Molecular profiling of non-small cell lung carcinomas
A genome-wide DNA methylation analysis

Rejane Hughes Carvalho

ISBN: 978-90-8570-975-6

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

The research described in this thesis was financially supported by the Netherlands Genomics Initiative (NGI) through the Cancer Genomics Centre (CGC).

Printed by: Wöhrmann Print Service, Zutphen

Cover and layout: R. Hughes Carvalho

Molecular Profiling of Non-small Cell Lung Carcinomas A Genome-wide DNA Methylation Analysis

Moleculaire profielen van het niet-kleincellig longcarcinoom
Een genoom-brede analyse van DNA methylering

Thesis

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

Prof.dr. H.G. Schmidt

and in accordance with the decision of the Doctorate Board

The public defence shall be held on
Wednesday, 22 February at 11:30 hours

by

Rejane Hughes Carvalho

Born in Salvador, Brazil



Doctoral Committee

Promoter:	Prof.dr. J.N.J. Philipsen
Other Members:	Prof.dr. R.W. Hendriks
	Prof.dr. D.N. Meijer
	Prof.dr. H.R. Delwel
Co-promoter:	Dr. J.G.J.V. Aerts

To my mom

It does not matter where, nor does when, your presence will always be a constant in my life. For your example of strengh and courage, your teaching and encouragments, your dedication and love. I own to you everything I have achieved, everything I will accomplish, everything I am – Today and evermore you are with me.

Contents

Abberviations 09
Preface and scope of the thesis 11
Chapter 1: Introduction **14**
Non-small cell lung carcinomas 14
NSCLC: A complicated name for a complex disease 15
Adenocarcinomas 16
Squamous cell carcinomas 17
Large cell carcinomas 17
NSCLC: molecular mechanisms 18
Environmental factors 18
Genetic factors 20
EGFR and KRAS 20
P53 22
Other common and less common genetic aberrations 22
Epigenetic factors 23
DNA methylation as epigenetic regulatory mechanism **23**
Epigenetics 23
MicroRNAs 24
Histones: covalent and non-covalent modifications 27
DNA methylation **28**
DNA methylation in normal development 29
DNA methylation in cancer 29
DNA methylation in NSCLC 32
DNA methylation as biomarker 34
Epigenetic therapies and prospects for NSCLC 35
Profiling DNA methylation 36
Methylation-specific PCR (MSP) 37
Bisulfite sequencing 38
Unveiling the epigenome: genome-wide methylation profiles 39
Chapter 2: Genome-wide methylation profiling of NSCLC **59**
Chapter 3: Novel methylated regions for use as NSCLC epigenetic tumor markers **95**
Chapter 4: General discussion and prospects **133**
Chapter 5: Addendum **145**
Summary/samenvating 146
Curriculum vitae 150
PhD portfolio 151
Acknowledgment 153

Abbreviations

General

5-azaC	5-azacytidine
5-aza-dC	5-aza-2'-deoxycytidine
5-mC	5-methylcytosine
ABI	Applied biosystems
ADC	Adenocarcinoma
ASR	Age-standardized rates
ATS	American Thoracic Society
AUC	Area under the curve
BAC	Bronchioloalveolar carcinoma
BISMA	Bisulfite sequencing DNA methylation analysis
CF	coupling factor
CIMP	CpG island methylator phenotype
COPD	Chronic obstructive pulmonary disease
DMR	Differentially methylated region
DNA	Deoxyribonucleic acid
dNTPs	Deoxyribonucleotide triphosphate
ERS	European Respiratory Society
ES cells	Embryonic stem cells
ETS	Environmental tobacco smoke
FDA	Food and drug administration
GA	Genome analyzer
GWMP	Genome wide methylation profile
HGNC	HUGO Gene Nomenclature Committee
HTS	High throughput sequencing
IASLC	International Association for the Study of Lung Cancer
IHEC	International Human Epigenome Consortium
LCC	Large cell carcinoma
LINE	Long interspersed nuclear element
LOH	Loss of heterozygosity
MBD	Methyl-binding domain
MeDIP	Methyl-DNA immunopreciptation
MEDIPS	methylated DNA immunoprecipitation statistics
MethyCap	Methyl-DNA Capture
miRNA	Micro-RNA
MSP	Methylation-specific PCR
MZ	Monozygotic twins
NGS	Next generation sequencing
NIH	National Institute of Health
NSCLC	Non-small cell lung carcinoma
NSCLC-NOS	NSCLC Non-otherwise specified
PAM	Partitioning Around medoids
PASW	Predictive analytics software
PCR	Plymerase chain reaction
QUMA	Quantification tool for methylation analysis
rms	Relative methylation score
RNA	Ribonucleic acid
SBL	Sequencing by ligation
SBS	Sequencing by synthesis
SCC	Squamous cell carcinoma
SCLC	Small cell lung carcinoma
SMART	Sensitive melting analysis after real time
SNP	Single nucleotide polymorphism
SVM	Support vector machines
TKI	Tyrosine kinase inhibitors
TSG	Tumor suppressor genes
TSS	Transcription starting sites
UCSC	University of California, Santa Cruz
UUD	Universal unmethylated DNA
WHO	World Health Organization

Genes investigated in this thesis

ANKRD13B	Ankyrin repeat domain 13B
BARX1	BARX homeobox 1
CDKN2A	Cyclin-dependent kinase inhibitor 2A (p16)
DMRTA2	DMRT-like family A2
DOT1L	DOT1-like, histone H3 methyltransferase
EN1	Engrailed homeobox 1
FAM78B	Family with sequence similarity 78, member B
FGF12	Fibroblast growth factor 12
GAS1	Growth arrest-specific 1
GATA3	GATA binding protein 3
GRIK2	Glutamate receptor, ionotropic, kainate 2
HIST1H3C	Histone cluster 1, H3c
HOXA1	Homeobox A1
LIMK1	LIM domain kinase 1
MSC	Musculin
ONECUT2	One cut homeobox 2
PARP15	Poly (ADP-ribose) polymerase family, member 15
PRDM14	PR domain containing 14
RASSF1	Ras association domain family member 1
RAX	Retina and anterior neural fold homeobox
SEPT9	Septin 9
TBX15	T-box 15
ZIC4	Zic family member 4

Genes and proteins mentioned in this thesis

ALK	Anaplastic lymphoma receptor tyrosine kinase
AMPK	Protein kinase, AMP-activated, alpha 2 catalytic subunit
APC	Adenomatous polyposis coli
BRAF	v-raf murine sarcoma viral oncogene homolog B1
CDH1	Cadherin 1, type 1, E-cadherin
CDH13	Cadherin 13, H-cadherin
CK5/9	Cytokeratin 5/6
CTCF	CCCTC-binding factor
DAPK	Death-associated protein kinase 1
DNMT	DNA methyltransferase
EGFR	Epidermal growth factor receptor
EML4	Echinoderm microtubule associated protein like 4
FAS	TNF receptor superfamily, member 6
HAT	Histones acetyltransferases
HDAC	Histone deacetylases
HDM	Histone demethylases
HER2	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2
HMT	Histone methyltransferases
IGF2	insulin-like growth factor 2
KRAS	v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog
LKB1	Serine/threonine kinase 11
MET	Met proto-oncogene (hepatocyte growth factor receptor)
MGMT	O-6-methylguanine-DNA methyltransferase
MYC	v-myc myelocytomatosis viral oncogene homolog
NKX2-1	NK2 homeobox 1 = Thyroid transcription factor 1
OTX1	Orthodenticle homeobox 1
p53	Tumor protein p53
p63	Tumor protein p63
PI3K,	Phosphoinositide-3-kinase
PIK3CA	Phosphoinositide-3-kinase, catalytic alpha polypeptide
PKR	Double-stranded RNA-dependent protein kinase
RB	Retinoblastoma 1
RUNX2	Runt-related transcription factor 2
TNF	Tumor necrosis factor
TTF-1	Thyroid transcription factor 1 = NK2 homeobox 1

Preface

In this booklet I will take you on the journey of my doctoral degree. The main subjects are both intriguing and controvert: epigenetic regulation by DNA methylation, focused on a common type of lung cancer called non-small cell lung carcinoma. For those who are already in the field of oncology or epigenetics, I will try to make the reading of this thesis as fascinating as the subjects have become to me in the past four years. For those who have no interest in or understanding of those particular subjects, but for whatever reason decide to read it, I hope the reading will be light enough to get you through, to leave you with an impression of how much is happening in our cells and why understanding these processes is important, not only for basic science, but also for the clinical applications that can arise from it.

Scope of this thesis

DNA methylation is a signaling marker used by the cell to control gene expression, to keep genes silenced or active. It is an important part of what is called epigenetic controlling mechanisms (epi- Greek: επί- over, above, outer). We are just beginning to understand the intricate processes involving this type of epigenetic regulation and its role in normal development and carcinogenesis. In most types of cancers, a combination of environmental, genetic and epigenetic factors is implicated in initiating and maintaining the disease. Non-small cell lung carcinoma is a complex malignancy, posing challenges to accurate diagnosis and prognosis, which are essential for the correct choice of treatment. Despite recent advances in molecular biology and oncology, diagnosis is still largely based on classical histological classification.

Profiling DNA methylation is crucial to find novel regions which not only may be involved in the biology of the tumor, but could also provide sensitive and specific markers for tumor phenotyping. Such markers may be a useful tool in cancer diagnosis and prognosis due to the stability, inheritability and tissue–specificity of DNA methylation. Profiling DNA methylation in cancer in a genome-wide manner has become an area of intensive research efforts, but as yet there is no standardized technology to generate genome-wide profiles, nor is there standardized statistical methodology to analyze the data. Because of this, genome-wide profiles are often biased, reflected by the small number of regions that overlap between different studies.

At present, there is not a single truly genome-wide profile of DNA methylation in NSCLC. In addition, there is no methylation marker or a panel of markers with sufficient sensitivity and specificity for clinical use. Therefore the main purpose of the research described here was to provide genome-wide DNA methylation maps of

NSCLC and normal lung tissues, for identification of differentially methylated regions (DMRs) that can be used to develop sensitive biological markers for NSCLC. This may enable development of a panel of markers that can be used for non-invasive diagnosis and histological classification of NSCLC.

In **Chapter 1** I introduce you to general concepts in non-small cell lung cancer (NSCLC), how epigenetic regulation by DNA methylation might be involved, and why DNA methylation can be useful in NSCLC studies. In **Chapter 2** we devised a combination of assays and methodologies which were used to profile NSCLCs and lung samples. Firstly, we isolated methylated DNA from lung and tumor tissues from patients and fully sequenced them with high-throughput technology, providing sequence reads representing the entire genome. We then made use of a novel normalization method, using unique controls: totally unmethylated genomic DNA to set up the lowest (background) levels and totally methylated genomic DNA for the highest (reads saturation) methylation levels. The controls were submitted the same process as tumor and lung samples. Next, we compared the normalized reads between normal and tumor samples using a very stringent statistical analysis. The combination of these methods led to the generation of very trustful profiles. These profiles are a valuable resource for the search of tumor and histological markers. Next, we used these profiles to select DMRs which are poorly and never investigated. In **Chapter 3** we examined the methylation status of 25 DMRs from our profiles in 61 NSCLC samples, searching for sensitive and specific markers. Most of the DMRs we investigated showed higher specificity and sensibility than most of methylated markers reported in the literature to date. Subsequently we tested different combinations of these DMRs to find the best panel to predict NSCLC tumors. We propose a panel of 5 markers that is 100% specific and has 94% sensitive to distinguish between tumor and normal tissue. Finally, in **Chapter 4**, I present a general discussion of the results described in Chapters 2 and 3, and discuss how our findings can be used to further advance NSCLC research.



1

Introduction

Non-small cell lung carcinomas
and epigenetic regulation
by DNA methylation

Since my research involved two subjects that are not necessarily linked to each other, in the following pages I will give you background knowledge in both. In the first part, I will start by telling you the reasons why we chose to investigate Non-small cell lung carcinomas (NSCLC), and by doing so you will understand what type of cancer NSCLC is and get to know the nomenclature of lung tumors and their classification, which is relevant because it reflects the complexity involved in studying NSCLC. I will lead you through the importance of correct diagnosis and, consequently, the necessity of good tumor markers. I will describe the molecular mechanisms that we know play a role in this disease, and how DNA methylation fits in this context. Here starts the second subject of my research, where I will introduce you to epigenetics and epigenetic regulation by DNA-methylation; what we know about it, but also how much we do not yet know about this regulatory mechanism. The reasons why DNA methylation makes good markers and which ones are already identified in NSCLC are discussed. Finally, I will briefly describe the drawbacks and advantages of the different methods currently used to profile DNA methylation.

1 - Non-small cell lung carcinomas

Why studying NSCLC? I could just say that any research in cancer is self-explanatory. You do not need to be in the medical research field to know how deadly this disease is, and how important it is to develop the best treatments and better diagnostic methods. However, with regard to lung cancer in general, and NSCLC in particular, there is an even more pressing reason: NSCLC is not only the most frequent type of cancer world-wide, but also the deadliest one (Ferlay et al., 2010). Figure 1 depicts the incidence and mortality of the 10 most frequent types of cancer. Lung cancer stands out with the highest incidence and mortality rates world-wide, considering both sexes and all ages (Figure 1A).

Figure 1B represents the comparison of the age-standardized rates (ASR) of the highest incidence cancers in The Netherlands and in Brazil, showing that even in the countries where lung cancer is not the most frequent type of tumor, it is still the deadliest malignancy. Although the rates presented in figure 1 represents lung cancer in general, NSCLC accounts for about 80% of all types of lung cancer (Jemal et al., 2006).

The reason lung cancers are so deadly is mainly due to the lack of symptoms in early stages of the disease. Very often, when the first symptoms appear, cancer has already spread to other parts of the body, with complains originating from the secondary (metastasized) affected sites. However, even when an early diagnosis is made and the tumor surgically resected, relapse often occurs within 5 years after first diagnosis (Chhatwani et al., 2009). Only about 15% of NSCLC patients survive more than 5 years (Jemal et al., 2006), and for patients in advanced stages where no surgery is recommended, prognosis has a 5-year survival of less than 1% (Nykamp, 2010).

In general, survival prognosis is based in the combination of three main factors: size of the tumor (scale from T1 to T4), presence/absence and number of regional lymph node metastasis (N0-N3), presence/absence of distant metastasis (M0-M1). Survival rates for NSCLC patients are shown in Table 1 (Baba et al., 2011; Goldstraw et al., 2007; Pfannschmidt et al., 2007; Scagliotti et al., 2011; Shepherd et al., 2007; Wang et al., 2010). The rates have a large variation within each stage, due not only to differences between studies but also to the heterogeneity of tumors.

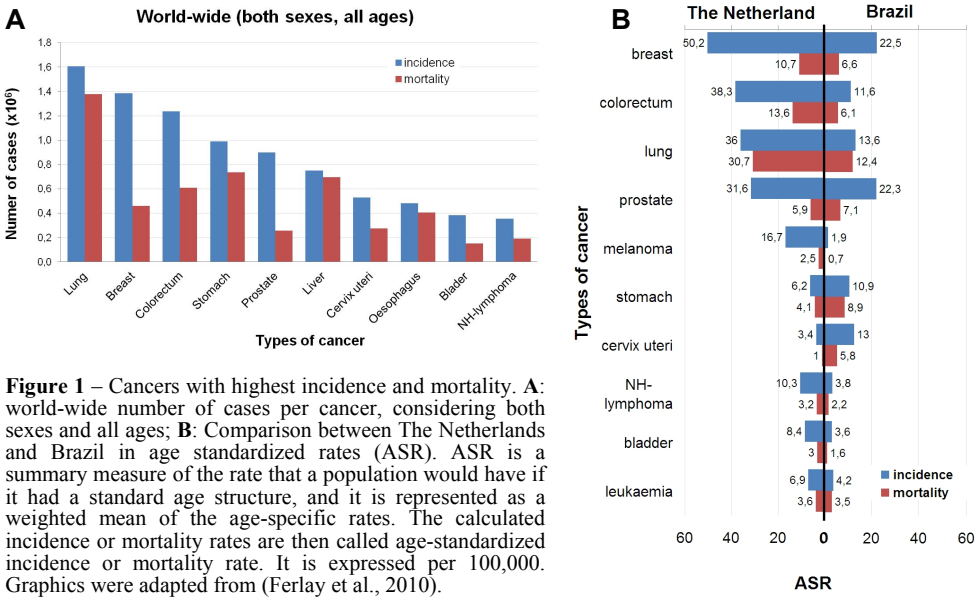


Figure 1 – Cancers with highest incidence and mortality. **A**: world-wide number of cases per cancer, considering both sexes and all ages; **B**: Comparison between The Netherlands and Brazil in age standardized rates (ASR). ASR is a summary measure of the rate that a population would have if it had a standard age structure, and it is represented as a weighted mean of the age-specific rates. The calculated incidence or mortality rates are then called age-standardized incidence or mortality rate. It is expressed per 100,000. Graphics were adapted from (Ferlay et al., 2010).

1.2 - NSCLC: A complicated name for a complex disease.

There are many different types of lung cancer. According to the World Health Organization (WHO) (2004), lung cancers are divided in two major groups: Small Cell Lung Carcinoma (SCLC) and Non-Small Cell Lung Carcinomas. SCLC is a very distinctive type of cancer, with well-characterized histological features (Stahel, 1992). NSCLC, on the other hand, receives this name because it involves several other subtypes and variants of lung cancer which are distinct from SCLC. The three main subtypes of NSCLC are Adenocarcinomas (ADC), Squamous cell carcinomas (SCC) and Large cell carcinomas (LCC).

1.2a - Adenocarcinomas

ADC is the most common type of lung cancer world-wide, accounting for nearly 40% of NSCLC cases (Travis, 2002). It is often associated with smoking habits, however it is the lung cancer with highest incidence among never-smokers (Subramanian and Govindan, 2007) and it is also the most frequent in women (Sagerup et al., 2011). ADC arises frequently in the periphery of the lung, and the main sites of metastasis are adrenal glands, liver, bone and brain. Its cells often show glandular formation, papillary structures and/or mucin production (Travis et al., 2011).

Table1: 5-years survival rates in NSCLC

Stages	Grades			5-years survival rate
	T	N	M	
IA	1	0	0	49-82%
IB	2a	0	0	45-68%
IIA	1	0	0	30-50%
	2a	1	0	
	2b	0	0	
IIB	2b	1	0	30-40%
	3	0	0	
IIIA	1 or 2	2	0	14-35%
	3	1 or 2	0	
	4	0 or 1	0	
IIIB	3	3	0	< 5% *
	4	2 or 3	0	
IV	1 to 4	0 to 3	1	~1% §

* Median survival of 13 months.
§ Median survival of 8 months.

ADC is the most heterogeneous subtype of NSCLC, with a large range of variation in histological, pathologic and molecular features. Thyroid transcription factor 1 (TTF-1/ NKX2-1) and napsin-A normally show as positive staining in ADC cells. (Roh et al., 2011; Yatabe et al., 2002). TTF-1 is particularly used for identification of lung as primary tumor site in metastasized areas, with immunological staining showing positive in 75 to 85% of metastasized ADC cases (Yatabe et al., 2002). Within ADCs, Bronchioloalveolar carcinoma (BAC) is the subtype that shows the widest differences in response to treatment, indicating a large heterogeneity within the tumors belonging to this sub-class (Travis et al., 2011).

In 2011, the International Association for the Study of Lung Cancer (IASLC), the American Thoracic Society (ATS), and the European Respiratory Society (ERS) reviewed the classification system and propose changes regarding the classification of adenocarcinomas (Travis et al., 2011). The proposed new classification is still based in histopathological criteria; however it integrates a multidisciplinary investigation, including histological, cytological, radiological and molecular approaches, mainly concerning BAC sub-classification, aimed at reducing misdiagnosis and improving patient care.

1.2b - Squamous cell carcinomas

SCC accounts for roughly 30% of all lung cancers. It is strongly related with smoking habits and most frequent diagnosed in men (Travis, 2002). The main initiating site for SCC is the central bronchus, more specifically the central portion of the tracheo-bronchial tree, although it can be also found in the periphery of this region (Holland, 2000). SCCs have some distinctive features, such as the presence of necrosis localized in the middle of the tumor which may result in central cavitation, and the production of keratin and desmosomes, also known as intercellular bridges, by the tumor cells. Tumors also frequently have an appearance of either spindle-like cell shapes or they present peripheral palisading patterns (Sakurai et al., 2004), where the cells in periphery of the tumor acquire a palisading distribution.

In immunohistochemistry staining, p63 and cytokeratin 5/6 (CK5/6) are the most reliable markers for SCC (Camilo et al., 2006). In cases where the particular features of ADC and/or SCC cannot be clearly observed under normal histological staining, a mixed panel of ADC and SCC markers (TTF-1/CK5/6 or p63/napsin-A) may be used, since those markers are often mutually exclusive (Rossi et al., 2009), except for tumors containing a mix of SCC and ADC or LCC.

1.2c - Large cell carcinomas

About 10% of all NSCLCs are diagnosed as LCC. In general, LCC is a poorly differentiated tumor type, presenting as large round cells with large nuclei and prominent nucleoli, which normally shows as a peripheral mass with pronounced necrosis without cavitations. They normally arise in the peripheral areas of the lung, show rapid growth and early metastasis (Travis, 2002).

This type of NSCLC is very often used in exclusion diagnosis, where the cytological features of ADC or SCC cannot be found. Under routine histological analysis, LCC shows no evidence of glandular or squamous differentiation, characteristics present in either one of the other two subtypes. However, studies using electron microscopy shows that those differentiating features are observed in most of the tumors classified as LCC (Said et al., 1983).

Not uncommonly, LCC diagnosis changes to ADC or SCC when a larger tumor sample becomes available and a more detailed analysis can be done. Reports have shown that immunohistochemistry analysis will be positive for intracellular mucin or cytokeratins in about 50% of the LCC cases (Delmonte et al., 1986), indicating poorly differentiated ADC or SCC in place of truly LCC.

Despite the division of NSCLC in these three major subgroups, you have by now an idea of how heterogeneous NSCLC and each of its subgroups are. This forms a major problem for diagnosis, which is currently based on histological classification and immunohistochemical staining. Due to the enormous variations in tumor proprieties

found in each subtype, not infrequently two pathologists give different diagnosis for the same tumor, where LCC or the terminology NSCLC-NOS (non-otherwise specified) is often used in diagnosis (Travis et al., 2011).

Studies of somatic mutations (Coate et al., 2009), copy number variation (Sy et al., 2004) and expression profiles (Hou et al., 2010) show that tumors belonging to the same traditional histological classification can display different molecular signatures. This indicates that, although those tumors are part of the same group, they may present different etiologies, behave differently in growth, predisposition to develop metastasis and response to treatment.

1.3 - NSCLC: molecular mechanisms

As for other types of cancers, there is not one cause for NSCLC. Initiation, development and maintenance of the disease are complex processes involving many factors, and interactions between them. Those factors can be of three different orders: environmental, genetic and epigenetic.

