics Leiden University Medical Center, the Netherlands

Other Experience:

2000 – 2002	Resident, Department of Gynecology & Obstetrics, the 1st affiliated hospital of Lanzhou University, China.
1999 – 2000	Intern, the 1st affiliated hospital of Lanzhou University Lanzhou, China.

List of publications:

1. van Loo PF, Mahtab EA, Wisse LJ, **Hou J**, Grosveld F, Suske G, Philipsen S, Gittenberger-de Groot AC. Transcription factor Sp3 knockout mice display serious cardiac malformations. *Mol Cell Biol.* 2007 27:8571-8582.

2. Soler E, Andrieu-Soler C, de Boer E, Bryne JC, Thongjuea S, Stadhouders R, Palstra RJ, Kockx C, van IJcken W, **Hou J**, Steinhoff C, Rijkers E, Lenhard B, Grosveld F. The Genome Wide Dynamics of the Binding of the Ldb1 Complex during Erythroid Differentiation. *Genes Dev.* 2010, accepted for publication.

3. **Hou J**, Aerts J, den Hamer B, van IJcken W, den Bakker M, Riegman P, van der Leest C, van der Spek P, Foekens JA, Hoogsteden HC, Grosveld F, Philipsen S. Expression-based classification of non-small cell lung carcinomas and survival prediction. 2010, Manuscript submitted for publication.

4. **Hou J**, Lambers M, Hughes Carvalho R, den Hamer B, Riegman P, van IJcken W, den Bakker M, van der Spek P, Hoogsteden HC, Hegmans J, Grosveld F, Aerts J, Philipsen S.Expression profiling-based prediction of the putative response of NSCLC patients to Pemetrexed therapy. 2010, Manuscript in preparation.

5. Patent: **J. Hou**, J.G.J.V. Aerts, F.G. Grosveld, J.N.J. Philipsen. Tumor gene profile. GB0904957.8. April 2009.

PhD Portfolio Summary

Summary of PhD training and teaching activities

Name PhD student: Jun Hou Erasmus MC Department: Cell Biology Research School: Postgraduate school Molecular Medicine	PhD period: June 2004 – January 2010 Promotor(s): Prof.dr.J.N.J.Philipsen Prof.dr. F.G. Grosveld Co-promotor: M.D. PhD J.G.J.V.Aerts

1. PhD training

	Year	ECTs
General academic skills		
- Save laboratory techniques	2004	0.3
- Molecular and Cell Biology. Rotterdam	2005	6
- Basic and Translational Oncology. Leiden	2006	1.5
- Biomedical Research Techniques. Rotterdam	2006	1
- English Biomedical Writing and Communication. Rotterdam	2008	4
In-depth courses		
- Analysis of microarray gene expression data. Leiden	2005	1.2
- Nexus Training Course. Rotterdam	2008	0.6
- Applied Bioinformatics 'Finding your way in biological information'. Rotterdam	2009	0.6
- Next-generation Sequencing data analysis. Leiden	2009	1
- MathWorks training on Fundamentals and Statistical Methods (Matlab). Rotterdam	2009	1
Poster / Presentations		
- Winter school Klein Walsertal	2005	1.5
- 7th Joint Medisch Genetisch Centrum-Cancer Research UK Graduate Student Conference, Cambridge, UK	2006 2007	1 0.3
- Bridge meetings on bioinformatics. Rotterdam		

Conferences		
- CGC/CBG meeting 'Molecular Mechanisms and Mouse Models in Cancer. Amsterdam	2008	0.3
- NCSB Kick-Off Symposium. Noordwijkerhout	2009	0.3
Seminars and workshops		
- Annual MGC-Symposium (MGC- Medisch-Genetisch Centrum Zuid-West Nederland)	2004 - 2008	1.2
- 6th Joint Medisch Genetisch Centrum-Cancer Research UK Graduate Student Conference, Liege, Belgium	2005	1
- Writing successful grant proposals. Rotterdam	2009	0.3