1.3.1 - Environmental factors

Not surprisingly smoking is the primary risk factor in lung cancer etiology. This association was already reported in 1950 (Doll and Hill, 1950). It is estimated that between 80% and 90% of lung cancers are resulting from smoking habits (Bryant and Cerfolio, 2007), and smokers have between 14 and 23 times higher risk of developing lung cancer than non-smokers . This observed risk variation is a reflection of the differences in level of tobacco consumption, duration of smoking habits (Peto, 1986), and starting age (Wiencke et al., 1999).

Some subtypes of NSCLC are more associated with smoking habits than others. While SCC is prevalently seen in smokers, ADC is more frequent in non-smokers (Figure 2). SCC used to be the most frequent type of NSCLC till middle of the 80's, but nowadays ADC has surpassed SCC incidence. This reverse in incidence rates is thought to be mainly due to change in smoking policies, addition of filters and changes in carcinogenic particles, such as introduction of low tar cigarettes (Harris et al., 2004).

Smoking does not only influence the probability of developing lung cancer, but it also affects the prognosis. Studies with lung cancer patients indicate that never-smokers have a better prognostic than smokers (Nordquist et al., 2004) and even former-smokers (Tammemagi et al., 2004). Smoking cessation decreases substantially the cumulative risk to develop lung cancer, but the risk for former smokers is still higher than that for never-smokers. Studies have also shown that not only active smoking, but also passive smoking, also called second hand smoke or environmental tobacco smoke (ETS), presents risk. ETS increases the risk of developing lung cancer by 25% (Taylor et al., 2007), and heavy daily exposure in childhood has a 2-fold increased risk to develop lung cancer in adulthood (Olivo-Marston et al., 2009).

Tobacco contains more than 40 carcinogen chemicals, from which at least 20 are strongly related to lung oncogenesis (Hecht, 2003). Constant exposure to those compounds, such as polycyclic aromatic hydrocarbons present in tobacco, frequently leads to DNA damage and gene mutation (Gyorffy et al., 2008).

Although **smoking** is the principal environmental risk factor, there are many others described for lung cancer, such as exposure to ionizing radiation, pollution, pesticide and particularly, **radon**, a natural occurring radioactive gas resulted from the decay of uranium and thorium (Samet, 2011; Samet et al., 2009). Chronic exposure to high levels of radon is considered the second predominant risk-factor in lung carcinogenesis, and exposure at home is responsible for 9% of lung cancer deaths in Europe (Darby et al., 2005). Other associations have also been found: **air pollution** such as from **nitrogen oxides** and **particles from traffic** (Lissowska et al., 2005; Vineis and Husgafvel-Pursiainen, 2005) have been associated to increased lung cancer incidence; exposure to **industrial carcinogens** (Taeger et al., 2006), **herbicides** and **pesticides** (Samanic et al., 2006) are also related.

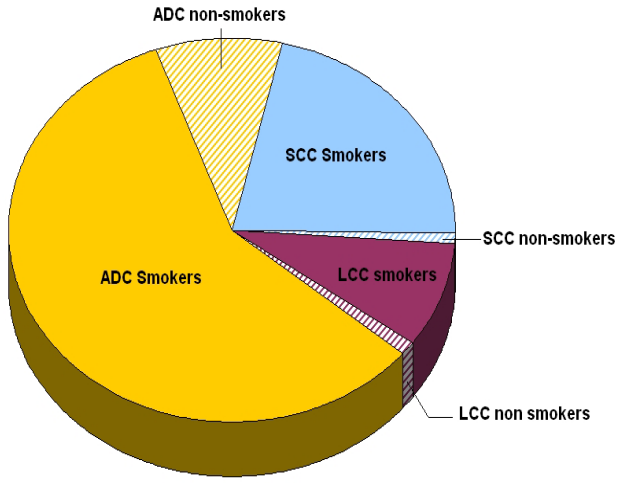


Figure 2: Incidence of NSCLC subtypes in smokers and non-smokes (Kenfield et al., 2008).

Risk-factors can also be disease related. Several **viral and bacterial infections** have been reported to be associated with increased risk of developing lung cancer (Liang et al., 2009; Littman et al., 2004); chronic obstructive pulmonary disease (**COPD**) is considered a major risk factor (Young et al., 2011), pneumonia and tuberculosis have also been positively associated (Brenner et al., 2011). These diseases

and the other factors mentioned above affect the lung tissue, exposing its cells to stress and cumulative DNA damage.

However, genetic predisposition may also play a role in NSCLC oncogenesis, since lung cancer is also observed in non-smokers and people with no history of exposure to other environmental risk-factors. Moreover, independently of what causes it, alterations in genes are the general underlying mechanism of oncogenesis.

1.3.2 - Genetic factors

Studies report a 2.5 fold increased risk to develop lung cancer in people with a family history of the disease (Nitadori et al., 2006). However, to date no particular genes have been defined as heritable cause. It is well known that several genetic aberrations are associated with NSCLC. Translocations, amplifications, loss of heterozygosity (LOH), somatic mutations, SNPs and a whole range of other genetic alterations have been reported. Expression profile studies demonstrate that hundreds of genes are differentially expressed in NSCLC when compared to their healthy counterparts. Some of those genes are associated with tumor signatures, others with survival, prognosis, and response to treatment (Hou et al., 2010). For most of those genes, there is no clear understanding of their role in the disease. They could be responsible for development and maintenance of the cancer cells, or their expression might have changed as a "bystander" effect caused by tumor-promoting aberrations in other genes.

I will discuss here genes with established roles in initiating, developing and/or maintaining the oncogenic state in NSCLC, focusing on the ones which are the most commonly known and strongly associated with NSCLC. Genes mutated or abnormally expressed in NSCLC are mostly concentrated in the EGFR and other kinase pathways [Figure 3], such as EGFR, KRAS, BRAF, PI3K, HER2 (ERBB2), MET and EML4/ALK. Other genes also found to be often mutated in NSCLC are P53, TTF-1 and LKB1.

1.3.2a - EGFR and KRAS

EGFR and KRAS harbor the most studied somatic mutations in NSCLC, and they are essentially mutually exclusive in NSCLC. The EGFR pathway regulates key processes in oncogenesis, such as proliferation, apoptosis, migration and angiogenesis. Mutations in the EGFR tyrosine kinase domain (exons 18-20) have an activating effect, leading to tumor proliferation and progression. Because genetic aberrations in EGFR are mostly found in its tyrosine kinase domain, tumors harboring them are target for therapies based in tyrosine kinase inhibitors (TKI), such as geftinib and erlotinib, that block activation of the pathway by preventing coupling of the epidermal growth factor. (Nguewa et al., 2011). Those mutations are associated with a more favorable prognosis (Linardou et al., 2009), and are mostly observed in ADC (about 40%), never-smokers, females and East Asians (Sequist et al., 2007). Amplifications and copy number

variations, however, are more often found in SCC, and associated with later stages, invasion and metastasis (Coate et al., 2009). Therefore, EGFR has become a predictive marker for prognosis, where some mutations (G719, L858R, L861Q and exon 19 deletion) are indicators of good response to TKI therapy (Ciardiello and Tortora, 2008; Kim et al., 2008), while others (insertions in exon 20 and T790M) are characterized as acquired resistance to treatment with those drugs (Pao et al., 2005). No response to TKIs is observed in patients with mutations in KRAS. This gene is a member of the RAS family of protooncogenes, and its pathway is downstream of EGFR (Herbst et al., 2008), consequently the signaling cascade is still activated independently of EGFR inhibitors. Mutations in KRAS are present in about 30% of ADCs and are rare in the other subtypes, are more often found in smokers and Caucasians (Finberg et al., 2007) and are associated with poor prognosis in NSCLC (Johnson et al., 2001).

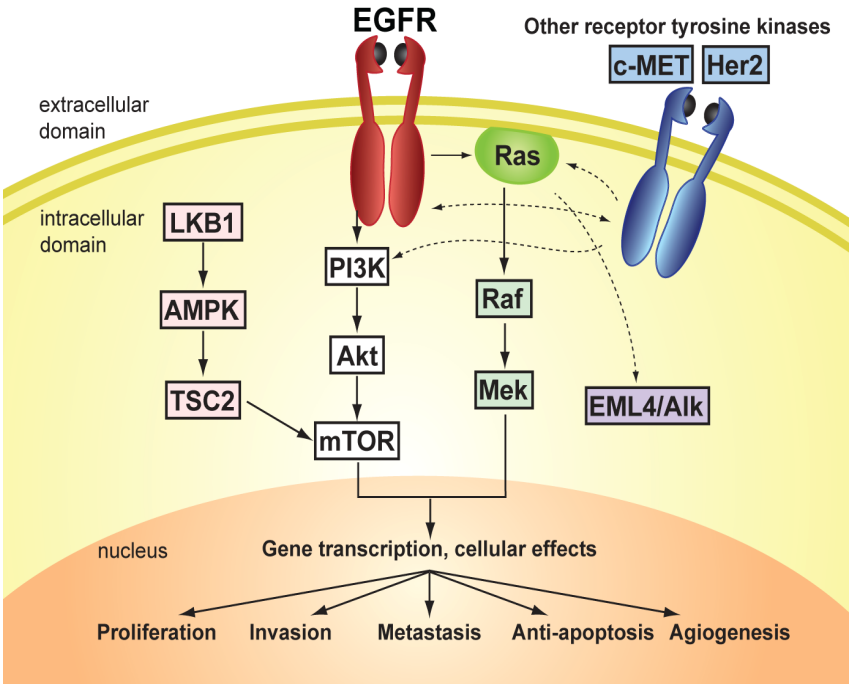


Figure 3: EGFR signaling pathway. Most common genetic aberrations in NSCLC are in the genes involved in EGFR pathway. Mutations that activate E(Allison, 2010; Hirsch et al., 2008)GFR or any of the other downstream pathways lead to proliferation, invasion, metastasis, resistance to apoptosis and angiogenesis. Adapted from (Allison, 2010; Hirsch et al., 2008)

1.3.2b - P53

Mutations in the gene that encodes the p53 protein (TP53) are the most frequent and reported mutation in cancer in general (Hainaut et al., 1997; Hainaut and Wiman, 2009; Hofseth et al., 2004; Olivier et al., 2010; Vogelstein et al., 2000), and NSCLC is not an exception with about 50% of the tumors showing some type of p53 abnormality (Olivier et al., 2010; Robles and Harris, 2010; Soussi et al., 2006). This tumor suppressor protein is a transcription factor which is activated in the presence of multiple cellular stresses, such as DNA damage (Mogi and Kuwano, 2011). Upon activation, p53 regulates downstream genes involved in cell cycle arrest, DNA repair, senescence and apoptosis, playing a major role in inhibiting tumor growth and development (Jin and Levine, 2001; Levine, 1997; Riley et al., 2008; Vogelstein et al., 2000). The most common mutations found in NSCLC are missense mutations affecting the DNA binding domain of the protein, and LOH (Baumann et al., 2009; Steels et al., 2001; Viktorsson et al., 2005). A particular mutation (GC to TA transversion) specifically correlates with tobacco smoking, since it is not frequently found in either other types of cancer or non-smokers (Toyooka et al., 2003b). Therefore TP53 mutations are slightly more frequently observed in SCC than in ADC (Lacroix et al., 2008), and they are also more often found in invasive than in noninvasive ADC (Huang et al., 1998; Terasaki et al., 2003; Yim et al., 2007; Yoshida et al., 2007), indicating a correlation between p53 mutation and poor prognosis (Baumann et al., 2009; Steels et al., 2001; Viktorsson et al., 2005). Despite the immense amount of data on p53 mutations in NSCLC, there is not enough overlap between them to suggest that p53 or any of its mutations can be used as a reliable predictive prognostic marker or as a therapeutic target (Travis et al., 2011).

1.3.2c - Other common and less common genetic aberrations

As mentioned in the beginning of this topic, genetic aberrations involving not only EGFR and KRAS are important players in NSCLC tumorgenesis. Alterations in their downstream effectors as well as in other kinases pathways are frequently found in NSCLC. **LKB1** (STK11) is found to be mutated in 34% of ADCs and 19% of SCCs (Ji et al., 2007). It is suggested that LKB1 acts as a tumor suppressor interacting with p53 and modulating AMPK activity (Ding et al., 2008), through a mechanism not yet understood. It is positively associated with KRAS and negatively with EGFR mutations (Finberg et al., 2007). Two other tyrosine kinase pathways altered in NSCLC are the hepatocyte growth factor receptor (encoded by the gene **MET**) and **HER2**. Amplification and point mutations in MET are found in about 20% of NSCLCs (Beau-Faller et al., 2008), while only 2% are detected for HER2, though HER2 protein over expression is also observed in 20% of NSCLCs (Stephens et al., 2004). Mutations in PI3K are not common; however amplification of the subunit that encodes the catalytic domain of PI3K (**PIK3CA**) are observed specially in SCC, and associated with increased expression of PI3K activity (Ji et al., 2011). **TTF1** is another gene often

found to be amplified in NSCLC (Berghmans et al., 2006), and used as an immunohistochemical marker for ADCs (Terry et al., 2010). In 2007, a chromosomal rearrangement was identified in a subset of NSCLC, which culminate in fusion of EML4 and ALK genes (Soda et al., 2007). **EML4/ALK fusion** results in gain of function, giving rise to oligomerization of the protein and activation of the kinase (Choi et al., 2008; Soda et al., 2007). This mutation is rare in SCC, and found in about 7% of ADCs (Soda et al., 2007). It is also more often associated with early onset of the disease (Shaw et al., 2009; Takahashi et al., 2010), and in never or light smokers (Horn and Pao, 2009; Koivunen et al., 2008). Tumors negative for EGFR or KRAS mutations may present EML4/ALK fusions (Soda et al., 2007).

Abnormal gene expression is also found in several other genes in NSCLC. However, the changes observed in those genes do not involve their coding sequence, such as mutations or amplifications. Rather, they influence gene expression through what is called epigenetic mechanisms.

1.3.3 - Epigenetic factors

In NSCLC, DNA methylation is a major epigenetic factor implicated in aberrant gene expression. Many genes, including tumor suppressor genes, are epigenetically inactivated by hypermethylation. Hypomethylation of genomic elements is reported to be associated with genomic instability in NSCLC. Since DNA methylation in NSCLC is the subject of this thesis, I will address this issue in more detail, where you will get to know about epigenetics, the most common epigenetic modifications and the role of DNA methylation in normal development, oncogenesis in general and NSCLC in particular.

2. DNA methylation as epigenetic regulatory mechanism

DNA methylation is a type of epigenetic modification. Therefore, to give you a better understanding of the how's and why's of DNA methylation, I will first briefly introduce you to epigenetics and its basic concepts, giving you in the way a quick overview of other epigenetic mechanisms that may cooperate with DNA methylation to coordinate chromatin architecture, remodelling and epigenetically regulate cellular processes.

2a - Epigenetics

The term epigenetics was conceived by the developmental biologist Conrad Waddington in 1942, who first defined it as "the branch of biology which studies the causal interactions between genes and their products, which bring the phenotype into being" (Waddington, 1942). Epigenetics is now commonly known as 'the study of heritable changes in genome function that occur without a change in DNA sequence'. In simpler words, for an element or event to be considered epigenetic, it has to meet

three basic criteria: 1) influence gene expression; 2) not affect the DNA sequence and 3) be inheritable. In recent years, our understanding of epigenetic biological processes has multiplied. The newly emerging knowledge needs to be fitted in the existing concepts or new concepts need to be created. Because of this, the strict definition of epigenetics which includes the heritability principle has come under attack. Many modifications considered as epigenetic are relatively unstable and short lived, such as some histone modifications; others are acquired in the course of development and can be changed, e.g. by environmental influences. All those variables has led to discussions about the definition of epigenetics, which mechanisms may or may not be classified as epigenetic, and how rigorously inheritable they must be. Despite the growing polemic, epigenetic mechanisms are currently divided into four main classes: DNA methylation; histone modifications (covalent and non-covalent) and microRNAs. Table 2 shows the main properties of each class (Sharma et al., 2010).

DNA methylation and histone modifications are the two most important epigenetic events, because they have a direct role in gene expression and chromatin remodeling (Figure 4A), while microRNAs play an indirect role by regulating proteins involved in DNA methylation and histone modifications (Friedman et al., 2009; Garzon et al., 2009). I highlight in the following two topics the roles of microRNAs and histone modifications, giving more emphasis to DNA methylation. All four mechanisms interact with each other, showing that the epigenetic machinery is an integrated regulatory mechanism.

2b - MicroRNAs

MicroRNAs (miRNAs) were first described in the beginning of 90's (Lee et al., 1993), but it needed another 10 years to unveil their role in regulating gene expression (Reinhart et al., 2000). MiRNAs are small non-coding RNAs with a sequence complementary to that of the target messenger RNA (mRNA). MiRNA forms complexes with other proteins and upon binding to the target mRNA, they inhibit protein translation by promoting mRNA degradation or by blocking progression of the protein synthesis. MiRNAs have been described to be involved in a wide variety of cellular processes, from differentiation to apoptosis (Feng et al., 2011; Li and Carthew, 2005; Schmittgen, 2008). As epigenetic regulators they play an indirect role by targeting proteins involved in DNA methylation and those responsible for histone modifications (Friedman et al., 2009; Garzon et al., 2009). They are also known to be regulated by DNA methylation. Hypermethylation has been observed in the DNA sequences that encode some microRNAs (Dickstein et al., 2010; Wong et al., 2011), resulting in their silencing and consequently disrupting the biological process regulated by them.

Table 2: Epigenetic mechanism in 3 types of normal mammalian cells

		All cell types	ES cells	Somatic cells
DNA methylation		Stable heritable modification; Gene silencing; Chromatin organization; Imprinting, X-chromosome inactivation, silencing of repetitive elements; Mediated by DNMTs	Bimodal distribution pattern: Global CpG methylation CpG islands unmethylated; Pluripotency gene promoters unmethylated	Tissue-specific methylation of some CpG islands and most non-CpG island promoters; Pluripotency gene promoters methylated
Histone modifications	Covalent	Labile heritable modification; Gene silencing and gene activation. Specific distribution patterns of histone marks contribute to chromatin organization. Mediated by HMTs, HDMs, HATs and HDACs etc.	Bivalent domains - active and repressive marks at promoters of developmentally important genes Plastic epigenome	Loss of bivalency and restricted epigenome Establishment of tissue-specific monovalent H3K27me and H3K4me domains Presence of large organized chromatin H3K9 modifications
	Non-covalent (positioning and variants)	Labile epigenetic regulatory mechanism; Gene silencing and gene activation; Both sliding of existing and incorporation of new nucleosomes Mediated by ATP-dependent chromatin-remodeling complexes	Epigenetic patterns and combination among them make up the epigenome of each cell type. Embryonic Stem cells (ES cells) shows a more plastic epigenome, necessary for keeping the pluripotency characteristic of those cells. Epigenome of somatic cells are more rigid and most of its structures are passed to the daughter cells. DNMTs: DNA methyltransferases; HMTs: histone methyltransferases; HDMs: histone demethylases; HATs: histones acetyltransferases; HDACs: histone deacetylases.	
miRNAs		Labile epigenetic regulatory mechanism; Gene silencing; Tissue-specific expression; Can be epigenetically regulated		

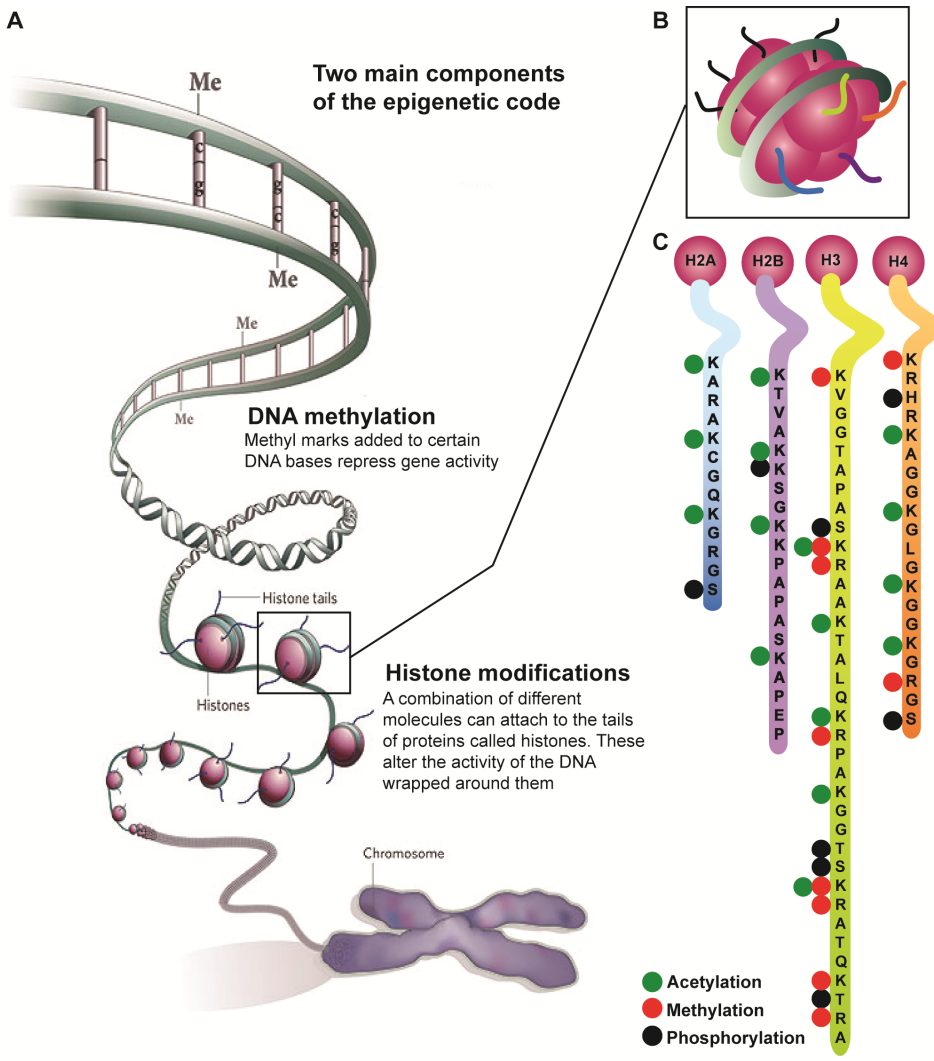


Figure 4: The epigenetic code. **A**: DNA methylation and histone modifications are the two main components of the epigenetic gene regulation machinery. **B**: Histone octamer showing the DNA wrapped around it (dark green) and the tails of each histone (blue, purple, light green and orange). **C**: The three most important histone tails modifications (acetylation, methylation and phosphorylation), showing their distribution, and which aminoacids residues they are commonly associated to. Adapted from (Qiu, 2006), with permission.

2c - Histones: covalent and non-covalent modifications

Histones are the core proteins that constitute nucleosomes, the basic structures where DNA is wrapped around to form chromatin. A nucleosome is composed of 4 different histones (H2A, H2B, H3 and H4) with two copies each, so a nucleosome is an octamer of histones (Figure 4B). When the histones are packed into nucleosomes, the N-termini become exposed forming tail-like structures aptly termed "histone tails". The tails of the histones can undergo different modifications, where chemical groups are attached covalently to the amino acids exposed in the tail. Many different modifications of the histone tails have been described (Kouzarides, 2007). Figure 4C depicts the three best studied modifications of the tail of each histone. Those **covalent modifications** play an important role in regulating cellular processes and chromatin rearrangement. They are associated with activation or repression of gene transcription, depending on which modification is present in which amino acid. Acetylation is often associated with activation of transcription (Schubeler et al., 2000; Sternglanz, 1996), while methylation is linked to repression or activation depending on the position and the level of the methyl group. Promoters of repressed genes are enriched for trimethylation (me3) of the lysine 9 (K9) in the tail of the histone 3 (H3) - H3K9me3 and also H3K27me3(Hublitz et al., 2009; Stewart et al., 2005); while active transcription is observed when H3K4me3 is present (Liang et al., 2004). These modifications work via signaling and recruitment of other proteins, allowing or blocking transcription. They may change and form different patterns, depending on their location in the genome. The patterns constituted by those modifications are known as the histone-code, and it has been suggested that they may determine cell identity (Shukla et al., 2008).

Considering that the DNA is wound around the nucleosomes, it becomes clear that their positioning affects accessibility of the DNA. Regulatory regions of genes are often nucleosome-free; hence those regions become exposed for binding of transcription factors (Jiang and Pugh, 2009). Shifting of nucleosomes along the DNA strand (nucleosome remodeling) is a physical process. It is executed by nucleosome remodeling enzymes, and called non-covalent for not involving a chemical bond of the modification group in the protein amino acid sequence. Another **non-covalent modification** is the substitution of the core histones for other variants such as H2A.Bbd, H2A.Z, H2A.X, macroH2A and H3.3. For example substitution of the H2A core histone by its variant H2A.Z is thought to facilitate nucleosome folding and to inhibit form ation of highly condensed chromatin and protect against DNA methylation (Zilberman et al., 2008).

While histone modifications (covalent and non-covalent) may participate as individual players in modulating chromatin structure, they also interact with DNA methylation as part of the mammalian epigenetic machinery.

2.1 - DNA methylation

In contrast to methylation in histone tails, which may enable or repress gene transcription, methylation of DNA is associated with gene silencing. In mammals, DNA methylation occurs by the addition of a methyl group to the position 5 of cytosine residues therefore named 5-methylcytosine (5-mC) (Figure 5A), and especially in those followed by a guanine: the so called CpG (Figure 5B).

These CpG dinucleotides are spread all over the genome in an unequal manner. Genomic regions presenting particularly high CpG concentrations are known as CpG islands. In the human genome about 60% of transcription starting sites (TSS) coincide with CpG islands (Bird, 1986), suggesting these might be important regions for epigenetic regulation by DNA methylation. However, many genomic regulatory regions are CpG island-free and methylation is not restricted to CpG-rich regions. Patterns of DNA methylation across the genome are established in early development and changes in those patterns may lead to diverse types of diseases.

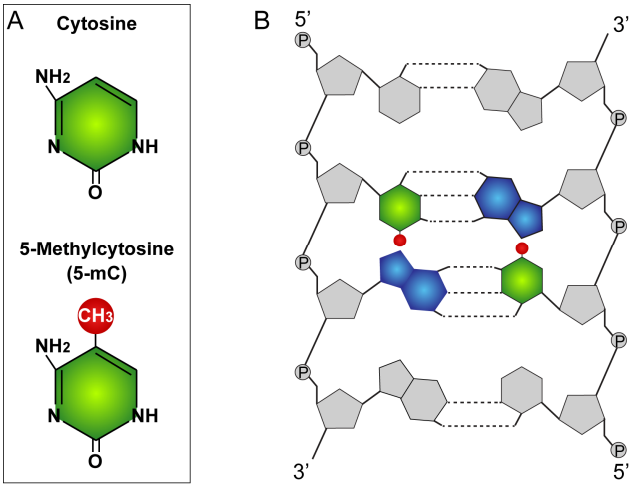


Figure 5: Cytosine methylation. A: Chemical structure of unmethylated cytosine (above panel), and 5-methylcytosine (bottom panel) depicting in red the methyl group. B: Schematic representation of small section of a double strand DNA, depicting a 5-methylcytosine nucleotide (green with a red circle representing the methyl group), followed by a guanine in the same strand. The "p" from CpG is to express the phosphor group which links the two nucleotides

2.1a - DNA methylation in normal development

What causes the different cells in our body to exert different functions, considering that their genetic make-up is equal? How can a single cell (the zygote) give rise to all different cells in the body? The fates of differentiated (somatic cells) and undifferentiated cells (e.g. embryonic stem cells – ES cells) are determined in the embryonic stages and maintained by intricate genetic and epigenetic mechanisms. DNA methylation is a key player in these processes (Morgan et al., 2005).

Genome imprinting, X chromosome inactivation and transposon silencing are well characterized mechanisms in which DNA methylation has a central role in normal development (Bird, 2002; Li, 2002). Another less well characterized role of DNA methylation is its involvement in the process of epigenetic reprogramming, where DNA methylation marks are lost and regained during embryogenesis. In early embryonic stage, DNA methylation marks are removed to ensure pluripotency of the embryonic stem cells (Gu et al., 2011). In the course of development, during differentiation stages, pluripotency genes are silenced by re-establishment of DNA methylation (in cooperation with histone modifications), and tissue-specific patterns are formed (Filion et al., 2006). The mechanism by which DNA methylation regulates developmental genes during embryogenesis and how genomic demethylation occurs is mostly unclear. It has been recently proposed that DNA is not demethylated during epigenetic reprogramming, but instead they receive a hydroxyl group, which would make them "invisible" (Branco et al., 2011) to regulatory proteins. Independent of the biochemical mechanism, there is no doubt that DNA methylation is essential for normal development. Studies have shown that mice depleted for the enzymes responsible for methylating DNA die before or right after birth (Li et al., 1992; Okano et al., 1999). Methylation is a process orchestrated by enzymes called DNA methyltransferases (DNMTs). It takes place after DNA synthesis and is regulated by three DNMTs: DNMT1, DNMT3A and DNMT3B. DNMT1 acts during DNA replication by recognizing the hemimethylated DNA strand and adding the methyl group on the corresponding CpG in the newly synthesized DNA strand (Damelin and Bestor, 2007; Takebayashi et al., 2007). DNMT3A and DNMT3B are *de novo* methylators and replication-independent; they operate either in presence or absence of hemimethylated DNA. These three enzymes act cooperatively to ensure maintenance and inheritance of the DNA methylation patterns that differentiated cells acquire during development (Okano et al., 1999). After those patterns are established, other proteins - the methyl binding domain proteins (MBDs) - recognize and bind methylated DNA, recruit other proteins to the site thereby forming repression complexes (Lopez-Serra and Esteller, 2008) in a finely coordinated operation to guarantee normal cell function.

2.1b - DNA methylation in cancer

There are two global processes by which aberrant DNA methylation contributes to tumorgenesis: genome-wide hypomethylation and site-specific hypermethylation

(Figure 6). In normal cells, proto-oncogenes, repetitive elements, transposons and imprinted genes are kept silenced by methylation. Abnormal hypomethylation of those regions may lead to their activation, contributing to cancer development through different processes. Activation of proto-oncogenes and imprinted genes by hypermethylation drives pathological cellular growth; as exemplified by hypomethylation of the genes involved in the Jak-STAT pathway in breast cancer putative breast cancer stem cells (Hernandez-Vargas et al., 2011), and loss of imprinting of the insulin growth factor 2 (IGF2) in colorectal cancer (Cui et al., 2003). Hypomethylation of repetitive elements result in genomic instability, facilitating chromosome breaks and rearrangements (Esteller, 2007). Likewise, activation of viral sequences and transposons (endoparasitic elements) by hypomethylation of their sequences also contributes to genomic instability due to their genomic translocation (Miura et al., 2001). While in cancer hypomethylation may be found in several regions across the genome; hypermethylation is usually a more site specific event. CpG islands in gene promoters are by far the best studied elements. Abnormal hypermethylation of promoters, mainly those located in CpG islands, is found in many tumor suppressor genes, and is thought to be an early event in tumorgenesis (Balaguer et al., 2010; Huang et al., 2011; Ignatov et al., 2010). Several genes have been reported to be aberrantly methylated in tumors, mainly those that regulate important pathways in tumorgenesis, such as cell cycle control, growth signaling, angiogenesis, cell adhesion, apoptosis and DNA repair. The basic mechanism behind gene silencing by hypermethylation is block of transcription starting sites (TSSs). Complexes of proteins are recruited to the site of DNA methylation and, integrated with histone modifications and chromatin remodeling, preventing activation of TSS (Chatagnon et al., 2011; Fuks et al., 2003).

About 60% of promoters in the human genome are embedded in CpG islands (Wang and Leung, 2004). It has been shown that hypermethylation also regulates expression of genes that do not have promoters associated with CpG Island (Han et al., 2011), though the importance of non-CpG island promoters methylation on regulating gene expression has not been fully established. Moreover, genomic regions distant from genes are also observed to be hypermethylated, indicating that in addition to promoter regions, abnormal methylation influences other regulatory binding sites such as enhancers and insulators, as it has been shown for CTCF and for long-range gene silencing (Coolen et al., 2010; Court et al., 2011).

Although hypermethylation is undoubtedly responsible for silencing of tumor suppressor genes, the process that leads to aberrant methylation or why specific sequences become targets for hypermethylation is still poorly understood. It has been reported that H3K27me3 is frequently observed in hypermethylated regions in cancer cells (Schlesinger et al., 2007). H3K27me3 is a histone mark in ES cells, suggesting an association between epigenetics mechanisms in cancer and pathways in stem cell maintenance (Ohm et al., 2007). Another hint comes from the CpG island methylator

phenotype (CIMP) hypothesis. It has been observed that subsets of CpG islands become methylated in a synchronized manner (Teodoridis et al., 2008), which suggests that hypermethylation in cancers may be a coordinated process targeting certain regions. Another target theory comes from the observations that some oncogenic proteins may direct the recruitment of DNMTs to specific sequences, targeting specific promoters (Di Croce et al., 2002).

The hypotheses discussed above provide directions to the way methylation is set up and maintained in cancer. Nevertheless we are far from fully understanding all the nuances of this specific tumorgenesis mechanism. In particular, we are only beginning to unravel the role of hypermethylation in gene-free regions in the genome (shores) and the biological processes that lead to and control abnormal methylation in tumor cells.

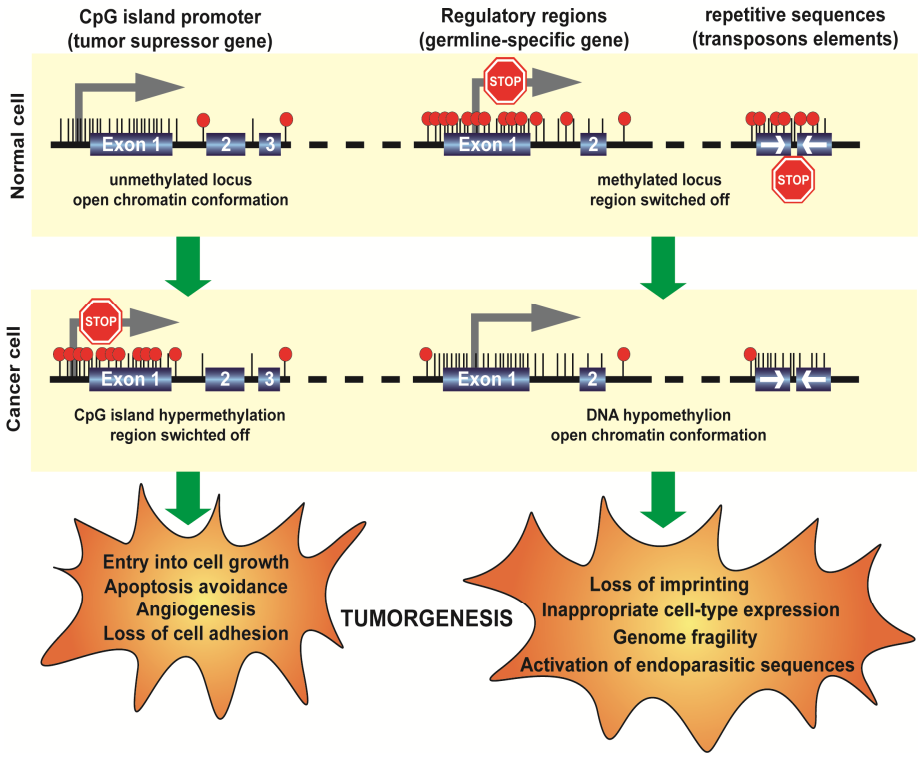


Figure 6: DNA methylation in normal cells and tumorgenesis. In normal cells (above panel) promoter of tumor suppressor genes and tissue specific genes are unmethylated and can be transcribed (grey arrow), while germ-line specific genes and transposon elements sequences are methylated and kept silenced. In tumors (bottom panel) tumor suppressor genes becomes hypermethylated and can therefore not be transcribed anymore, while germ-line specific genes and repetitive sequences lose methylation marks, becoming hypomethylated and therefore active. Both: aberrant hypermethylation and hypomethylation may lead to tumorgenesis . Adapted from (Esteller, 2007).

2.1c - DNA methylation in NSCLC

In NSCLC, as for other cancers, the most observed and documented epigenetic abnormality is hypermethylation in CpG islands of TSSs. Studies in NSCLC has shown that CpG island promoter hypermethylation occur preferentially in genes involved in later stages of somatic development, indicating that silencing of those genes may disrupt maintenance of a differentiated cellular state. Several studies comparing NSCLC to normal lung tissues or cell lines reported a large number of genes being differently methylated in the tumor; e.g. APC, CDKN2A (p16), FHIT, RARβ-2, RASSF1, CDH13, DAPK, MGMT, and others. The samples used for detecting the methylation status of those genes varied from primary tumor tissue to sputum, bronchial lavage, serum and even condensed exhaled breath. Although the rate of hypermethylation in the samples varies among the studies, because different sample types and different methodologies are used, some genes still recur as highly methylated independently of sample type or method. They are mainly genes involved in cell cycle arrest, DNA repair, apoptosis, invasion and proliferation (Belinsky et al., 2002; Honorio et al., 2003; Kersting et al., 2000; Lamy et al., 2002; Soria et al., 2002).

There are intrinsic and extrinsic factors that may lead to incorrect DNA synthesis or DNA damage. For ensuring a normal cell/tissue/organism function, each cell needs to be able to identify and correct/repair the mistakes or eliminate the malfunctioning cells. During the cell cycle, there are two main check points where DNA synthesis is controlled, mistakes corrected and damage repaired. When a problem is found, the cell cycle is stopped, a mechanism known as cell cycle arrest, where mistakes can be corrected and damage repaired. When cells cannot either undergo cell arrest and/or DNA cannot be repaired, another backup mechanism enters in action: apoptosis, which leads to cell death and consequently elimination of defective cells. In NSCLC, many genes involved in regulating one or more of the pathways involved in the above described mechanisms are aberrantly methylated, resulting in abnormal cell growth.

CDKN2A is a tumor suppressor gene inactive by hypermethylation in 60-70% of NSCLCs (Belinsky, 2004). This gene encodes for the INK4A protein, an inhibitor of cyclin dependent kinases CDK2, 4 and 6, via a pathway involving another tumor suppressor, the retinoblastoma protein (RB), which plays a role in cell cycle control (Lukas et al., 1995; Weinberg, 1995). Cyclin-dependent kinases phosphorylate RB, impeding binding of RB to its effector, resulting in disruption of cell cycle G1 phase check point (Lukas et al., 1995). Therefore, CDKN2A is an important player in allowing cells to undergo G1 checkpoint control. Alterations in this pathway via one or more of its players are observed in many types of cancer, including SCLC and NSCLC. Interestingly CDKN2A hypermethylation is almost never found in SCLC, where mutations in RB are predominantly observed (Harbour et al., 1988). Another tumor suppressor frequently methylated in NSCLC is APC (Toyooka et al., 2003b). APC is a negative regulator of the Wnt pathway, a pathway well-known to be involved in cancer

development and progression (Virmani et al., 2000), where β-catenin is the end effector, leading to expression of many genes involved in cell growth, such as MYC, a protooncogene. In normal cells, β-catenin levels are kept under control by a degradation complex in which APC is an essential player. Silencing of APC leads to increased levels of β-catenin and consequently uncontrolled cell growth (Usadel et al., 2002).

Genes involved in DNA repair (e.g. MGMT) and apoptosis (e.g. DAPK) are examples of genes recurrently found to be aberrantly methylated in NSCLC (Gu et al., 2006). Cells are protected from carcinogens as alkylating agents by enzymes such as MGMT, an O6 methylguanine-DNA-methyltransferase. These enzymes act by removing adducts from the O6 position in the guanine. Failure to remove these adducts is associated with mutations in genes (Belinsky, 2004). MGMT is methylated in about 24 to 48% of ADCs (Esteller et al., 2000; Pulling et al., 2003; Zochbauer-Muller et al., 2001), resulting in failure of this repair mechanism. Methylation of MGMT is associated with mutations in KRAS (Esteller et al., 2000; Park et al., 2001). Apoptosis is a complex cell death mechanism controlled by many signaling molecules, including members of the TNF (tumor necrosis factor) family. DAPK positively mediate apoptosis via induction of γ-interferon, TNF-α and FAS signaling (Cohen et al., 1997; Cohen et al., 1999). DAPK silencing is thought to disrupt the pathway, enabling cells to escape apoptotic signals. DAPK hypermethylation is found in about 30-50% of NSCLC (Kim et al., 2001; Tang et al., 2000; Toyooka et al., 2003a).

Under the topic “molecular mechanism” of this thesis, I showed that the RAS signaling pathway has a major role in controlling cell proliferation, invasion and apoptosis. Not surprisingly, abnormal expression of molecules involved in this pathway may lead to disruption of one or more of those processes, resulting in tumorgenesis. RAS association domain family protein 1 (RASSF1) is thought to be a RAS effector (Dammann et al., 2000). Not much is known about the exact role of RASSF1 in those cellular processes, but loss of or decrease in RASSF1 activity by inactivating mutations, LOH and hypermethylation is associated with different types of cancer (Agathanggelou et al., 2005; Bergqvist et al., 2010; Chen et al., 2005; Marsit et al., 2004). RASSF1 hypermethylation is observed in about 30% of NSCLCs (Wang et al., 2011).

Not only protein-coding genes are aberrantly methylated in NSCLC, microRNAs are also observed with different methylation levels in NSCLC tumors, such as miRNA124a and miRNAs-34b/c. In other types of cancer MiRNA-124 is thought to play a role in regulating CDK6 is the RB pathway (Agirre et al., 2009), while the miRNA-34 family plays a part in the p53 pathway (Bommer et al., 2007; Corney et al., 2007). Both pathways are directly related to mechanisms involved in tumorgeneis.

Many other genes/regions have been reported to be hypermethylated in NSCLC, and considering that those methylation changes are observed not only between tumor and normal tissues, but also between tumor subtypes and different cellular stages, hypermethylated genes have been proposed as indicators for diagnosis and prognosis.

2.3 - DNA methylation as biomarker

The search for DNA methylation as biomarker in NSCLC has lead to an explosion of studies screening primary tumors and distant sources for differentially methylated genes. Methylated markers for histological phenotype, early detection, prognosis and survival have been proposed, some with more success than others.

Many studies have reported genes with difference in the frequency of their methylation status between histological subtypes of NSCLC. Among other, genes such as APC and RUNX2 were found significantly more frequently hypermethylated in ADC, while DAPK1 and CDH1 presented higher methylation frequencies in SCC (Grote et al., 2005; Gu et al., 2006; Niklinska et al., 2009; Toyooka et al., 2001). Although the differences reported were statistically significant, there is discordance between these studies regarding not only the genes involved but also how significant the differences are. This discordance between genes and their significance indicates that not a single gene, but a panel of genes may be needed for discriminating different histological subtypes.

In contrast, methylation of CDKN2A is consistently reported as an early event in tumorgenesis. A study using four tissues at distinct stages of cellular differentiation (from basal cell hyperplasia to squamous cell carcinoma) observed positive association between methylation levels of CDKN2A and the differentiated cellular state, with 17% methylation frequency in basal-cell hyperplasia and 60% in SCC (Belinsky et al., 1998). In another study using cancer-free subjects, bronchial biopsies were performed in smokers and non-smokers, identifying CDKN2A methylation in 18% of the smoker and in none of the never-smokers (Belinsky et al., 2002). Detection of methylated genes in distant sources, such as bronchial lavages, sputum and serum is frequent. Data indicate that methylation of those genes is an early event, making them a good marker for early diagnosis. In one study, methylation of MGMT and CDKN2A was detected in the sputum of all SCC patients examined. Moreover, methylation of MGMT could be detected in the sputum from 5 months up to 3 years before tumors were clinically detected (Palmisano et al., 2000). Although the methylation frequency of marker genes is generally lower in distant sources than in primary tumors, it is almost never seen in healthy controls.

Aberrant hypermethylation can also be used as indicator of poor prognosis or survival. In 2000, a study showed association of DAPK1 methylation and survival in patients with stage I NSCLC, where hypermethylation of this gene was an indication of tumor aggressiveness (Tang et al., 2000). Since then, many studies have reported

correlation between methylation status of several genes (APC, CDH13, DAPK, DLEC1, FHIT, MGMT, RUNX3) and poor prognosis in NSCLC (Kim et al., 2003; Kim et al., 2005; Lu et al., 2004; Maruyama et al., 2004; Seng et al., 2008).

The run for finding methylation markers in NSCLC is not only for improving early diagnosis, for better detection of risk factors and to improve patient outcome. Methylation markers may also provide clues to key pathways that are defective in the tumors. The discovery of EGFR and KRAS mutations, for example, was a hallmark for providing a new line of therapy based on kinase inhibitors and predicting treatment response in patients. Regarding epigenetic markers in NSCLC, although specificity and sensitivity are still an issue, with the increasing data from methylation analysis of NSCLCs new opportunities for better diagnosis, prognosis and therapy are emerging.

2.4 - Epigenetic therapies and prospects for NSCLC

Methylation in the DNA and modifications of the histone tails are reversible modification due to their covalent chemical bonds, therefore a good direction for epigenetic based therapy. DNA demethylating agents, such as 5-azanucleosides and inhibition of histone deacetylases rely on the reversibility of those bonds. Drugs that inhibit DNA methyaltion by blocking DNMT activity such as 5-azacytidine (5-AzaC) and 5-aza-2'-deoxycytidine (5-Aza-dC) have long been know and used in assays investigating the role of DNA methylation (Cihak, 1974). In dividing cells these compounds are incorporated into the DNA (5-Aza and 5-AzadC) and RNA (5-Aza). During replication, DNMTs recognize the hemi-methylated strand and bind to the newly incorporated cytosine to add a methyl-group. When 5-AzaCs or 5-AzadCs are incorporated in place of cytosines, DNMTs become irreversible bound to those compounds. This depletes the level of DNMTs in the cells, inhibiting DNA methylation (Christman, 2002).