Acknowledgements

To be top in science and medicine would never be the first and sole goal in my life. During my five (+) years in Rotterdam, I not only developed a research interest, but more importantly ascertained which personalities I did encourage and will still head for. I would like to acknowledge everyone who played a role in my professional and personal development.
First of all, I would like to thank my promoters Prof.dr. Sjaak Philipsen and Prof.dr. Frank Grosveld for offering me the opportunity to work as a PhD student at the department of Cell biology, Erasmus MC, Rotterdam.
Sjaak, I cannot find a proper word to express my gratitude to you. You have being patiently guided and seen me through to the end of my thesis, accepted my successful and unsuccessful trials and adaptations to be a PhD student. You are more than supportive and advisory as a promoter. What I learnt from you is not limited to science. The more cherished by me is how to be a modest and understanding person.
I also would like to express my gratitude to Prof.dr. Frank Grosveld for supporting my work in your department and continuous instruction to my project.
My copromotor Joachim, you always direct me to arise clinic relevant questions and to adjust my studies to be more patient favored. I benefit a lot from you insightful comments on my work and my thesis. You thinking all about patients remind me what a real doctor is in my deep hearts. I respect you as a bridge between clinics and me.
Many thanks to my reading Committee, Peter, Guido, and Els who reviewed my thesis at short notice. Your comments improved the manuscript a lot. And special thanks to your valuable advice on my future career.
I would like to thank the other members of my Committte, for taking time to read my thesis and to review my work.
Eva and Sahar, my real friends and paranimfen. Thank you for sharing everything with me, happiness and sadness, and for supporting me in my defense. Eva, none else except you can understand my willingness to work on relieving patients, and also my love to dogs. Sahar, your great courage and maturity should never be underestimated. I learnt a lot from you. And I am also trying to follow you to work harder and harder. It is great to experience PhD period along the way with all of you.
Bianca, without you, I could not finish the work in this thesis. It is lucky to spend my first few years with you. You are not only a colleague to me. I will always remember the good time we spent together and of course your help

with work and others. Wish you all the best.
Margaretha and Joost, thank you for your great collaboration. Your excellent expertise makes Alimta study complete with beautiful IHC.
I would like to acknowledge Prof.dr. Micheal den Bakker. You put so much effort into TMA assay and are always willing to have discussions and to answer my questions.
Wilfred, Christel, and Zeliha, your efficiency and professional assistance are highly appreciated.
Cor, thank you for your support and useful discussion. Without clinical information you provided I cannot perform any analyses. Rejane, thanks for your hands with PCR.
Nynke, you are the most generous people I have seen. Your help to other people is ever-unconditional. You deserve everything good including a so lovely husband, Ton. Wish you success in each trip around the world.
Mirijam, Anton, and all people from Bioinformatics, your friendly support made my work much easier.
I would also mention my colleagues and ex-colleagues. Rita, although you left shortly after I started in this group, but I would never forget your kindness to a freshman who was in a bit fear in a new environment. Thamar, although we are not working closely together, but I respect your brilliant expertise and knowledge, and also your effort to all your students. From Laura, I could learn to be strong and to persist in my own beliefs. It is a pleasure to be a group with you: Roy, Harald, Teus, Sylvia, Pavlos, Divine, and Anna.
I also learnt from Charlotte and Eric, your 120% dedication to work is always an example to me. Athina and Umut, thanks for conversations about your work with me, it is glad to work with you.
Marike and Jasperina, thank you for enormous help from my first day in Rotterdam.
江涛和石莹，很难得在这里能和你们像家人一样一起分担生活中的喜怒哀乐。天娜，吕鹏，陈韬，余晓，童淼，刘凡，你们予我亦师亦友。谢谢与你们一起的快乐时光，谢谢天娜在答辩时支持我。
‘谁言寸草心，报得三春晖’，亲爱的爸爸妈妈，以及我最爱的亲人们，你们的爱和支持是我最宝贵的所有，而有一天能成为你们的骄傲永远是我的目标；我的旺财和球球，爱你们和你们的依赖，是我每天的动力；商鹏，能坚持到今天，是因为有你在我身边 -- 死生契阔，与子成说；执子之手，与子偕老。

Jun / 珺