Their application as therapeutic agents was initially a challenge, mainly due to the high toxicity. Because only about 10% is incorporated in the DNA (Digel and Lubbert, 2005), low dosage was thought to be inefficient; while in large dosage the toxicity was too high for clinical use. The break-through came with low-dosage regimes, with positive responses in patients with myelodysplastic syndrome (MDS), a pre-state of leukemia (Fenaux et al., 2009). There are two DNA demethylating drugs currently approved and in use for MDS: 5-aza-cytidine (under the commercial name of Vidaza) and decitabine, a 5-aza-2'-deoxycytidine (Dacogen). Another success of epigenetic therapeutics is drugs targeting histone modifications, most notably histone deacetylases (HDAC) inhibitors. Vorinostat (Zolinza) and romidepsin (Istodax) are two FDA approved drugs for the treatment of cutaneous T cell lymphoma.

Multiple epigenetic drugs are currently undergoing clinical trials for different types of cancer (Esteller, 2011). For NSCLC, however the effects of epigenetic drugs in patients have yet to be proved. Despite the efficacy shown by these agents in

NSCLC cell lines (Zhu et al., 2001) and animal models (Belinsky et al., 2003), clinical trials have not reflected the same outcome (Momparler et al., 2000). Recently a combination of AzaC and HDAC inhibitors was tested in a phase II trial in advanced NSCLC patients that were previously treated with other therapies and relapsed. Good tolerance to the treatment were observed (Juergens et al., 2008). Currently, another phase II clinical trial is ongoing with, combining 5-azacitidine and Entinostat in resected Stage I Non-small Cell Lung Cancer (Juergens and Raines, 2010). For NSCLC, there is still a long way to go, demanding more clinical trials and strategies to improve use of epigenetic drugs, including determination of the optimal dosage and the most effective combination of different epigenetic agents, possibly in combination with standard therapies. Considering the speed at which new knowledge is emerging in the field of epigenetic alterations in NSCLC, the future of epigenetic therapy for NSCLC holds great promise.

2.5 - Profiling DNA methylation

There are a large number of methodologies to investigate DNA methylation, varying from diverse approaches of sample preparation to different choice for analysis of the methylated loci, and all combinations between them. Until a few years go, research in this field was directed to interrogating DNA methylation at specific sites or genes. Although those methods have provided useful information about the role of DNA methylation and methylation patterns in normal cells and in cancer, they do not give insight into the overall mechanisms involved in this type of epigenetic regulation. These technologies could not provide information on the differences between normal and cancer cells in a genome-wide manner. With the development of new sequencing technology, studies of epigenetic marks including DNA methylation have gone from single genes to cover the entire genome.

I will explain the basics of DNA methylation technologies, focus on the methods used in my work and the most commonly used genome-wide approaches, summarizing the most important aspects. This will give you enough background to understand the work described in this thesis. Extensive reviews on this subject can be found (Laird, 2010; Sharma et al., 2010; Suzuki and Bird, 2008).

All approaches used for profiling methylation, either single gene or genome-wide, make use of specific properties of the methylated CpG on the DNA. Some restriction enzymes, for example, cannot digest methylated CpG while their homologue can, as is the case for the HpaII and MspI **restriction endonucleases**. Both enzymes recognize the same sequence: CCGG. However, while MspI digest the DNA independently of CpG methylation; HpaII is blocked by the presence of a methyl group in the “CpG” dinucleotide. BstUI is another example; its activity is blocked by methylation of CGCG sequences. The fact that those enzymes produce different fragments when incubated with the same DNA sample has been used to investigate the

DNA methylation status of specific restriction sites (Kawai et al., 1994; Xiong and Laird, 1997).

Another basic technology was discovered more than 40 years ago, when in 1970 three groups observed changes in cytosine nucleotides when DNA was treated with sodium bisulfate (Hayatsu et al., 1970; Mukai et al., 1970) . They observed that treatment with sodium bisulfite would change cytosines to uracil residues. Further studies demonstrated that methylated cytosines were refractory to this reaction. These observations led to the use of **bisulfite conversion** to study DNA methylation, and have revolutionized epigenetic studies of DNA methylation. In contrast to restriction digestion, sodium bisulfite converts unmethylated cytosines to uracil, leaving methylated cytosines intact, providing information at single nucleotide resolution.

Alternative methods were established based on the ability of naturally occurring proteins, such as MBDs, to recognize methylated DNA, and antibodies against 5-mC. Both methods involve use of beads coated with the MBDs or 5-mC antibody for affinity selection of methylated DNA. The MBD-based method is called Methyl Capture (**MethylCap**), and the antibody-based method is known as **MeDIP**, for Methylated DNA immunoprecipitation. Each method has its particularities, though both provide an approach for isolating methylated DNA and therefore concentrating the investigation on regions where methylation occurs. Either one may be used in combination with bisulfite DNA conversion. The methods described below form the basis of most DNA methylation studies.

2.5a - Methylation-specific PCR (MSP)

MSP is a single-gene/fragment investigation, involving bisulfite conversion of DNA followed by PCR amplification (Herman et al., 1996). This method makes use of the differences between the converted and unconverted cytosines by bisulfite treatment. For investigating the methylation status of a certain gene fragment, two sets of primers are designed within the regions where methylation is expected to be found. One of the primers sets will have thymines in place of cytosines to amplify converted unmethylated DNA. The other set will match the original DNA sequence where cytosines that were methylated will be kept as cytosines. DNA is then amplified by PCR using both primer sets. Consequently, the gene fragment analyzed will be amplified depending on its methylation status (Figure 6). Results can be visualized by loading the PCR products on an agarose gel. MSP is a qualitative method, rather than quantitative. It is not possible to determine the rate of methylation accurately, or to determine the methylation status of individual CpGs. However, this technique has a number of advantages that make it widely used for single fragment investigation: 1) it is a low cost method 2) it is not labor-intensive and can be easily set up; 3) depending on primer design, MSP can be a highly sensitive and specific reaction and 4) the analysis of the results is simple and does not require specialized software or computer skills.

In order to transform MSP into a more quantitative approach, other methods have arisen that are based on real-time quantitative PCR. MethylLight (Eads et al., 2000) and HeavyMethyl (Cottrell et al., 2004) make use of fluorescent probes that are differentially labeled for detection of methylated or unmethylated site. Although measurements can be made when each fluorescent probes is added to the newly formed strand, makes use of probes which correspond to few sequences of methylated or unmethylated CpG, therefore they fail to detect when single or small numbers of CpG are modified. MethylQuant, also known as Methyl-SybrGreen, is another assay developed for quantification of MSP (Thomassin et al., 2004). It uses two sets of primers and fluorescent DNA dyes in place of probes. Differences in fragment amplification for each primer are measured following standard qPCR analysis. Although efficient to detect small changes in the status of methylated/unmethylated CpGs, this method is influenced by other issues, such as primer specificity, formation of primer dimers, and the interference of the fluorescent dye with the PCR reactions. All of those will result in undesired differences in qPCR readings. Sensitive Melting Analysis after Real Time (SMART)-MSP does not use fluorescent dyes, but relies on difference in melting curve reads for methylated and unmethylated fragments (Kristensen et al., 2008). This method is also influenced by primer specificity and formation of primer dimers. Moreover, although cost are still relatively low for these methods, comparing to standard MSP there is an increased requirement for specific reagents, PCR machines and analysis software. Therefore standard MSP is still the most feasible for diagnostic applications, in particular when considering low cost settings.

2.5b - Bisulfite sequencing

Bisulfite sequencing can be used for a single gene fragment and for genome-wide analysis. It involves bisulfite treatment of DNA samples followed by sequencing of the converted samples. For single gene analysis, after the samples are bisulfite treated, the fragment of interest is amplified using a single primer set devoid of CpGs, to ensure amplification regardless of methylation status. The amplified product is then ligated into an appropriate vector and cloned in *E. coli*. Two approaches can be used at this stage: 1) culture of single bacterial colonies, followed by extraction of the vectors and sequencing; or 2) colony PCR, followed by sequencing of the amplified product. The major advantage of bisulfite sequencing is that it enables quantitative assessment of the methylation status of individual CpGs in populations of cells. Furthermore, if single nucleotide polymorhpisms are present, allele-specific determination of the methylation status is possible. This is widely employed for investigation of imprinted loci. There are also drawbacks: bisulfite sequencing is a labor intensive technique with moderate costs, demanding specific software for analysis of the sequenced fragments.

2.5c - Unveiling the epigenome: genome-wide methylation profiles

For generating DNA methylation profiles the most broadly used approach to date is investigation of numerous loci spread across the genome. Although several groups have claimed this to be genome-wide (Kim et al., 2010; Rauch et al., 2007; Watanabe et al., 2011), it does not genuinely cover the whole genome, but it is restricted to a wide-spread selection of specific genes or genomic regions. The reason for its wide employment is the relative easiness of the method, together with a straight forward analysis and relatively moderate cost. Real genome-wide screenings are still highly expensive and standardized normalization inexistent. However with the increasing use of this methodology, costs are decreasing and normalization methods are being developed that allow correct interpretation of the reads and production of reliable data. An increasing number of groups have attempted and succeeded in providing genome-wide methylation profiles of different normal tissues and tumors, using diverse methodologies (Buckley et al., 2011; Fackler et al., 2011; Hill et al., 2011; Xu et al., 2011). However, because real genome-wide methylation profiles that are performed in different studies use different methods and technologies, data comparison is a problem that remains to be solved.

Regarding truly genome-wide methylation profiles, there are two methods widely used: array based and next-generation sequencing based. Array based focus in the hybridization of the DNA to pre-designed probe sequences. Those pre-designed sequences can cover the entire genome but are usually limited to well-known genomic regions, such as gene promoters and CpG islands. The next-generation sequencing technology does not discriminate which genomic regions are being investigated, providing a truly genome-wide screening of methylation patterns. Those methods can be used with either bisulfite converted DNA or methylated DNA enrichment by MeDIP or MethylCap assays. When bisulfite conversion is used for genome-wide analysis, instead of amplifying a gene or fragment of interest, the converted DNA is amplified using unspecific primers and sequenced using next generation sequencing or hybridized to an array. Arrays can be also used in combination with MeDIP and MethyCap, but the predominant approach chosen following those two assays is next-generation sequencing, in which case they are known as MeDIP-seq and MethylCap-seq. In both methods, genomic DNA from the tissues/samples of interest is enriched for the methylated fragments and the methyl-enriched fraction is ultra-deep sequenced, resulting in a genomic map specific for the methylated regions. The difference between the two methods lays on the ability of MBD proteins to bind double strand DNA, while anti 5-mC have affinity for denatured, single-stranded DNA. For most next generation sequencing technologies DNA, fragments needs to undergo a library preparation step, which can only be performed in double strand DNA. Therefore, for MeDIP-seq (antibody), library preparation needs to be executed before the sample undergoes methyl-enrichment. Because MethylCap works on double strand DNA, no previous

library preparation is needed. Figure 7 depicts a flow chart of the assays, methods and systems discussed here.

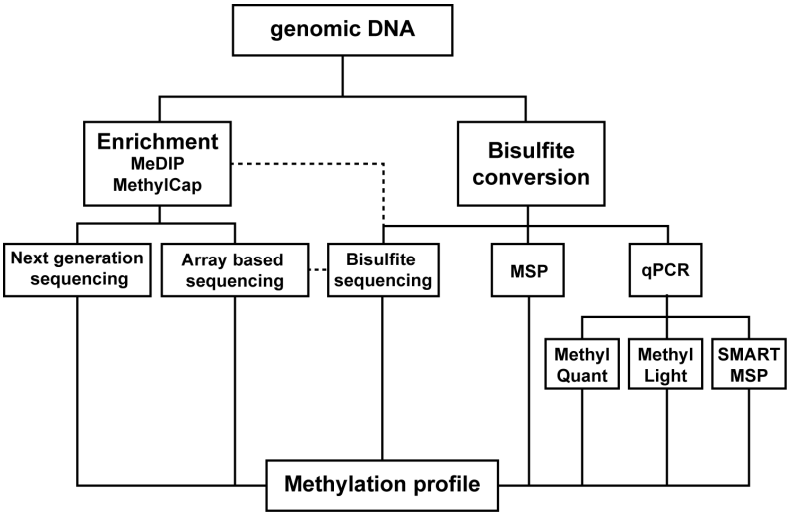


Figure 7: Methylation profile flow chart: schematic representation of the assays, methods and combinations among them to generate methylation profiles.

There are some drawbacks on using NGS. First, it is still a high cost method: although costs are decreasing, it is still the most expensive among the genome-wide systems. Second, the high demand on computer power: the number of sequence reads produced is in the order of billions and with improved technologies this is still increasing, resulting in terabytes of data that need to be stored and analyzed. Third, specific software, programs and bioinformatics specialized staff and tools are needed for analysis of the large amount of data in order to generate the genome-wide methylation profiles.

NGS technologies have been developed by different companies using different platforms to provide ultra deep DNA sequencing. I will briefly describe the technology behind each of the four most used systems: Illumina with GAIIx (HiSeq2000 is the newest version): controls the ultra-deep sequencing market nowadays; Roche with 454: the first system to become available on the market; Life technologies/Applied Biosystems with SOLiD and Bioscience with Helicos.

The Illumina platform is based on sequencing by synthesis (SBS) technology in solid phase (Figure 8). Prior to sequencing, fragmented DNA strands are ligated to

adaptors flanking each end, which will enable pairing to a complementary oligonucleotide in a flow cell. This technology supports parallel sequencing using reversible terminators, which are fluorescently-labeled terminators molecules bound to each dNTP. All four fluorescent dNTPs, together with DNA polymerase and the adaptors-ligated fragments are added to flow-cells channels, where bridged amplifications happen in clusters (solid-phase). Detection of single bases results as images are captured when each dNTP is incorporated, followed by terminator cleavage to allow incorporation of the next fluorescently-labeld dNTP (SBS). Most publications in the past years using Illumina platform were performed with the Genome Analyzer IIx (GAIIx). Currently the HiSeq2000 offers improved performance (Illumina, 2011).

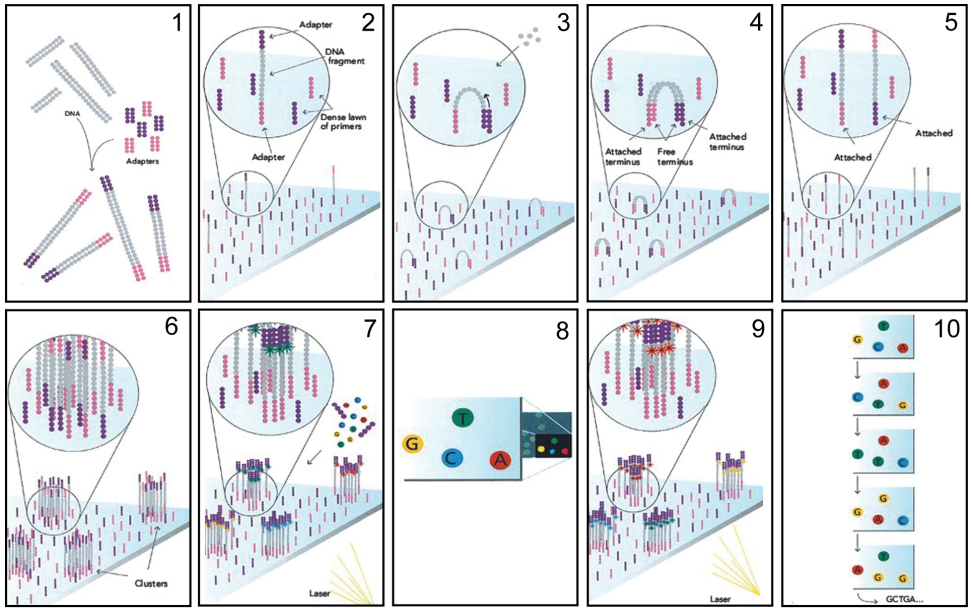


Figure 8. Illumina sequencing technology. The HiSeq2000 and GAIIx make use of the SBS in solid-phase technology, which is based on formation of fragment clusters in a flow cell (solid-phase) and addition of fluorescently labeled dNTPs bound to terminators, which allows capture of the signal. The terminators are cleaved off and the next fluorescently labeled dNTPs bound to terminators are added. 1) igate adaptor to both end of fragmented DNA samples; 2) Denature and bind single stranded template to the surface of the flow cell; 3) Add unlabeled dinucleotides and enzymes to initiate bridge amplification; 4) incorporate dinucleotides to make double stand; 5) denaturation leaves single stranded templates attached to the flow cell; 6) several million clusters of the fragmented DNA template is formed in each channel of the flow cell; 7) addition of the four labeled reversible terminators, primers and DNA polymerase. Initiation of the first chemical cycle and first sequencing cycle with laser excitation; 8) after laser excitation, image from each cluster is captured and based identified; 9) second sequencing cycle, repetition of step 7; 10) After all bases are added and image captured from each sequencing cycle, images are processed and base sequences from each cluster determined.

Roche 454 sequencing is also a SBS technology, however with emulsion-based clonal amplification, where PCR library is made in beads emulsified with amplification reagents in water-in-oil mixtures. This creates millions of clonally copied fragments on each bead. Beads are then loaded individually into holes in a PicoTiterPlate where pyrosequencing allows incorporation of chemiluminescent nucleotides (Figure 9), emitting signals which are registered by a camera (Roche, 2011).

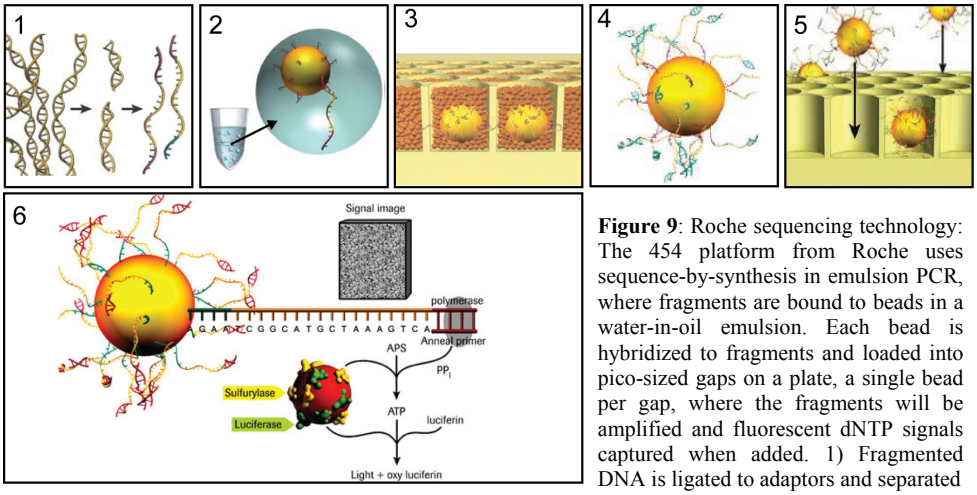


Figure 9: Roche sequencing technology: The 454 platform from Roche uses sequence-by-synthesis in emulsion PCR, where fragments are bound to beads in a water-in-oil emulsion. Each bead is hybridized to fragments and loaded into pico-sized gaps on a plate, a single bead per gap, where the fragments will be amplified and fluorescent dNTP signals captured when added. 1) Fragmented DNA is ligated to adaptors and separated in single strands; 2) Fragments are bound to beads under emulsion. Beads are captured in droplets of PCR-reacion mix in oil emulsion; 3) Beads are loaded into the slides, one bead per well. Amplification occurs resulting in beads carrying ten million copies of a unique DNA template. Emulsion is broken, DNA strands denatured and each bead placed in single wells in the picotiter plates with enzymes and additional beads, coupled with sulphurylase and luciferase, image is captured by a camera detecting the light generated by the addition of each nucleotide in the pyrosequencing reaction.

The SOLiD platform by Life Technologies uses also beads-in-emulsion, but with a sequencing by ligation approach. The clonally amplified fragments on the beads undergo a chemical modification in their 3' end to allow attachment of the beads to a glass slide, where primers on the plates are hybridized to adaptors from the fragments on the beads. It is called sequence by ligation because it involves fluorescent di-bases probes, which ligate to the growing sequencing primer. A single probe gives information on every first and second base they ligate to. After the fluorescent di-base is ligated, its signal emission is detected, and then the di-base cleaved. This cycle is repeated for the length of the fragment (Figure 10) (AppliedBiosystems, 2011).

Helicos is the only platform that uses a single strand DNA molecule approach. As for Illumina HiSeq2000 and Roche 454, it is also based on sequence-by-synthesis reactions, but in Helicos, single strand DNA harboring a poly A adaptor is loaded into

to a flow cell with the complementary poly T sequence. During the sequencing reaction, fluorescently-labeled nucleotides are incorporated one at a time forming a complementary strand using the original single stranded sequence as template (Figure 11). This system is also the only one which does not require amplification of the DNA fragments (Helicos, 2011).

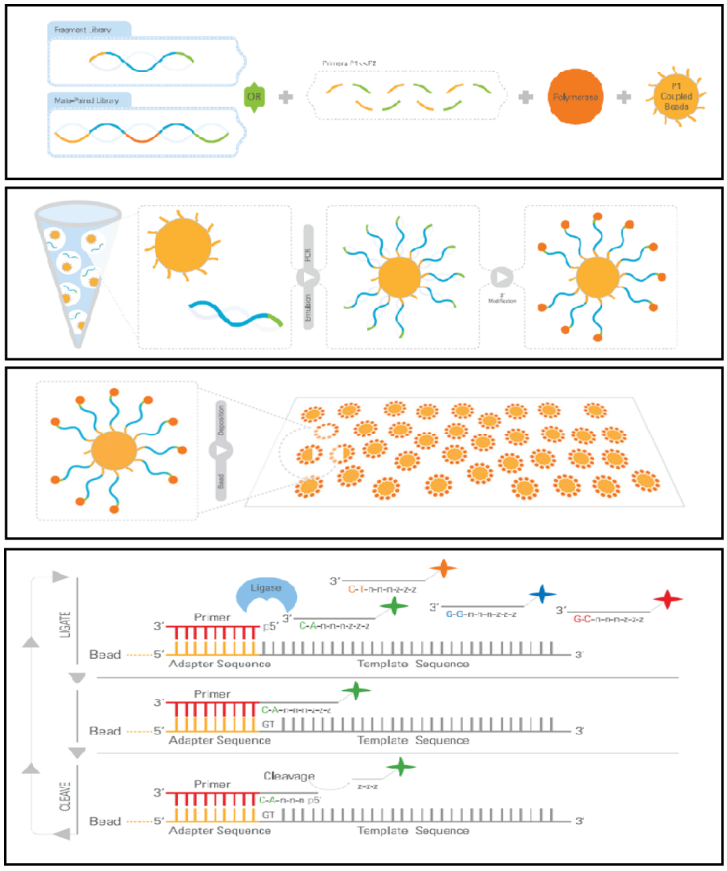


Figure 10: ABI sequencing technology. The SOLiD platform uses SBL in emulsion PCR, where fragments are bound to beads in a water-in-oil emulsion. 3' ends of fragments are modified to allow hybridization to a glass plate. Fluorescently labeled di-bases probes are ligated to the complementary di-base in the target strand, when signal emitted is captured. Each probe gives information about two bases and each base in a fragment is read twice. The cycle is repeated for the length of the fragment.

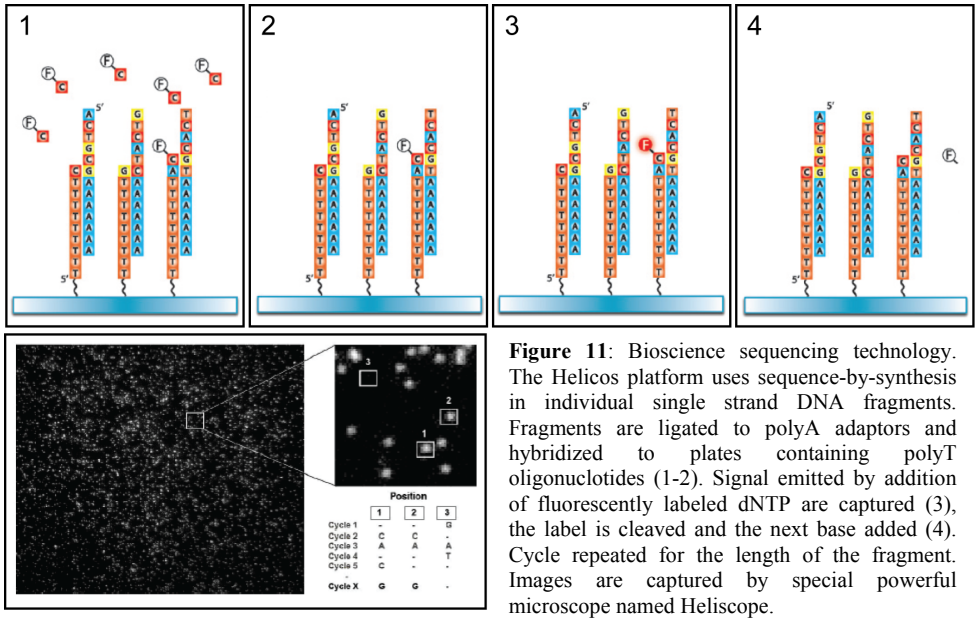


Figure 11: Bioscience sequencing technology. The Helicos platform uses sequence-by-synthesis in individual single strand DNA fragments. Fragments are ligated to polyA adaptors and hybridized to plates containing polyT oligonuclotides (1-2). Signal emitted by addition of fluorescently labeled dNTP are captured (3), the label is cleaved and the next base added (4). Cycle repeated for the length of the fragment. Images are captured by special powerful microscope named Heliscope.

Table 3: Comparison of sequencing platforms

Platform	Techno-logy	Read lenght (bp)	(Gb/ run) #	Run time (days)#	Machine cost (U$ x10^3) ¥	Advantages and disadvantages		Biological applications§
						+	-	
GAIIx	SBS solid-phase	2x 150	80	2 - 9	545	Most used system (for GAIIx, previous platform)	Low capacity of Multi-plexing	Variant discovery y whole-genome resequencing; whole-exome capture;
454 GS FLX	SBS emulsion	700*	1	0.9	500	Long reads; fast run times	High reagent costs; high error in homopoly-mer repeats	Indirect *de novo* assemblies; medium scale exome capture
SOLiD	SBL emulsion	60	30	6	595	Di-base probes improve error correction	Long run times	Variant discovery y whole-genome resequencing; whole-exome capture
Helicos	SBS single molecule	32*	37	2.5 - 8.0	999	No amplifica-tion bias	Higher error rates	Sequence-based methods.

* Average read length; # throughput and time can vary depending on the assay used, e.g. single run, matched pair; ¥prices in January 2010; § Most common applications. SBS:sequence-by-synthesis; SBL: sequence-by-ligation [adapted from (Metzker, 2010).

Each of the above described platforms has its advantages and disadvantages, and may be more suitable depending on the biological application (Table 3). However, no matter which method, approach or platform is chosen, the quick advances in genome-wide methylation profiles are opening new possibilities for DNA methylation analysis and giving better insight on how methylation influences normal development and its involvement in tumorgenesis.

In the next chapters I will describe how we used genome-wide technology in combination with the Methyl-Capture assay to accomplish the aims of this project. Our main goal was to profile genome-wide DNA methylation of NSCLC and normal lung tissues, and to identify differentially methylated regions that may be used to develop sensitive biological markers for NSCLC. For this we devised novel and reliable methods of normalization and analysis, which allowed us to generate extensive and comprehensive methylation maps of the tumors and lung tissues. We tested 25 novel markers and derived a panel of 5 markers capable of distinguishing tumors from normal tissues with high sensitivity and specificity.

References

Agathanggelou, A., Cooper, W.N., and Latif, F. (2005). Role of the Ras-association domain family 1 tumor suppressor gene in human cancers. Cancer Res *65*, 3497-3508.

Agirre, X., Vilas-Zornoza, A., Jimenez-Velasco, A., Martin-Subero, J.I., Cordeu, L., Garate, L., San Jose-Eneriz, E., Abizanda, G., Rodriguez-Otero, P., Fortes, P*., et al.* (2009). Epigenetic silencing of the tumor suppressor microRNA Hsa-miR-124a regulates CDK6 expression and confers a poor prognosis in acute lymphoblastic leukemia. Cancer Res *69*, 4443-4453.

Allison, M. (2010). Turning the tide in lung cancer. Nat Biotechnol *28*, 999-1002.

AppliedBiosystems, L.T.-. (2011). The 5500 Series Genetic Analysis Systems: Next-Generation Sequencing (Life Technologies - Applied Biosystems). http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/next-generation-systems.html. Accession date: 19/11/2011.

Baba, T., Uramoto, H., Kuwata, T., Oka, S., Shigematsu, Y., Nagata, Y., Shimokawa, H., Takenoyama, M., Hanagiri, T., and Tanaka, F. (2011). A study of surgically resected peripheral non-small cell lung cancer with a tumor diameter of 1.0 cm or less. Scand J Surg *100*, 153-158.

Balaguer, F., Link, A., Lozano, J.J., Cuatrecasas, M., Nagasaka, T., Boland, C.R., and Goel, A. (2010). Epigenetic silencing of miR-137 is an early event in colorectal carcinogenesis. Cancer Res *70*, 6609-6618.

Baumann, M., Zips, D., and Appold, S. (2009). Radiotherapy of lung cancer: technology meets biology meets multidisciplinarity. Radiother Oncol *91*, 279-281.

Beau-Faller, M., Ruppert, A.M., Voegeli, A.C., Neuville, A., Meyer, N., Guerin, E., Legrain, M., Mennecier, B., Wihlm, J.M., Massard, G*., et al.* (2008). MET gene copy number in non-small cell lung cancer: molecular analysis in a targeted tyrosine kinase inhibitor naive cohort. J Thorac Oncol *3*, 331-339.

Belinsky, S.A. (2004). Gene-promoter hypermethylation as a biomarker in lung cancer. Nat Rev Cancer *4*, 707-717.

Belinsky, S.A., Klinge, D.M., Stidley, C.A., Issa, J.P., Herman, J.G., March, T.H., and Baylin, S.B. (2003). Inhibition of DNA methylation and histone deacetylation prevents murine lung cancer. Cancer Res *63*, 7089-7093.

Belinsky, S.A., Nikula, K.J., Palmisano, W.A., Michels, R., Saccomanno, G., Gabrielson, E., Baylin, S.B., and Herman, J.G. (1998). Aberrant methylation of p16(INK4a) is an early event in lung cancer and a potential biomarker for early diagnosis. Proc Natl Acad Sci U S A *95*, 11891-11896.

Belinsky, S.A., Palmisano, W.A., Gilliland, F.D., Crooks, L.A., Divine, K.K., Winters, S.A., Grimes, M.J., Harms, H.J., Tellez, C.S., Smith, T.M.*, et al.* (2002). Aberrant promoter methylation in bronchial epithelium and sputum from current and former smokers. Cancer Res *62*, 2370-2377.

Berghmans, T., Paesmans, M., Mascaux, C., Martin, B., Meert, A.P., Haller, A., Lafitte, J.J., and Sculier, J.P. (2006). Thyroid transcription factor 1--a new prognostic factor in lung cancer: a meta-analysis. Ann Oncol *17*, 1673-1676.

Bergqvist, J., Latif, A., Roberts, S.A., Hadfield, K.D., Lalloo, F., Howell, A., Evans, D.G., and Newman, W.G. (2010). RASSF1A polymorphism in familial breast cancer. Fam Cancer *9*, 263-265.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. Genes Dev *16*, 6-21.

Bird, A.P. (1986). CpG-rich islands and the function of DNA methylation. Nature *321*, 209-213.

Bommer, G.T., Gerin, I., Feng, Y., Kaczorowski, A.J., Kuick, R., Love, R.E., Zhai, Y., Giordano, T.J., Qin, Z.S., Moore, B.B.*, et al.* (2007). p53-mediated activation of miRNA34 candidate tumor-suppressor genes. Curr Biol *17*, 1298-1307.

Branco, M.R., Ficz, G., and Reik, W. (2011). Uncovering the role of 5-hydroxymethylcytosine in the epigenome. Nat Rev Genet.

Brenner, D.R., McLaughlin, J.R., and Hung, R.J. (2011). Previous lung diseases and lung cancer risk: a systematic review and meta-analysis. PLoS One *6*, e17479.

Bryant, A., and Cerfolio, R.J. (2007). Differences in epidemiology, histology, and survival between cigarette smokers and never-smokers who develop non-small cell lung cancer. Chest *132*, 185-192.

Buckley, P.G., Das, S., Bryan, K., Watters, K.M., Alcock, L., Koster, J., Versteeg, R., and Stallings, R.L. (2011). Genome-wide DNA methylation analysis of neuroblastic tumors reveals clinically relevant epigenetic events and large-scale epigenomic alterations localized to telomeric regions. Int J Cancer *128*, 2296-2305.

Camilo, R., Capelozzi, V.L., Siqueira, S.A., and Del Carlo Bernardi, F. (2006). Expression of p63, keratin 5/6, keratin 7, and surfactant-A in non-small cell lung carcinomas. Hum Pathol *37*, 542-546.

Chatagnon, A., Perriaud, L., Nazaret, N., Croze, S., Benhattar, J., Lachuer, J., and Dante, R. (2011). Preferential binding of the methyl-CpG binding domain protein 2 at methylated transcriptional start site regions. Epigenetics *6*.

Chen, Y.J., Tang, Q.B., and Zou, S.Q. (2005). Inactivation of RASSF1A, the tumor suppressor gene at 3p21.3 in extrahepatic cholangiocarcinoma. World J Gastroenterol *11*, 1333-1338.

Chhatwani, L., Cabebe, E., and Wakelee, H.A. (2009). Adjuvant treatment of resected lung cancer. Proc Am Thorac Soc *6*, 194-200.

Choi, Y.L., Takeuchi, K., Soda, M., Inamura, K., Togashi, Y., Hatano, S., Enomoto, M., Hamada, T., Haruta, H., Watanabe, H.*, et al.* (2008). Identification of novel isoforms of the EML4-ALK transforming gene in non-small cell lung cancer. Cancer Res *68*, 4971-4976.

Christman, J.K. (2002). 5-Azacytidine and 5-aza-2'-deoxycytidine as inhibitors of DNA methylation: mechanistic studies and their implications for cancer therapy. Oncogene *21*, 5483-5495.

Ciardiello, F., and Tortora, G. (2008). EGFR antagonists in cancer treatment. N Engl J Med *358*, 1160-1174.

Cihak, A. (1974). Biological effects of 5-azacytidine in eukaryotes. Oncology *30*, 405-422.

Coate, L.E., John, T., Tsao, M.S., and Shepherd, F.A. (2009). Molecular predictive and prognostic markers in non-small-cell lung cancer. Lancet Oncol *10*, 1001-1010.

Cohen, O., Feinstein, E., and Kimchi, A. (1997). DAP-kinase is a Ca2+/calmodulin-dependent, cytoskeletal-associated protein kinase, with cell death-inducing functions that depend on its catalytic activity. EMBO J *16*, 998-1008.

Cohen, O., Inbal, B., Kissil, J.L., Raveh, T., Berissi, H., Spivak-Kroizaman, T., Feinstein, E., and Kimchi, A. (1999). DAP-kinase participates in TNF-alpha- and Fas-induced apoptosis and its function requires the death domain. J Cell Biol *146*, 141-148.

Coolen, M.W., Stirzaker, C., Song, J.Z., Statham, A.L., Kassir, Z., Moreno, C.S., Young, A.N., Varma, V., Speed, T.P., Cowley, M.*, et al.* (2010). Consolidation of the cancer genome into domains of repressive chromatin by long-range epigenetic silencing (LRES) reduces transcriptional plasticity. Nat Cell Biol *12*, 235-246.

Corney, D.C., Flesken-Nikitin, A., Godwin, A.K., Wang, W., and Nikitin, A.Y. (2007). MicroRNA-34b and MicroRNA-34c are targets of p53 and cooperate in control of cell proliferation and adhesion-independent growth. Cancer Res *67*, 8433-8438.

Cottrell, S.E., Distler, J., Goodman, N.S., Mooney, S.H., Kluth, A., Olek, A., Schwope, I., Tetzner, R., Ziebarth, H., and Berlin, K. (2004). A real-time PCR assay for DNA-methylation using methylation-specific blockers. Nucleic Acids Res *32*, e10.

Court, F., Baniol, M., Hagege, H., Petit, J.S., Lelay-Taha, M.N., Carbonell, F., Weber, M., Cathala, G., and Forne, T. (2011). Long-range chromatin interactions at the mouse Igf2/H19 locus reveal a novel paternally expressed long non-coding RNA. Nucleic Acids Res *39*, 5893-5906.

Cui, H., Cruz-Correa, M., Giardiello, F.M., Hutcheon, D.F., Kafonek, D.R., Brandenburg, S., Wu, Y., He, X., Powe, N.R., and Feinberg, A.P. (2003). Loss of IGF2 imprinting: a potential marker of colorectal cancer risk. Science *299*, 1753-1755.

Damelin, M., and Bestor, T.H. (2007). Biological functions of DNA methyltransferase 1 require its methyltransferase activity. Mol Cell Biol *27*, 3891-3899.

Dammann, R., Li, C., Yoon, J.H., Chin, P.L., Bates, S., and Pfeifer, G.P. (2000). Epigenetic inactivation of a RAS association domain family protein from the lung tumour suppressor locus 3p21.3. Nat Genet *25*, 315-319.

Darby, S., Hill, D., Auvinen, A., Barros-Dios, J.M., Baysson, H., Bochicchio, F., Deo, H., Falk, R., Forastiere, F., Hakama, M.*, et al.* (2005). Radon in homes and risk of lung cancer: collaborative analysis of individual data from 13 European case-control studies. BMJ *330*, 223.

Delmonte, V.C., Alberti, O., and Saldiva, P.H. (1986). Large cell carcinoma of the lung. Ultrastructural and immunohistochemical features. Chest *90*, 524-527.

Di Croce, L., Raker, V.A., Corsaro, M., Fazi, F., Fanelli, M., Faretta, M., Fuks, F., Lo Coco, F., Kouzarides, T., Nervi, C.*, et al.* (2002). Methyltransferase recruitment and DNA hypermethylation of target promoters by an oncogenic transcription factor. Science *295*, 1079-1082.

Dickstein, J., Senyuk, V., Premanand, K., Laricchia-Robbio, L., Xu, P., Cattaneo, F., Fazzina, R., and Nucifora, G. (2010). Methylation and silencing of miRNA-124 by EVI1 and self-renewal

exhaustion of hematopoietic stem cells in murine myelodysplastic syndrome. Proc Natl Acad Sci U S A *107*, 9783-9788.

Digel, W., and Lubbert, M. (2005). DNA methylation disturbances as novel therapeutic target in lung cancer: preclinical and clinical results. Crit Rev Oncol Hematol *55*, 1-11.

Ding, L., Getz, G., Wheeler, D.A., Mardis, E.R., McLellan, M.D., Cibulskis, K., Sougnez, C., Greulich, H., Muzny, D.M., Morgan, M.B.*, et al.* (2008). Somatic mutations affect key pathways in lung adenocarcinoma. Nature *455*, 1069-1075.

Doll, R., and Hill, A.B. (1950). Smoking and carcinoma of the lung; preliminary report. Br Med J *2*, 739-748.

Eads, C.A., Danenberg, K.D., Kawakami, K., Saltz, L.B., Blake, C., Shibata, D., Danenberg, P.V., and Laird, P.W. (2000). MethyLight: a high-throughput assay to measure DNA methylation. Nucleic Acids Res *28*, E32.

Esteller, M. (2007). Cancer epigenomics: DNA methylomes and histone-modification maps. Nat Rev Genet *8*, 286-298.

Esteller, M. (2011). Cancer Epigenetics for the 21st Century: What's Next? Genes Cancer *2*, 604-606.

Esteller, M., Toyota, M., Sanchez-Cespedes, M., Capella, G., Peinado, M.A., Watkins, D.N., Issa, J.P., Sidransky, D., Baylin, S.B., and Herman, J.G. (2000). Inactivation of the DNA repair gene O6-methylguanine-DNA methyltransferase by promoter hypermethylation is associated with G to A mutations in K-ras in colorectal tumorigenesis. Cancer Res *60*, 2368-2371.

Fackler, M.J., Umbricht, C.B., Williams, D., Argani, P., Cruz, L.A., Merino, V.F., Teo, W.W., Zhang, Z., Huang, P., Visvananthan, K.*, et al.* (2011). Genome-wide methylation analysis identifies genes specific to breast cancer hormone receptor status and risk of recurrence. Cancer Res *71*, 6195-6207.

Fenaux, P., Mufti, G.J., Hellstrom-Lindberg, E., Santini, V., Finelli, C., Giagounidis, A., Schoch, R., Gattermann, N., Sanz, G., List, A.*, et al.* (2009). Efficacy of azacitidine compared with that of conventional care regimens in the treatment of higher-risk myelodysplastic syndromes: a randomised, open-label, phase III study. Lancet Oncol *10*, 223-232.

Feng, S., Cong, S., Zhang, X., Bao, X., Wang, W., Li, H., Wang, Z., Wang, G., Xu, J., Du, B.*, et al.* (2011). MicroRNA-192 targeting retinoblastoma 1 inhibits cell proliferation and induces cell apoptosis in lung cancer cells. Nucleic Acids Res *39*, 6669-6678.

Ferlay, J., Shin, H.R., Bray, F., Forman, D., Mathers, C., and Parkin, D.M. (2010). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int J Cancer *127*, 2893-2917.

Filion, G.J., Zhenilo, S., Salozhin, S., Yamada, D., Prokhortchouk, E., and Defossez, P.A. (2006). A family of human zinc finger proteins that bind methylated DNA and repress transcription. Mol Cell Biol *26*, 169-181.

Finberg, K.E., Sequist, L.V., Joshi, V.A., Muzikansky, A., Miller, J.M., Han, M., Beheshti, J., Chirieac, L.R., Mark, E.J., and Iafrate, A.J. (2007). Mucinous differentiation correlates with absence of EGFR mutation and presence of KRAS mutation in lung adenocarcinomas with bronchioloalveolar features. J Mol Diagn *9*, 320-326.

Friedman, J.M., Liang, G., Liu, C.C., Wolff, E.M., Tsai, Y.C., Ye, W., Zhou, X., and Jones, P.A. (2009). The putative tumor suppressor microRNA-101 modulates the cancer epigenome by repressing the polycomb group protein EZH2. Cancer Res *69*, 2623-2629.

Fuks, F., Hurd, P.J., Wolf, D., Nan, X., Bird, A.P., and Kouzarides, T. (2003). The methyl-CpG-binding protein MeCP2 links DNA methylation to histone methylation. J Biol Chem *278*, 4035-4040.

Garzon, R., Liu, S., Fabbri, M., Liu, Z., Heaphy, C.E., Callegari, E., Schwind, S., Pang, J., Yu, J., Muthusamy, N.*, et al.* (2009). MicroRNA-29b induces global DNA hypomethylation and tumor

suppressor gene reexpression in acute myeloid leukemia by targeting directly DNMT3A and 3B and indirectly DNMT1. Blood *113*, 6411-6418.

Goldstraw, P., Crowley, J., Chansky, K., Giroux, D.J., Groome, P.A., Rami-Porta, R., Postmus, P.E., Rusch, V., and Sobin, L. (2007). The IASLC Lung Cancer Staging Project: proposals for the revision of the TNM stage groupings in the forthcoming (seventh) edition of the TNM Classification of malignant tumours. J Thorac Oncol *2*, 706-714.

Grote, H.J., Schmiemann, V., Geddert, H., Rohr, U.P., Kappes, R., Gabbert, H.E., and Bocking, A. (2005). Aberrant promoter methylation of p16(INK4a), RARB2 and SEMA3B in bronchial aspirates from patients with suspected lung cancer. Int J Cancer *116*, 720-725.

Gu, J., Berman, D., Lu, C., Wistuba, II, Roth, J.A., Frazier, M., Spitz, M.R., and Wu, X. (2006). Aberrant promoter methylation profile and association with survival in patients with non-small cell lung cancer. Clin Cancer Res *12*, 7329-7338.

Gu, P., Xu, X., Le Menuet, D., Chung, A.C., and Cooney, A.J. (2011). Differential recruitment of methyl CpG-binding domain factors and DNA methyltransferases by the orphan receptor germ cell nuclear factor initiates the repression and silencing of Oct4. Stem Cells *29*, 1041-1051.

Gyorffy, E., Anna, L., Kovacs, K., Rudnai, P., and Schoket, B. (2008). Correlation between biomarkers of human exposure to genotoxins with focus on carcinogen-DNA adducts. Mutagenesis *23*, 1-18.

Hainaut, P., Soussi, T., Shomer, B., Hollstein, M., Greenblatt, M., Hovig, E., Harris, C.C., and Montesano, R. (1997). Database of p53 gene somatic mutations in human tumors and cell lines: updated compilation and future prospects. Nucleic Acids Res *25*, 151-157.

Hainaut, P., and Wiman, K.G. (2009). 30 years and a long way into p53 research. Lancet Oncol *10*, 913-919.

Han, H., Cortez, C.C., Yang, X., Nichols, P.W., Jones, P.A., and Liang, G. (2011). DNA methylation directly silences genes with non-CpG island promoters and establishes a nucleosome occupied promoter. Hum Mol Genet *20*, 4299-4310.

Harbour, J.W., Lai, S.L., Whang-Peng, J., Gazdar, A.F., Minna, J.D., and Kaye, F.J. (1988). Abnormalities in structure and expression of the human retinoblastoma gene in SCLC. Science *241*, 353-357.

Harris, J.E., Thun, M.J., Mondul, A.M., and Calle, E.E. (2004). Cigarette tar yields in relation to mortality from lung cancer in the cancer prevention study II prospective cohort, 1982-8. BMJ *328*, 72.

Hayatsu, H., Wataya, Y., Kai, K., and Iida, S. (1970). Reaction of sodium bisulfite with uracil, cytosine, and their derivatives. Biochemistry *9*, 2858-2865.

Hecht, S.S. (2003). Tobacco carcinogens, their biomarkers and tobacco-induced cancer. Nat Rev Cancer *3*, 733-744.

Helicos (2011) (Cambridge, MA, USA). http://www.helicosbio.com/Products/HelicosregGenetic AnalysisSystem/tabid/140/Default.aspx. Accession date: 26/11/2011.

Herbst, R.S., Heymach, J.V., and Lippman, S.M. (2008). Lung cancer. N Engl J Med *359*, 1367-1380.

Herman, J.G., Graff, J.R., Myohanen, S., Nelkin, B.D., and Baylin, S.B. (1996). Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. Proc Natl Acad Sci U S A *93*, 9821-9826.

Hernandez-Vargas, H., Ouzounova, M., Le Calvez-Kelm, F., Lambert, M.P., McKay-Chopin, S., Tavtigian, S.V., Puisieux, A., Matar, C., and Herceg, Z. (2011). Methylome analysis reveals Jak-STAT pathway deregulation in putative breast cancer stem cells. Epigenetics *6*, 428-439.

Hill, V.K., Ricketts, C., Bieche, I., Vacher, S., Gentle, D., Lewis, C., Maher, E.R., and Latif, F. (2011). Genome-wide DNA methylation profiling of CpG islands in breast cancer identifies novel genes associated with tumorigenicity. Cancer Res *71*, 2988-2999.

Hirsch, F.R., Herbst, R.S., Olsen, C., Chansky, K., Crowley, J., Kelly, K., Franklin, W.A., Bunn, P.A., Jr., Varella-Garcia, M., and Gandara, D.R. (2008). Increased EGFR gene copy number detected by fluorescent in situ hybridization predicts outcome in non-small-cell lung cancer patients treated with cetuximab and chemotherapy. J Clin Oncol *26*, 3351-3357.

Hofseth, L.J., Hussain, S.P., and Harris, C.C. (2004). p53: 25 years after its discovery. Trends Pharmacol Sci *25*, 177-181.

Holland, J.F. (2000). Cancer medicine, Vol Chapter 88: Cancer of the Lung, 5th edn (Hamilton, Ontario ; New York, B.C. Decker).

Honorio, S., Agathanggelou, A., Schuermann, M., Pankow, W., Viacava, P., Maher, E.R., and Latif, F. (2003). Detection of RASSF1A aberrant promoter hypermethylation in sputum from chronic smokers and ductal carcinoma in situ from breast cancer patients. Oncogene *22*, 147-150.

Horn, L., and Pao, W. (2009). EML4-ALK: honing in on a new target in non-small-cell lung cancer. J Clin Oncol *27*, 4232-4235.

Hou, J., Aerts, J., den Hamer, B., van Ijcken, W., den Bakker, M., Riegman, P., van der Leest, C., van der Spek, P., Foekens, J.A., Hoogsteden, H.C*., et al.* (2010). Gene expression-based classification of non-small cell lung carcinomas and survival prediction. PLoS One *5*, e10312.

Huang, C.L., Taki, T., Adachi, M., Konishi, T., Higashiyama, M., Kinoshita, M., Hadama, T., and Miyake, M. (1998). Mutations of p53 and K-ras genes as prognostic factors for non-small cell lung cancer. Int J Oncol *12*, 553-563.

Huang, L.W., Pan, H.S., Lin, Y.H., Seow, K.M., Chen, H.J., and Hwang, J.L. (2011). P16 methylation is an early event in cervical carcinogenesis. Int J Gynecol Cancer *21*, 452-456.

Hublitz, P., Albert, M., and Peters, A.H. (2009). Mechanisms of transcriptional repression by histone lysine methylation. Int J Dev Biol *53*, 335-354.

Ignatov, A., Bischoff, J., Ignatov, T., Schwarzenau, C., Krebs, T., Kuester, D., Costa, S.D., Roessner, A., Semczuk, A., and Schneider-Stock, R. (2010). APC promoter hypermethylation is an early event in endometrial tumorigenesis. Cancer Sci *101*, 321-327.

Illumina (2011). Technology / Sequencing Technology (http://www.illumina.com/systems/genome_analyzer_iix.ilmn). Accession date: 24/11/2011.

Jemal, A., Siegel, R., Ward, E., Murray, T., Xu, J., Smigal, C., and Thun, M.J. (2006). Cancer statistics, 2006. CA Cancer J Clin *56*, 106-130.

Ji, H., Ramsey, M.R., Hayes, D.N., Fan, C., McNamara, K., Kozlowski, P., Torrice, C., Wu, M.C., Shimamura, T., Perera, S.A*., et al.* (2007). LKB1 modulates lung cancer differentiation and metastasis. Nature *448*, 807-810.

Ji, M., Guan, H., Gao, C., Shi, B., and Hou, P. (2011). Highly frequent promoter methylation and PIK3CA amplification in non-small cell lung cancer (NSCLC). BMC Cancer *11*, 147.

Jiang, C., and Pugh, B.F. (2009). Nucleosome positioning and gene regulation: advances through genomics. Nat Rev Genet *10*, 161-172.

Jin, S., and Levine, A.J. (2001). The p53 functional circuit. J Cell Sci *114*, 4139-4140.

Johnson, L., Mercer, K., Greenbaum, D., Bronson, R.T., Crowley, D., Tuveson, D.A., and Jacks, T. (2001). Somatic activation of the K-ras oncogene causes early onset lung cancer in mice. Nature *410*, 1111-1116.

Juergens, R.A., and Raines, C.P. (2010). Trial of Adjuvant Combined Epigenetic Therapy With 5-azacitidine and Entinostat in Resected Stage I Non-small Cell Lung Cancer Versus Standard Care (J1037), NIH, ed.

Juergens, R.A., Vendetti, F., Coleman, B., Sebree, R.S., Rudek, M.A., Belinsky, S.A., Brock, M.V., Herman, J.G., B., B.S., and Rudin, C.M. (2008). Phase I trial of 5-azacitidine (5AC) and SNDX-275 in advanced lung cancer (NSCLC) Journal of Clinical Oncology *26*.

Kawai, J., Hirose, K., Fushiki, S., Hirotsune, S., Ozawa, N., Hara, A., Hayashizaki, Y., and Watanabe, S. (1994). Comparison of DNA methylation patterns among mouse cell lines by restriction landmark genomic scanning. Mol Cell Biol *14*, 7421-7427.

Kenfield, S.A., Wei, E.K., Stampfer, M.J., Rosner, B.A., and Colditz, G.A. (2008). Comparison of aspects of smoking among the four histological types of lung cancer. Tob Control *17*, 198-204.

Kersting, M., Friedl, C., Kraus, A., Behn, M., Pankow, W., and Schuermann, M. (2000). Differential frequencies of p16(INK4a) promoter hypermethylation, p53 mutation, and K-ras mutation in exfoliative material mark the development of lung cancer in symptomatic chronic smokers. J Clin Oncol *18*, 3221-3229.

Kim, D.H., Kim, J.S., Ji, Y.I., Shim, Y.M., Kim, H., Han, J., and Park, J. (2003). Hypermethylation of RASSF1A promoter is associated with the age at starting smoking and a poor prognosis in primary non-small cell lung cancer. Cancer Res *63*, 3743-3746.

Kim, D.H., Nelson, H.H., Wiencke, J.K., Christiani, D.C., Wain, J.C., Mark, E.J., and Kelsey, K.T. (2001). Promoter methylation of DAP-kinase: association with advanced stage in non-small cell lung cancer. Oncogene *20*, 1765-1770.

Kim, E.H., Park, A.K., Dong, S.M., Ahn, J.H., and Park, W.Y. (2010). Global analysis of CpG methylation reveals epigenetic control of the radiosensitivity in lung cancer cell lines. Oncogene *29*, 4725-4731.

Kim, E.S., Hirsh, V., Mok, T., Socinski, M.A., Gervais, R., Wu, Y.L., Li, L.Y., Watkins, C.L., Sellers, M.V., Lowe, E.S*., et al.* (2008). Gefitinib versus docetaxel in previously treated non-small-cell lung cancer (INTEREST): a randomised phase III trial. Lancet *372*, 1809-1818.

Kim, J.S., Han, J., Shim, Y.M., Park, J., and Kim, D.H. (2005). Aberrant methylation of H-cadherin (CDH13) promoter is associated with tumor progression in primary nonsmall cell lung carcinoma. Cancer *104*, 1825-1833.

Koivunen, J.P., Mermel, C., Zejnullahu, K., Murphy, C., Lifshits, E., Holmes, A.J., Choi, H.G., Kim, J., Chiang, D., Thomas, R*., et al.* (2008). EML4-ALK fusion gene and efficacy of an ALK kinase inhibitor in lung cancer. Clin Cancer Res *14*, 4275-4283.

Kouzarides, T. (2007). Chromatin modifications and their function. Cell *128*, 693-705.

Kristensen, L.S., Mikeska, T., Krypuy, M., and Dobrovic, A. (2008). Sensitive Melting Analysis after Real Time- Methylation Specific PCR (SMART-MSP): high-throughput and probe-free quantitative DNA methylation detection. Nucleic Acids Res *36*, e42.

Lacroix, L., Commo, F., and Soria, J.C. (2008). Gene expression profiling of non-small-cell lung cancer. Expert Rev Mol Diagn *8*, 167-178.

Laird, P.W. (2010). Principles and challenges of genomewide DNA methylation analysis. Nat Rev Genet *11*, 191-203.

Lamy, A., Sesboue, R., Bourguignon, J., Dautreaux, B., Metayer, J., Frebourg, T., and Thiberville, L. (2002). Aberrant methylation of the CDKN2a/p16INK4a gene promoter region in preinvasive bronchial lesions: a prospective study in high-risk patients without invasive cancer. Int J Cancer *100*, 189-193.

Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell *75*, 843-854.

Levine, A.J. (1997). p53, the cellular gatekeeper for growth and division. Cell *88*, 323-331.

Li, E. (2002). Chromatin modification and epigenetic reprogramming in mammalian development. Nat Rev Genet *3*, 662-673.

Li, E., Bestor, T.H., and Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell *69*, 915-926.

Li, X., and Carthew, R.W. (2005). A microRNA mediates EGF receptor signaling and promotes photoreceptor differentiation in the Drosophila eye. Cell *123*, 1267-1277.

Liang, G., Lin, J.C., Wei, V., Yoo, C., Cheng, J.C., Nguyen, C.T., Weisenberger, D.J., Egger, G., Takai, D., Gonzales, F.A.*, et al.* (2004). Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome. Proc Natl Acad Sci U S A *101*, 7357-7362.

Liang, H.Y., Li, X.L., Yu, X.S., Guan, P., Yin, Z.H., He, Q.C., and Zhou, B.S. (2009). Facts and fiction of the relationship between preexisting tuberculosis and lung cancer risk: a systematic review. Int J Cancer *125*, 2936-2944.

Linardou, H., Dahabreh, I.J., Bafaloukos, D., Kosmidis, P., and Murray, S. (2009). Somatic EGFR mutations and efficacy of tyrosine kinase inhibitors in NSCLC. Nat Rev Clin Oncol *6*, 352-366.

Lissowska, J., Bardin-Mikolajczak, A., Fletcher, T., Zaridze, D., Szeszenia-Dabrowska, N., Rudnai, P., Fabianova, E., Cassidy, A., Mates, D., Holcatova, I.*, et al.* (2005). Lung cancer and indoor pollution from heating and cooking with solid fuels: the IARC international multicentre case-control study in Eastern/Central Europe and the United Kingdom. Am J Epidemiol *162*, 326-333.

Littman, A.J., White, E., Jackson, L.A., Thornquist, M.D., Gaydos, C.A., Goodman, G.E., and Vaughan, T.L. (2004). Chlamydia pneumoniae infection and risk of lung cancer. Cancer Epidemiol Biomarkers Prev *13*, 1624-1630.

Lopez-Serra, L., and Esteller, M. (2008). Proteins that bind methylated DNA and human cancer: reading the wrong words. Br J Cancer *98*, 1881-1885.

Lu, C., Soria, J.C., Tang, X., Xu, X.C., Wang, L., Mao, L., Lotan, R., Kemp, B., Bekele, B.N., Feng, L.*, et al.* (2004). Prognostic factors in resected stage I non-small-cell lung cancer: a multivariate analysis of six molecular markers. J Clin Oncol *22*, 4575-4583.

Lukas, J., Parry, D., Aagaard, L., Mann, D.J., Bartkova, J., Strauss, M., Peters, G., and Bartek, J. (1995). Retinoblastoma-protein-dependent cell-cycle inhibition by the tumour suppressor p16. Nature *375*, 503-506.

Marsit, C.J., Hasegawa, M., Hirao, T., Kim, D.H., Aldape, K., Hinds, P.W., Wiencke, J.K., Nelson, H.H., and Kelsey, K.T. (2004). Loss of heterozygosity of chromosome 3p21 is associated with mutant TP53 and better patient survival in non-small-cell lung cancer. Cancer Res *64*, 8702-8707.

Maruyama, R., Sugio, K., Yoshino, I., Maehara, Y., and Gazdar, A.F. (2004). Hypermethylation of FHIT as a prognostic marker in nonsmall cell lung carcinoma. Cancer *100*, 1472-1477.

Metzker, M.L. (2010). Sequencing technologies - the next generation. Nat Rev Genet *11*, 31-46.

Miura, A., Yonebayashi, S., Watanabe, K., Toyama, T., Shimada, H., and Kakutani, T. (2001). Mobilization of transposons by a mutation abolishing full DNA methylation in Arabidopsis. Nature *411*, 212-214.

Mogi, A., and Kuwano, H. (2011). TP53 mutations in nonsmall cell lung cancer. J Biomed Biotechnol *2011*, 583929.

Momparler, R.L., Eliopoulos, N., and Ayoub, J. (2000). Evaluation of an inhibitor of DNA methylation, 5-aza-2'-deoxycytidine, for the treatment of lung cancer and the future role of gene therapy. Adv Exp Med Biol *465*, 433-446.

Morgan, H.D., Santos, F., Green, K., Dean, W., and Reik, W. (2005). Epigenetic reprogramming in mammals. Hum Mol Genet *14 Spec No 1*, R47-58.

Mukai, F., Hawryluk, I., and Shapiro, R. (1970). The mutagenic specificity of sodium bisulfite. Biochem Biophys Res Commun *39*, 983-988.

Nguewa, P.A., Calvo, A., Pullamsetti, S.S., Banat, G.A., Grimminger, F., and Savai, R. (2011). Tyrosine kinase inhibitors with antiangiogenic properties for the treatment of non-small cell lung cancer. Expert Opin Investig Drugs *20*, 61-74.

Niklinska, W., Naumnik, W., Sulewska, A., Kozlowski, M., Pankiewicz, W., and Milewski, R. (2009). Prognostic significance of DAPK and RASSF1A promoter hypermethylation in non-small cell lung cancer (NSCLC). Folia Histochem Cytobiol *47*, 275-280.

Nitadori, J., Inoue, M., Iwasaki, M., Otani, T., Sasazuki, S., Nagai, K., and Tsugane, S. (2006). Association between lung cancer incidence and family history of lung cancer: data from a large-scale population-based cohort study, the JPHC study. Chest *130*, 968-975.

Nordquist, L.T., Simon, G.R., Cantor, A., Alberts, W.M., and Bepler, G. (2004). Improved survival in never-smokers vs current smokers with primary adenocarcinoma of the lung. Chest *126*, 347-351.

Nykamp, V. (2010). Surgical treatment of non-small cell lung cancer. S D Med *Spec No*, 57-59.

Ohm, J.E., McGarvey, K.M., Yu, X., Cheng, L., Schuebel, K.E., Cope, L., Mohammad, H.P., Chen, W., Daniel, V.C., Yu, W.*, et al.* (2007). A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. Nat Genet *39*, 237-242.

Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell *99*, 247-257.

Olivier, M., Hollstein, M., and Hainaut, P. (2010). TP53 mutations in human cancers: origins, consequences, and clinical use. Cold Spring Harb Perspect Biol *2*, a001008.

Olivo-Marston, S.E., Yang, P., Mechanic, L.E., Bowman, E.D., Pine, S.R., Loffredo, C.A., Alberg, A.J., Caporaso, N., Shields, P.G., Chanock, S.*, et al.* (2009). Childhood exposure to secondhand smoke and functional mannose binding lectin polymorphisms are associated with increased lung cancer risk. Cancer Epidemiol Biomarkers Prev *18*, 3375-3383.

Palmisano, W.A., Divine, K.K., Saccomanno, G., Gilliland, F.D., Baylin, S.B., Herman, J.G., and Belinsky, S.A. (2000). Predicting lung cancer by detecting aberrant promoter methylation in sputum. Cancer Res *60*, 5954-5958.

Pao, W., Miller, V.A., Politi, K.A., Riely, G.J., Somwar, R., Zakowski, M.F., Kris, M.G., and Varmus, H. (2005). Acquired resistance of lung adenocarcinomas to gefitinib or erlotinib is associated with a second mutation in the EGFR kinase domain. PLoS Med *2*, e73.

Park, T.J., Han, S.U., Cho, Y.K., Paik, W.K., Kim, Y.B., and Lim, I.K. (2001). Methylation of O(6)-methylguanine-DNA methyltransferase gene is associated significantly with K-ras mutation, lymph node invasion, tumor staging, and disease free survival in patients with gastric carcinoma. Cancer *92*, 2760-2768.

Peto, R. (1986). Influence of dose and duration of smoking on lung cancer rates. IARC Sci Publ, 23-33.

Pfannschmidt, J., Muley, T., Bulzebruck, H., Hoffmann, H., and Dienemann, H. (2007). Prognostic assessment after surgical resection for non-small cell lung cancer: experiences in 2083 patients. Lung Cancer *55*, 371-377.

Pulling, L.C., Divine, K.K., Klinge, D.M., Gilliland, F.D., Kang, T., Schwartz, A.G., Bocklage, T.J., and Belinsky, S.A. (2003). Promoter hypermethylation of the O6-methylguanine-DNA methyltransferase gene: more common in lung adenocarcinomas from never-smokers than smokers and associated with tumor progression. Cancer Res *63*, 4842-4848.

Qiu, J. (2006). Epigenetics: unfinished symphony. Nature *441*, 143-145.

Rauch, T., Wang, Z., Zhang, X., Zhong, X., Wu, X., Lau, S.K., Kernstine, K.H., Riggs, A.D., and Pfeifer, G.P. (2007). Homeobox gene methylation in lung cancer studied by genome-wide analysis with a microarray-based methylated CpG island recovery assay. Proc Natl Acad Sci U S A *104*, 5527-5532.

Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E., Horvitz, H.R., and Ruvkun, G. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature *403*, 901-906.

Riley, T., Sontag, E., Chen, P., and Levine, A. (2008). Transcriptional control of human p53-regulated genes. Nat Rev Mol Cell Biol *9*, 402-412.

Robles, A.I., and Harris, C.C. (2010). Clinical outcomes and correlates of TP53 mutations and cancer. Cold Spring Harb Perspect Biol *2*, a001016.

Roche (2011). http://www.my454.com/ (accession: 09/11/2011).

Roh, M.H., Schmidt, L., Placido, J., Farmen, S., Fields, K.L., Courey, A.J., Arenberg, D.A., and Knoepp, S.M. (2011). The application and diagnostic utility of immunocytochemistry on direct smears in the diagnosis of pulmonary adenocarcinoma and squamous cell carcinoma. Diagn Cytopathol.

Rossi, G., Pelosi, G., Graziano, P., Barbareschi, M., and Papotti, M. (2009). A reevaluation of the clinical significance of histological subtyping of non--small-cell lung carcinoma: diagnostic algorithms in the era of personalized treatments. Int J Surg Pathol *17*, 206-218.

Sagerup, C.M., Smastuen, M., Johannesen, T.B., Helland, A., and Brustugun, O.T. (2011). Sex-specific trends in lung cancer incidence and survival: a population study of 40,118 cases. Thorax *66*, 301-307.

Said, J.W., Nash, G., Tepper, G., and Banks-Schlegel, S. (1983). Keratin proteins and carcinoembryonic antigen in lung carcinoma: an immunoperoxidase study of fifty-four cases, with ultrastructural correlations. Hum Pathol *14*, 70-76.

Sakurai, H., Asamura, H., Watanabe, S., Suzuki, K., and Tsuchiya, R. (2004). Clinicopathologic features of peripheral squamous cell carcinoma of the lung. Ann Thorac Surg *78*, 222-227.

Samanic, C., Rusiecki, J., Dosemeci, M., Hou, L., Hoppin, J.A., Sandler, D.P., Lubin, J., Blair, A., and Alavanja, M.C. (2006). Cancer incidence among pesticide applicators exposed to dicamba in the agricultural health study. Environ Health Perspect *114*, 1521-1526.

Samet, J.M. (2011). Radiation and cancer risk: a continuing challenge for epidemiologists. Environ Health *10 Suppl 1*, S4.

Samet, J.M., Avila-Tang, E., Boffetta, P., Hannan, L.M., Olivo-Marston, S., Thun, M.J., and Rudin, C.M. (2009). Lung cancer in never smokers: clinical epidemiology and environmental risk factors. Clin Cancer Res *15*, 5626-5645.

Scagliotti, G.V., Pastorino, U., Vansteenkiste, J.F., Spaggiari, L., Facciolo, F., Orlowski, T.M., Maiorino, L., Hetzel, M., Leschinger, M., Visseren-Grul, C.*, et al.* (2011). Randomized Phase III Study of Surgery Alone or Surgery Plus Preoperative Cisplatin and Gemcitabine in Stages IB to IIIA Non-Small-Cell Lung Cancer. J Clin Oncol.

Schlesinger, Y., Straussman, R., Keshet, I., Farkash, S., Hecht, M., Zimmerman, J., Eden, E., Yakhini, Z., Ben-Shushan, E., Reubinoff, B.E.*, et al.* (2007). Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. Nat Genet *39*, 232-236.

Schmittgen, T.D. (2008). Regulation of microRNA processing in development, differentiation and cancer. J Cell Mol Med *12*, 1811-1819.

Schubeler, D., Francastel, C., Cimbora, D.M., Reik, A., Martin, D.I., and Groudine, M. (2000). Nuclear localization and histone acetylation: a pathway for chromatin opening and transcriptional activation of the human beta-globin locus. Genes Dev *14*, 940-950.

Seng, T.J., Currey, N., Cooper, W.A., Lee, C.S., Chan, C., Horvath, L., Sutherland, R.L., Kennedy, C., McCaughan, B., and Kohonen-Corish, M.R. (2008). DLEC1 and MLH1 promoter methylation are associated with poor prognosis in non-small cell lung carcinoma. Br J Cancer *99*, 375-382.

Sequist, L.V., Bell, D.W., Lynch, T.J., and Haber, D.A. (2007). Molecular predictors of response to epidermal growth factor receptor antagonists in non-small-cell lung cancer. J Clin Oncol *25*, 587-595.

Sharma, S., Kelly, T.K., and Jones, P.A. (2010). Epigenetics in cancer. Carcinogenesis *31*, 27-36.

Shaw, A.T., Yeap, B.Y., Mino-Kenudson, M., Digumarthy, S.R., Costa, D.B., Heist, R.S., Solomon, B., Stubbs, H., Admane, S., McDermott, U*., et al.* (2009). Clinical features and outcome of patients with non-small-cell lung cancer who harbor EML4-ALK. J Clin Oncol *27*, 4247-4253.

Shepherd, F.A., Crowley, J., Van Houtte, P., Postmus, P.E., Carney, D., Chansky, K., Shaikh, Z., and Goldstraw, P. (2007). The International Association for the Study of Lung Cancer lung cancer staging project: proposals regarding the clinical staging of small cell lung cancer in the forthcoming (seventh) edition of the tumor, node, metastasis classification for lung cancer. J Thorac Oncol *2*, 1067-1077.

Shukla, V., Vaissiere, T., and Herceg, Z. (2008). Histone acetylation and chromatin signature in stem cell identity and cancer. Mutat Res *637*, 1-15.

Soda, M., Choi, Y.L., Enomoto, M., Takada, S., Yamashita, Y., Ishikawa, S., Fujiwara, S., Watanabe, H., Kurashina, K., Hatanaka, H*., et al.* (2007). Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature *448*, 561-566.

Soria, J.C., Rodriguez, M., Liu, D.D., Lee, J.J., Hong, W.K., and Mao, L. (2002). Aberrant promoter methylation of multiple genes in bronchial brush samples from former cigarette smokers. Cancer Res *62*, 351-355.

Soussi, T., Ishioka, C., Claustres, M., and Beroud, C. (2006). Locus-specific mutation databases: pitfalls and good practice based on the p53 experience. Nat Rev Cancer *6*, 83-90.

Stahel, R.A. (1992). Morphology, surface antigens, staging, and prognostic factors of small cell lung cancer. Curr Opin Oncol *4*, 308-314.

Steels, E., Paesmans, M., Berghmans, T., Branle, F., Lemaitre, F., Mascaux, C., Meert, A.P., Vallot, F., Lafitte, J.J., and Sculier, J.P. (2001). Role of p53 as a prognostic factor for survival in lung cancer: a systematic review of the literature with a meta-analysis. Eur Respir J *18*, 705-719.

Stephens, P., Hunter, C., Bignell, G., Edkins, S., Davies, H., Teague, J., Stevens, C., O'Meara, S., Smith, R., Parker, A*., et al.* (2004). Lung cancer: intragenic ERBB2 kinase mutations in tumours. Nature *431*, 525-526.

Sternglanz, R. (1996). Histone acetylation: a gateway to transcriptional activation. Trends Biochem Sci *21*, 357-358.

Stewart, M.D., Li, J., and Wong, J. (2005). Relationship between histone H3 lysine 9 methylation, transcription repression, and heterochromatin protein 1 recruitment. Mol Cell Biol *25*, 2525-2538.

Subramanian, J., and Govindan, R. (2007). Lung cancer in never smokers: a review. J Clin Oncol *25*, 561-570.

Suzuki, M.M., and Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet *9*, 465-476.

Sy, S.M., Wong, N., Lee, T.W., Tse, G., Mok, T.S., Fan, B., Pang, E., Johnson, P.J., and Yim, A. (2004). Distinct patterns of genetic alterations in adenocarcinoma and squamous cell carcinoma of the lung. Eur J Cancer *40*, 1082-1094.

Taeger, D., Fritsch, A., Wiethege, T., Johnen, G., Eisenmenger, A., Wesch, H., Ko, Y., Stier, S., Michael Muller, K., Bruning, T., *et al.* (2006). Role of exposure to radon and silicosis on the cell type of lung carcinoma in German uranium miners. Cancer *106*, 881-889.

Takahashi, T., Sonobe, M., Kobayashi, M., Yoshizawa, A., Menju, T., Nakayama, E., Mino, N., Iwakiri, S., Sato, K., Miyahara, R., *et al.* (2010). Clinicopathologic features of non-small-cell lung cancer with EML4-ALK fusion gene. Ann Surg Oncol *17*, 889-897.

Takebayashi, S., Tamura, T., Matsuoka, C., and Okano, M. (2007). Major and essential role for the DNA methylation mark in mouse embryogenesis and stable association of DNMT1 with newly replicated regions. Mol Cell Biol *27*, 8243-8258.

Tammemagi, C.M., Neslund-Dudas, C., Simoff, M., and Kvale, P. (2004). Smoking and lung cancer survival: the role of comorbidity and treatment. Chest *125*, 27-37.

Tang, X., Khuri, F.R., Lee, J.J., Kemp, B.L., Liu, D., Hong, W.K., and Mao, L. (2000). Hypermethylation of the death-associated protein (DAP) kinase promoter and aggressiveness in stage I non-small-cell lung cancer. J Natl Cancer Inst *92*, 1511-1516.

Taylor, R., Najafi, F., and Dobson, A. (2007). Meta-analysis of studies of passive smoking and lung cancer: effects of study type and continent. Int J Epidemiol *36*, 1048-1059.

Teodoridis, J.M., Hardie, C., and Brown, R. (2008). CpG island methylator phenotype (CIMP) in cancer: causes and implications. Cancer Lett *268*, 177-186.

Terasaki, H., Niki, T., Matsuno, Y., Yamada, T., Maeshima, A., Asamura, H., Hayabuchi, N., and Hirohashi, S. (2003). Lung adenocarcinoma with mixed bronchioloalveolar and invasive components: clinicopathological features, subclassification by extent of invasive foci, and immunohistochemical characterization. Am J Surg Pathol *27*, 937-951.

Terry, J., Leung, S., Laskin, J., Leslie, K.O., Gown, A.M., and Ionescu, D.N. (2010). Optimal immunohistochemical markers for distinguishing lung adenocarcinomas from squamous cell carcinomas in small tumor samples. Am J Surg Pathol *34*, 1805-1811.

Thomassin, H., Kress, C., and Grange, T. (2004). MethylQuant: a sensitive method for quantifying methylation of specific cytosines within the genome. Nucleic Acids Res *32*, e168.

Toyooka, S., Toyooka, K.O., Maruyama, R., Virmani, A.K., Girard, L., Miyajima, K., Harada, K., Ariyoshi, Y., Takahashi, T., Sugio, K., *et al.* (2001). DNA methylation profiles of lung tumors. Mol Cancer Ther *1*, 61-67.

Toyooka, S., Toyooka, K.O., Miyajima, K., Reddy, J.L., Toyota, M., Sathyanarayana, U.G., Padar, A., Tockman, M.S., Lam, S., Shivapurkar, N., *et al.* (2003a). Epigenetic down-regulation of death-associated protein kinase in lung cancers. Clin Cancer Res *9*, 3034-3041.

Toyooka, S., Tsuda, T., and Gazdar, A.F. (2003b). The TP53 gene, tobacco exposure, and lung cancer. Hum Mutat *21*, 229-239.

Travis, W.D. (2002). Pathology of lung cancer. Clin Chest Med *23*, 65-81, viii.

Travis, W.D., Brambilla, E., Noguchi, M., Nicholson, A.G., Geisinger, K.R., Yatabe, Y., Beer, D.G., Powell, C.A., Riely, G.J., Van Schil, P.E., *et al.* (2011). International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. J Thorac Oncol *6*, 244-285.

Usadel, H., Brabender, J., Danenberg, K.D., Jeronimo, C., Harden, S., Engles, J., Danenberg, P.V., Yang, S., and Sidransky, D. (2002). Quantitative adenomatous polyposis coli promoter methylation analysis in tumor tissue, serum, and plasma DNA of patients with lung cancer. Cancer Res *62*, 371-375.

Viktorsson, K., De Petris, L., and Lewensohn, R. (2005). The role of p53 in treatment responses of lung cancer. Biochem Biophys Res Commun *331*, 868-880.

Vineis, P., and Husgafvel-Pursiainen, K. (2005). Air pollution and cancer: biomarker studies in human populations. Carcinogenesis *26*, 1846-1855.

Virmani, A.K., Rathi, A., Zochbauer-Muller, S., Sacchi, N., Fukuyama, Y., Bryant, D., Maitra, A., Heda, S., Fong, K.M., Thunnissen, F.*, et al.* (2000). Promoter methylation and silencing of the retinoic acid receptor-beta gene in lung carcinomas. J Natl Cancer Inst *92*, 1303-1307.

Vogelstein, B., Lane, D., and Levine, A.J. (2000). Surfing the p53 network. Nature *408*, 307-310.

Waddington, C.H. (1942). The epigenotype. Endeavour *1*, 18-20.

Wang, J., Wang, B., Chen, X., and Bi, J. (2011). The prognostic value of RASSF1A promoter hypermethylation in non-small cell lung carcinoma: a systematic review and meta-analysis. Carcinogenesis *32*, 411-416.

Wang, T., Nelson, R.A., Bogardus, A., and Grannis, F.W., Jr. (2010). Five-year lung cancer survival: which advanced stage nonsmall cell lung cancer patients attain long-term survival? Cancer *116*, 1518-1525.

Wang, Y., and Leung, F.C. (2004). An evaluation of new criteria for CpG islands in the human genome as gene markers. Bioinformatics *20*, 1170-1177.

Watanabe, K., Emoto, N., Hamano, E., Sunohara, M., Kawakami, M., Kage, H., Kitano, K., Nakajima, J., Goto, A., Fukayama, M.*, et al.* (2011). Genome structure-based screening identified epigenetically silenced microRNA associated with invasiveness in non-small-cell lung cancer. Int J Cancer.

Weinberg, R.A. (1995). The retinoblastoma protein and cell cycle control. Cell *81*, 323-330.

Wiencke, J.K., Thurston, S.W., Kelsey, K.T., Varkonyi, A., Wain, J.C., Mark, E.J., and Christiani, D.C. (1999). Early age at smoking initiation and tobacco carcinogen DNA damage in the lung. J Natl Cancer Inst *91*, 614-619.

Wong, K.Y., Yu, L., and Chim, C.S. (2011). DNA methylation of tumor suppressor miRNA genes: a lesson from the miR-34 family. Epigenomics *3*, 83-92.

Xiong, Z., and Laird, P.W. (1997). COBRA: a sensitive and quantitative DNA methylation assay. Nucleic Acids Res *25*, 2532-2534.

Xu, Y., Hu, B., Choi, A.J., Gopalan, B., Lee, B.H., Kalady, M.F., Church, J.M., and Ting, A.H. (2011). Unique DNA methylome profiles in CpG island methylator phenotype colon cancers. Genome Res.

Yatabe, Y., Mitsudomi, T., and Takahashi, T. (2002). TTF-1 expression in pulmonary adenocarcinomas. Am J Surg Pathol *26*, 767-773.

Yim, J., Zhu, L.C., Chiriboga, L., Watson, H.N., Goldberg, J.D., and Moreira, A.L. (2007). Histologic features are important prognostic indicators in early stages lung adenocarcinomas. Mod Pathol *20*, 233-241.

Yoshida, Y., Kokubu, A., Suzuki, K., Kuribayashi, H., Tsuta, K., Matsuno, Y., Kusumoto, M., Kanai, Y., Asamura, H., Hirohashi, S.*, et al.* (2007). Molecular markers and changes of computed tomography appearance in lung adenocarcinoma with ground-glass opacity. Jpn J Clin Oncol *37*, 907-912.

Young, R.P., Hopkins, R.J., Whittington, C.F., Hay, B.A., Epton, M.J., and Gamble, G.D. (2011). Individual and cumulative effects of GWAS susceptibility loci in lung cancer: associations after sub-phenotyping for COPD. PLoS One *6*, e16476.

Zhu, W.G., Dai, Z., Ding, H., Srinivasan, K., Hall, J., Duan, W., Villalona-Calero, M.A., Plass, C., and Otterson, G.A. (2001). Increased expression of unmethylated CDKN2D by 5-aza-2'-deoxycytidine in human lung cancer cells. Oncogene *20*, 7787-7796.

Zilberman, D., Coleman-Derr, D., Ballinger, T., and Henikoff, S. (2008). Histone H2A.Z and DNA methylation are mutually antagonistic chromatin marks. Nature *456*, 125-129.

Zochbauer-Muller, S., Fong, K.M., Virmani, A.K., Geradts, J., Gazdar, A.F., and Minna, J.D. (2001). Aberrant promoter methylation of multiple genes in non-small cell lung cancers. Cancer Res *61*, 249-255.



2

Genome-wide DNA methylation profiling of non-small cell lung carcinomas

Rejane Hughes Carvalho[1,2], Vanja Haberle[3], Jun Hou[1,4], Teus van Gent[1,4], Supat Thongjuea[3], Wilfred van IJcken[5], Christel Kockx[5], Rutger Brouwer[5], Erikjan Rijkers[5], Anieta Sieuwerts[6], John Foekens[6], Mirjam van Vroonhoven[5,7], Joachim Aerts[8], Frank Grosveld[1,2,4,9], Boris Lenhard[3] and Sjaak Philipsen[1,2,4]

[1] Department of Cell Biology, [2] Centre for Cancer Genomics,
[4] Dutch Consortium for System Biology, [5] Center for Biomics,
[6] Internal Oncology, [7] Department of Bioinformatics, [8] Department of Pulmonology,
[9] Center for Biomedical Genetics, ErasmusMC, Rotterdam, The Netherlands
[3] Uni BCCS, University of Bergen, Norway

(Submitted)

Abstract

Non-small cell lung carcinoma (NSCLC) is a complex malignancy that due to its heterogeneity and poor prognosis poses many challenges to diagnosis, prognosis and patient treatment. DNA methylation is an important mechanism of epigenetic regulation involved in normal development and cancer. It is a very stable and specific modification and therefore in principle a very suitable marker for epigenetic phenotyping of tumors. Here we present genome-wide DNA methylation analysis of NSCLC samples and paired healthy lung tissues, where we combine MethylCap and next generation sequencing (MethylCap-seq) to provide comprehensive DNA methylation maps of these tumors and paired healthy lung tissues. The MethylCap-seq data were validated by bisulfite sequencing and methyl-specific PCR of selected regions. Analysis of the MethylCap-seq data revealed strong positive correlation between replicate experiments and between paired tumor/adjacent lung samples. We identified 57 differentially methylated regions (DMRs) present in all NSCLC tumors analyzed by MethylCap-seq. While hypomethylated DMRs did not correlate to any particular functional category of genes, the hypermethylated DMRs were strongly associated with genes encoding transcriptional regulators. Furthermore, subtelomeric regions and satellite repeats were hypomethylated in the NSCLC samples. We also identified DMRs that were specific to two of the major subtypes of NSCLC, adenocarcinomas and squamous cell carcinomas. Collectively, we provide a resource containing genome-wide DNA methylation maps of NSCLC and normal lung tissues, and comprehensive lists of known and novel DMRs and associated genes in NSCLC.

Introduction

Non-small cell lung carcinoma (NSCLC) is a common malignancy characterized by a worldwide high incidence and low survival rate (Jemal et al., 2010). NSCLC is a heterogenic disease which is broadly classified in three major histopathological subtypes: adenocarcinoma (ADC), squamous cell carcinoma (SCC) and large cell carcinoma (LCC). This heterogeneity poses challenges for diagnosis and treatment, since each subtype presents with distinctive prognosis (Little, 2006) and the choice of therapeutic regimen is predominantly based on tumor subtype and staging parameters (Molina et al., 2008). Development of personalized diagnostics and therapy is leading the way to a new era that may overcome some of the difficulties in treating complex diseases such as NSCLC.

In the past decade, comparative gene expression profiles of tumors have been extensively studied (Hou et al., 2010; Larsen et al., 2007; Liang & Pardee, 2003), yielding useful insights in the molecular hallmarks of carcinogenesis (Kameyama et al., 2003; Petty et al., 2004). With the advent of next generation sequencing, genome-wide

screening has become an attractive tool for profiling tumors versus normal tissues cells (Ding et al., 2010; Kameyama et al., 2003). DNA methylation is a very stable epigenetic mark and next generation sequencing studies have recently shown that many genes are aberrantly methylated in cancer (Lister et al., 2009; Weber et al., 2005). Tissue specific DNA methylation patterns are stabilised during embryonic development, and faithfully maintained through cell divisions (Bird, 2002; Bird & Wolffe, 1999; Leonhardt et al., 1992). Nevertheless, established methylation patterns can be reprogrammed, with tumors cells undergoing DNA demethylation and *de novo* methylation through mechanisms not yet completely understood. *CDKN2A* and *RASSF1* are examples of genes found to be aberrantly methylated in a wide variety of tumors (Agathanggelou et al., 2001; Kohonen-Corish et al., 2007; Liu et al., 2005a; Pfeifer et al., 2002), and epigenetic silencing of these genes has also been reported in NSCLC (Dammann et al., 2000; Feng et al., 2008; Hsu et al., 2007; Vaissiere et al., 2009). The great majority of DNA methylation studies are concentrated on the analysis of CpG islands located in the promoter areas of pre-selected genes. However, differentially methylated areas may be located within genes and at large distances from the nearest neighbouring genes (Clark, 2007; Mayor et al., 2009). Although data on methylated genes in NSCLC are rapidly accumulating, unbiased data concerning specificity of the genome-wide distribution of methylated loci are still scarce.

In this study we used MethylCap and high-throughput sequencing (MethylCap-seq; (Brinkman et al., 2010)) to perform a genome-wide DNA methylation screening of NSCLC tumors and paired normal lung tissues. With this approach, we sought to identify genome-wide aberrant methylation patterns of NSCLC. Specific differentially methylated regions would be promising candidate molecular markers for non-invasive diagnostics using circulating tumor DNA, and increase the number of possible targets for epigenetic therapy.

Results

Methylation profiles in NSCLC: Study outline

We performed genome-wide DNA methylation analysis of NSCLC using the MethylCap assay followed by high-throughput sequencing: MethylCap-seq (Figure 1A). We used DNA isolated from 7 NSCLC tumors and paired normal lung tissues (Table 1 and Supplementary Table 1). As controls, we prepared fully methylated and fully unmethylated genomic DNA. The DNA samples were sheared and then enriched for methylated DNA using the MethylCap procedure (Brinkman et al., 2010). This is based on capture of methylated DNA by biotinylated methyl-binding domain protein (MBD), which is then retrieved by binding to streptavidin-coated beads (Figure 1B). The recovered DNA was directly sequenced using the Illumina GAIIx next generation sequencing platform. For each DNA sample, we performed two independent enrichment procedures and sequence runs. Total numbers of sequence

reads, mapped reads and unique reads for each sample are represented in supplementary table 2. For identification of differentially methylated regions (DMRs) we employed a rigorous normalization procedure and a variety of bioinformatics tools (Figure 1C). We used MethylCap-seq data obtained from artificially prepared fully unmethylated and fully methylated DNA samples for normalization of the data and assignment of DMRs. The bioinformatics approach is described in detail in Supplementary Methods and Supplementary Figure 1. Relative methylation scores were used to build an individual methylation profile for each sample. Subsequently, we compared the profiles to identify highly significant DMRs between tumors and paired healthy tissues, and between the subtypes of tumors. Finally, we validated the MethylCap-seq results by bisulfite sequencing and methyl-specific PCR of selected DMRs.

Table 1: Patient data used for MethylCap-seq

Sample ID	Tissue type	Case ID	Stage	Gender	Age at Diagnosis
2213N	healthy	SCC1	IIB	M	54,54
2214T	SCC				
2235N	healthy	SCC2	IIB	M	73,73
2236T	SCC				
2245N	healthy	ADC1	IB	F	66,26
2246T	ADC				
2255N	healthy	ADC2	IA	F	54,26
2256T	ADC				
2257N	healthy	ADC3	N/A	M	78,96
2258T	ADC				
2261N	healthy	SCC3	IIB	M	70,03
2262T	SCC				
22he	healthy	LCC1	IB	F	56,58
22tu	LCC				

SCC: Squamous cell carcinoma; ADC: Adenocarcinoma; LCC: Large cell carcinoma; N/A: data not available.

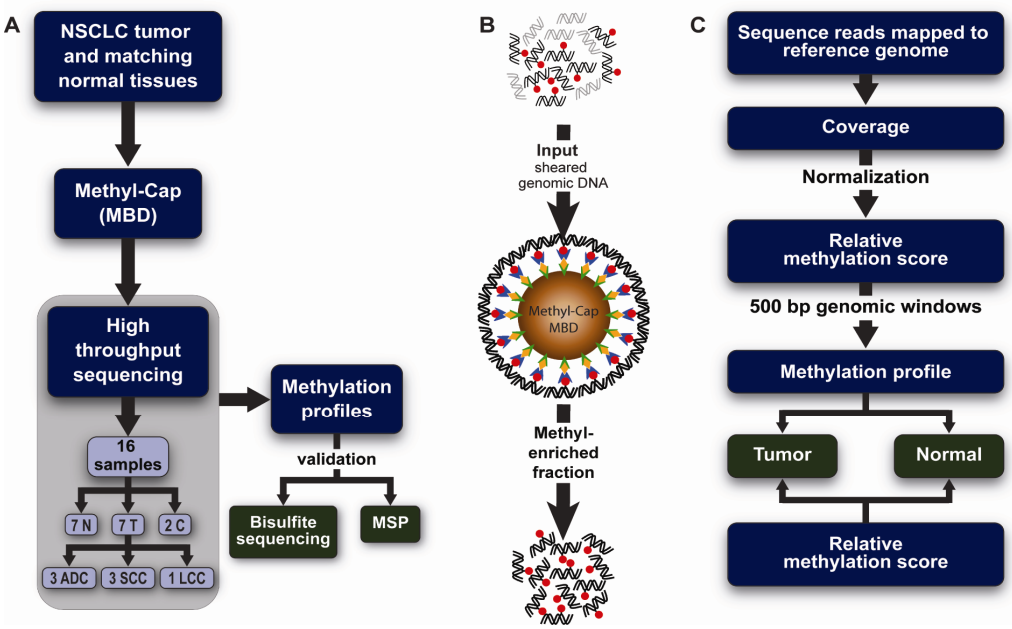


Figure 1: Experimental design for profiling of DNA methylation patterns in NSCLC.
(A) Overall view of the steps followed to generate the profiles (MBD: methyl-binding domain protein; T: tumor; N: Normal; ADC: adenocarcinoma; SCC: squamous cell carcinoma; LCC: large cell carcinoma); (B) MethylCap using methyl-binding domain proteins: sheared genomic DNA is used as input fraction (methyl groups in red) and incubated with beads (brown) coated with steptavidin (green) biotin (yellow)-MBD protein (blue) to capture methylated DNA. Captured fragments are subjected to high-throughput sequencing. (C) Summary of the bioinformatics approach used to generate the methylation profiles.

Global analysis of genome-wide methylation patterns of NSCLC

To asses the reproducibility of the MethylCap-seq procedure, we first performed a self/self comparison of the replicate experiments. This yielded an average Pearson's correlation coefficient of 0.89 (Figure 2A and Supplementary Figure 2), indicating excellent reproducibility between independent experiments. Next, we examined hyper- and hypomethylated areas by comparing the relative methylation scores of the tumors versus paired healthy lung tissues. This revealed an average correlation coefficient of 0.83, reflecting a generally high similarity in the methylation patterns of the matched tumor and healthy lung samples (Figure 2B and Supplementary Figure 3). Collectively, these analyses establish the high reproducibility and specificity of the MethylCap-seq procedure that we used.

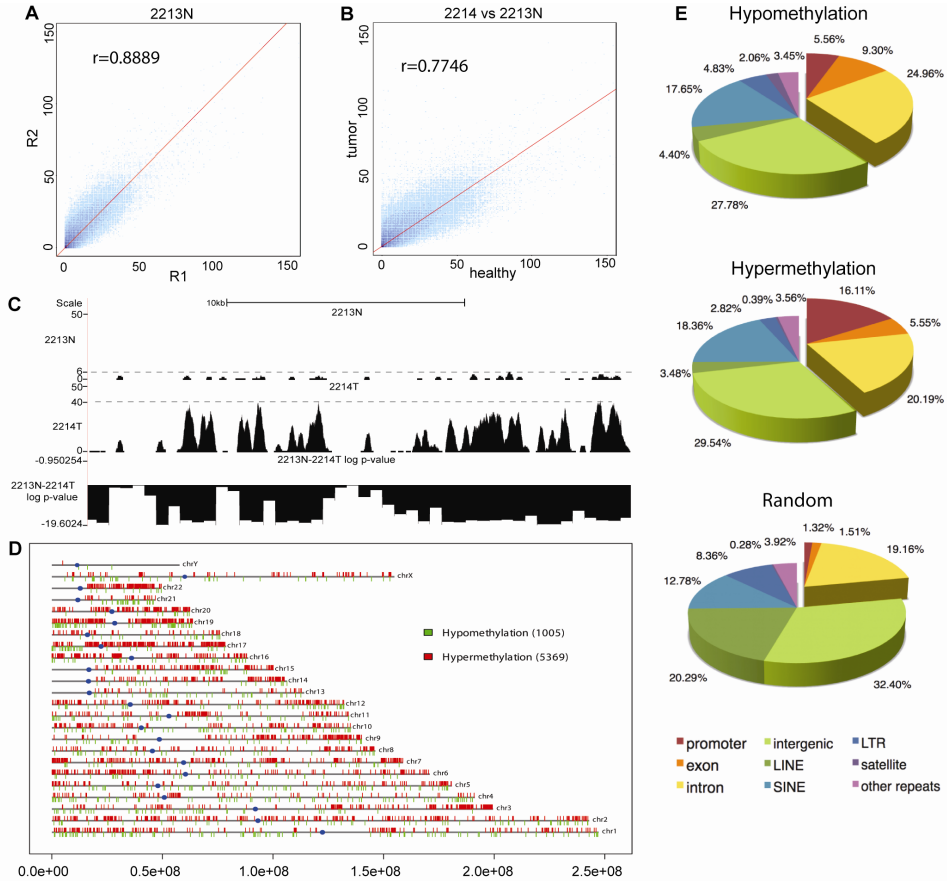


Figure 2: Global analysis of DNA methylation patterns in NSCLC. (A) Correlation between experimental replicates. Each point represents one 10bp genomic bin. Density of points (log10 scale) is shown in different shades of blue. Pearson's correlation coefficient is denoted in the top left corner of each scatter plot. (B) Correlation between tumor (x axis) and paired healthy lung tissue sample (y axis). (C) Raw reads from a normal/healthy sample (above panel) and its matching tumor sample (middle panel) in the position chr2:176,716,200-176,738,910 viewed in the UCSC genome browser; normalized p-value is depicted in the bottom panel. (D) Representation of the distribution of hypomethylated (green) en hypermethylated (red) regions across chromosomes in a tumor versus normal sample. (E) Composition of hypo- and hypermethylated regions in comparison to general distribution of MethylCap-seq reads obtained from the fully unmethylated DNA sample. Proportions of regions covered by distinct genomic features are shown. Promoters were defined as regions +/- 1 kb from all Ensembl TSSs. Distinct classes of repeats were retrieved from the UCSC Table Browser (RepeatMasker table). Repeats were excluded from all subsequent features. Genomic background consisted of 5000 regions of 500bp sampled randomly across the genome.

To allow visual inspection of the DMRs, raw reads and normalized log p-values were uploaded in the UCSC Genome browser, where DMRs can be observed (Figure 2C). To focus on the most significant DMRs, we used a very stringent p-value of 10^{-18} (Supplementary Figure 4). Non-repetitive regions with p-value $\leq 10^{-18}$ were projected onto the genome and visualized as genome-wide maps of differential methylation

(Figure 2D and Supplementary Figure 5). This revealed that in the non-repetitive fraction of the genome, tumors displayed higher overall hypermethylation then their paired healthy tissues. This is consistent with the notion that *de novo* methylation and hypermethylation of promoter CpG islands are associated with carcinogenesis (Keshet et al., 2006)

The DMRs were analyzed according to general hallmarks of their genomic localization (Figure 2E). Compared to the general distribution of MethylCap-seq reads obtained from the fully unmethylated DNA sample (pie chart labelled "random"), we found that promoter areas were particularly enriched among the DMRs (pie charts labelled "hypo" and "hyper" methylated). In addition, some specific repeat classes such as LINE and satellites were also differentially represented in the DMRs, with satellite repeats being particularly hypomethylated in tumors when compared to healthy lung tissues. In contrast, LINE repeats are relatively underrepresented in the DMRs (hypo and hypermethylation pie charts), indicating that the methylation status of the bulk of the LINE elements was similar between the tumors and paired healthy tissues.

Differentially Methylated Regions (DMRs) in NSCLC

We observed a total of 14742 DMRs in the 7 NSCLC tumor samples (Supplementary Table 3), i.e. regions found to be differentially methylated in at least one tumor when compared to its paired healthy tissue with a two-fold change difference in methylation signal. Performing unsupervised cluster analysis with these DMRs revealed that the samples clustered according to the histological subtype of the tumors (Figure 3A). In particular, hypermethylated DMRs were overrepresented in SCC samples. The dendogram of the DMRs indicated the presence of 9 major subgroups of DMRs. Two of these were composed of generally hypermethylated (red) or hypomethylated (green) DMRs. These were plotted on the chromosomes, revealing that the "red" DMRs were scattered throughout the chromosomes. In contrast, the "green" DMRs displayed a more localized distribution towards the chromosome ends (Figure 3A). Subtelomeric hypomethylation has been found to be associated with risk of developing cancer (Rauch & Pfeifer, 2010).

We found 57 DMRs present in all tumor samples, of which 51 were hypermethylated and 7 were hypomethylated. We also analyzed DMRs in relation to NSCLC subtypes. We found 287 DMRs unique for SCC and 26 DMRs unique for ADC (Supplementary File 1). These DMRs were classified by hyper- and hypomethylation and their genomic locations relative to genes (Figure 3B). We found that the ADC-specific DMRs had a genomic distribution similar to that observed for the DMRs present in all 7 tumors, while for SCCs DMRs within genes were concentrated in the gene body. SCC also presented considerable more subgroup-specific DMRs than ADC and "all-tumors"; this could be explained by the notion that SCC is a relatively homogeneous tumor type.

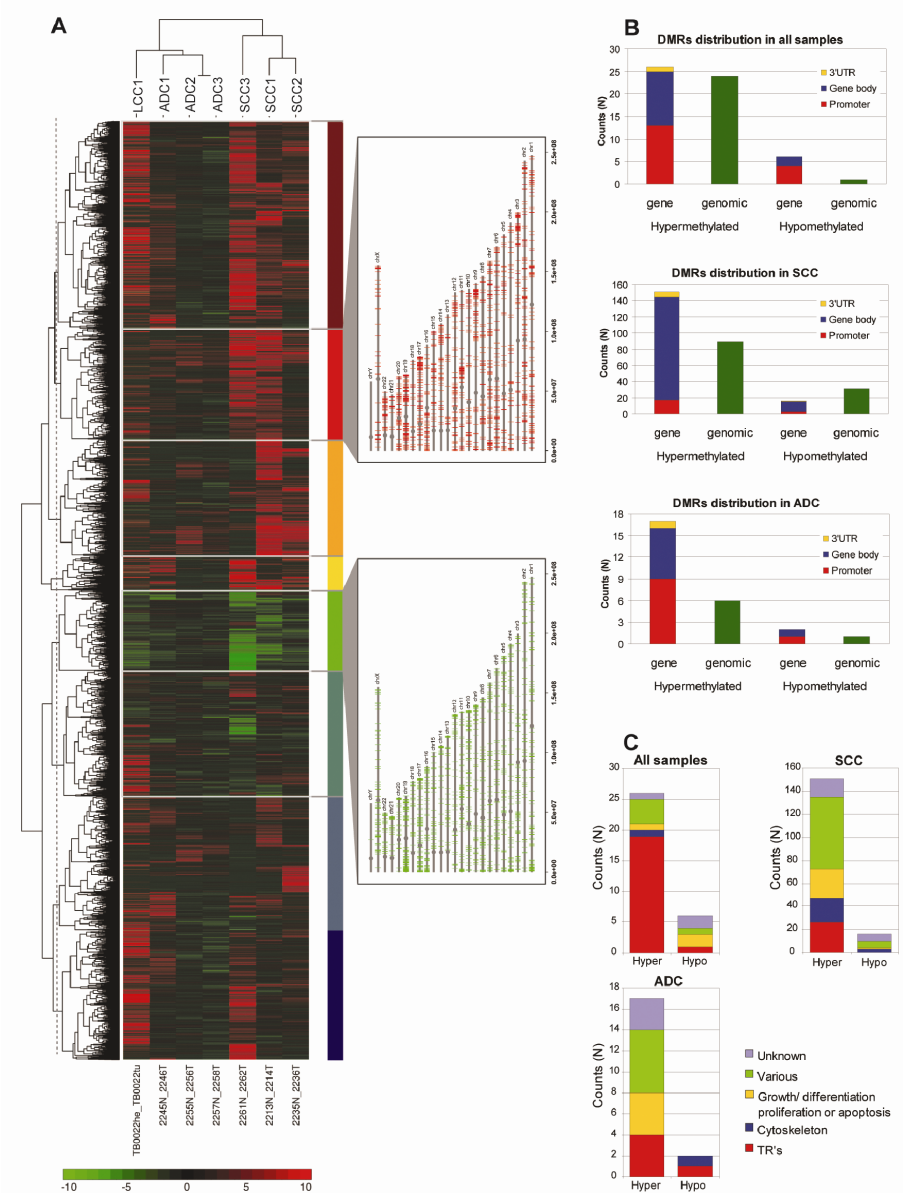


Figure 3: DMRs in NSCLC versus paired healthy lung tissue samples. (A): Heatmap of 14742 most significant DMRs, picturing regions with mostly hypermethylated (top), mostly hypomethylated (middle), and mixture (bottom). Color bar at the bottom represents log ratio of normalized signal in tumor vs. healthy (red - hypermethylation, green - hypomethylation). Dashed lines represent cuts in the dendogram, generating 9 groups of DMRs. The DMRs from the two most distinctive groups are depicted on the chromosomes. Red: cluster containing regions hypermethylated in all samples; Green: cluster containing regions hypomethylated in all samples. (B): DMRs distribution relative to gene position. Number of hypermethylated and hypomethylated genes and regions (outside gene area), and their distribution as in promoter, gene body and 3' UTR; (C): Gene function distribution of the differentially methylated genes. Number of hypermethylated and hypomethylated genes relative to gene classes or function. ADC: Adenocarcinoma; SCC: squamous cell carcinoma. TR: transcription regulator.

For the DMRs that fall in gene areas, we analyzed which types of genes were present in each of the 3 groups (Figure 3C and Supplementary File 1). We found that for the DMRs present in all tumor samples, more than 75% of the associated genes belonged to the class of transcriptional regulators. For SCC-specific DMRs, we observed a considerable heterogeneity in the functional categories of associated genes. Nevertheless, there was an overrepresentation of genes involved in transcriptional regulation, organization of the cytoskeleton and cell cycle regulation. We note that while some of the DMRs were associated with genes that had been previously reported to be differentially methylated in lung tumors, such as *APC*, *CDH13*, *CDKN2A*, *DAPK*, *hMLH1*, *HOX genes, OTX1*, and *HOX2,* and many others (Dammann et al., 2000; Rauch et al., 2007; Rauch et al., 2008; Schmidt et al., 2010; Zochbauer-Muller et al., 2001), we also found DMRs which, to the best of our knowledge, have not been previously reported to be associated with methylation in NSCLC or other types of cancer.

Validation of differentially methylated regions in NSCLC by bisulfite sequencing

To validate the results obtained with MethylCap-seq, we selected five fragments for analysis by bisulfite sequencing. We chose *CDKN2A* and *RASSF1* since hypermethylation of these genes has been widely reported in a variety of cancers including NSCLC (Agathanggelou et al., 2001; Dammann et al., 2000; Vaissiere et al., 2009), and the *EN1* promoter region since hypermethylation of *EN1* had been previously reported in a lung tumor cell line (Rauch et al., 2007) but not yet in primary NSCLC tumors. The other two fragments were located at positions chr2:119,331,343-119,331,692 and chr2:119,328,097-119,328,400 and we name then Frag_01 and Frag_02 respectively. Selection of the fragments for bisulfite sequencing analysis was based on the MethylCap-seq data (Figure 4).

To assure sufficient depth of coverage for quantitative analysis, we sequenced 36 individual clones from each fragment and each bisulfite-converted DNA sample. Figure 4B shows an example of the methylation status for one fragment of a paired sample. The methylation status of the 5 fragments in all 7 paired NSCLC/healthy lung tissue samples is summarized in Figure 4C; full data for all samples is available in Supplementary File 2 and Supplementary File 3. We also examined potential differences in methylation according to tumor histology. To perform this analysis we grouped the samples in four categories: all clones from the normal tissues of patients (ADC_N and SCC_N) and clones from tumors (ADC_T and SCC_T), (Figure 4D and Supplementary File 4). We observed a statistically significant difference (p-value =<0.0001) not only between tumors and healthy tissues, but also between ADC and SCC for all fragments analyzed by bisulfite sequencing. This identifies that these fragments as candidate markers for differentiating tumors *versus* healthy tissues, and also between tumor subtypes. Moreover this finding validates the MethylCap-seq data,

since the methylation status determined by bisulfite sequencing correlated quantitatively with the MethylCap-seq data (Figure 4E).

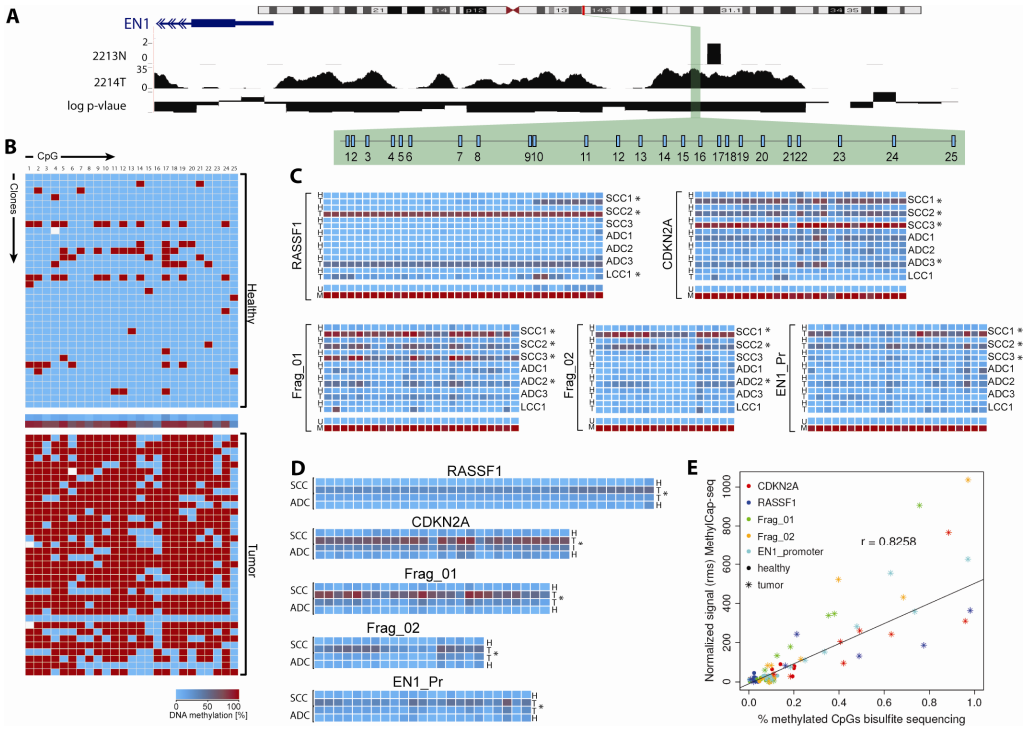


Figure 4: Bisulfite sequencing validation for MethylCap-Seq . (A) Sequence reads for the Frag_01 in a tumor (2214T) and matching normal tissue (2213N), plotted in the Genome Browser, showing the distribution of the 25 CpGs contained in the fragment highlighted in green. Position for this fragment is chr2:119,331,343-119,331,692. (B) Methylation status of each CpG in all 36 individually sequenced clones in the same samples and fragment shown in Fig. 4A. The middle row represents the average of methylation in all clones per CpG position. (C) Average of methylation of all clones sequenced per patient in each fragment (N=normal; T=tumor; U = control totally unmethylated DNA; M = control totally methylated DNA). (D) Average of the methylation status of the sum of all clones, grouped per histological subtype; comparison between ADC and SCC in all fragments were statistically significant. (E) Correlation between normalized methylation signal from MethylCap-seq and CpG methylation from bisulfite sequence. Different regions are shown in different colors; healthy samples are marked by dots and tumors by stars. Perason's correlation coefficient is denoted above de linear regression curve. *p-value =< 0.001.

Screening by methylation-specific PCR (MSP) verifies methylation differences between Normal and Tumor, and between tumor subtypes

To screen the methylation status of selected DMRs in a larger sample set we used methylation-specific PCR (MSP). We selected CDKN2A, Frag_01 and Frag_02 for this analysis, since they were most significant differentially methylated fragments in

the bisulphite sequencing experiments. We screened 96 samples (48 NSCLC and paired healthy lung samples), and used totally methylated and totally unmethylated DNA samples as controls for the specificity of the MSP assays. Results for the 7 paired samples used in MethylCap-seq are displayed in Figure 5. Sensitivity and specificity for the 3 fragments in all 96 samples are summarized in table 2 and shown in Supplementary Figure 6.

For each "M" primer set, amplification of the fragment is dependent on methylation; while each "U" primer set requires unmethylated DNA for amplification. In the 48 paired samples analyzed, *CDKN2A* showed frequent amplification with the M primer set in healthy lung samples (24/48), although the amplified bands were often of low intensity. Tumors had a much larger number displaying amplification with both M and U primer sets. 36 out of 47 tumors displayed amplification with the M primer, and two of the tumors displayed amplification only with the M primer set.

In contrast, Frag_01 presented no amplification with the M primer set in the healthy lung samples, while the Frag_02 M primer set only 9 out of 48 were amplified. In tumors, we observed fragment amplification in 29/48 samples for Frag_01 and 36/48 samples for Frag_02 with the M primer sets. Differences in methylation status showed significant in tumors versus healthy lung samples for all 3 fragments, with a p-value < 0.02 for CDKN2A and $p<0.0001$ for Frag_01 and Frag_02; Figure 5B). Frag_01 and Frag_02 present a much lower p-value, indicating they may be better tumor markers than CDKN2A. We also observed that significant differences between SCC and ADC were only found for in Frag_01 (Figure 5C). In conclusion, the MSP results validate the MethylCap-seq data, since all three fragments are consistent with the significant DMRs in all tumors, and Frag_01 appeared as part of the SCC-specific DMRs in the MethylCap-seq data.

Table 2: Frequency, sensitivity and sensibility of the 3 fragments in 48 patients.

	CDKN2A				**Frag_01**				**Frag_02**			
	n	(%)	Sens (%)	Spec (%)	n	(%)	Sens (%)	Spec (%)	n	(%)	Sens (%)	Spec (%)
Lung	24/48	50.0	75.0	27.1*	0/47	0.0	62	100.0§	9/48	18.8	75.0	56.3§
Tumor	36/48	75.0			29/47	59.6			36/48	75.0		
ADC	13/17	52.9	76.5	35.3	5/16	31.3	31.25	31.3	12/17	70.6	70.6	64.7
SCC	15/16	88.2	93.8	37.5	13/16	81.3	81.25	81.3*	14/16	87.5	87.5	75.0

Sensitivity is calculated using the ration between methylated (amplified) sample and the total amount of samples analyzed per fragment; specificity is ratio of the difference between amplification in tumor and in the matching healthy sample. Sens = sensitivity. Spec = specificity. Statistic significance was calculated using X^2. * = $p<0.02$; § = $p<0.0001$

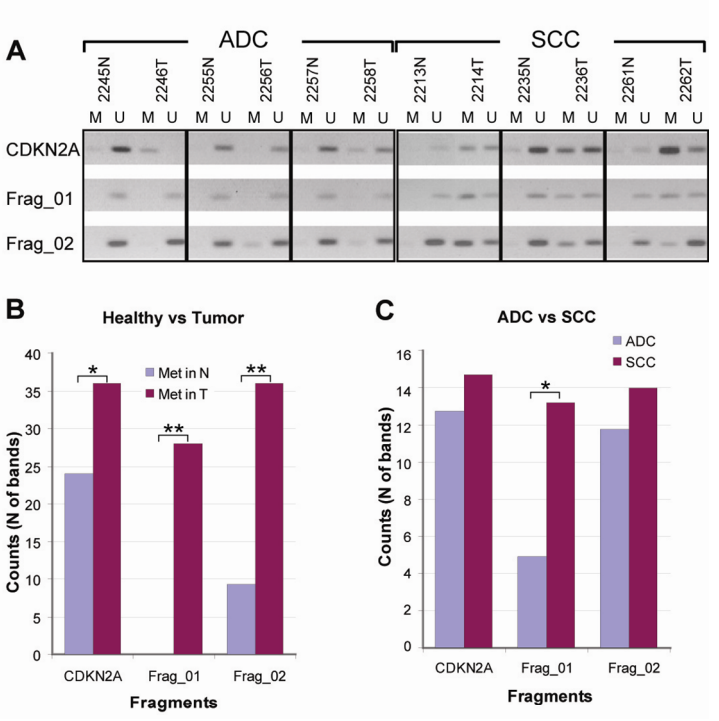


Figure 5: Methylation-specific PCR screening in 48 tumors and matching healthy tissues. (A) Cropped gel images for the 7 patients used for MethylCap-Seq, grouped by histological subtype (U: unmethylated; M: methylated; N: normal/healthy; T: tumor). (B) Histogram of the methylation status in normal/healthy samples and tumor samples in all 48 patients in the three fragments analyzed. (C) Histogram of the methylation status per histological subtype (* = p-value =< 0.02; **= p-value <0.001).

Discussion

In this study we performed an unbiased genome-wide DNA methylation analysis of 7 NSCLC cases. To date, genome-wide methylation studies in NSCLC have concentrated on specific areas of the genome (Rauch et al., 2008), promoter regions (Bibikova et al., 2006; Sano et al., 2007; Shames et al., 2006) or sets of pre-selected candidate genes (Feng et al., 2008; Zhang et al., 2011). Here we combined MethylCap with high-throughput sequencing (MethylCap-seq) to draw detailed methylome maps of NSCLC tumors and paired healthy lung tissues. The replicate experiments showed that the depth of coverage generated by MethylCap-seq was sufficient to capture the methylation status of the entire genome. Moreover, there was a very high degree of

correlation between the replicate experiments, demonstrating the robustness of the MethylCap-seq approach. As controls for normalization and validation purposes we made use of fully methylated and fully unmethylated DNA samples. The MethylCap-seq data of these controls provide a useful resource for benchmarking of the fully methylated *versus* fully unmethylated status of any particular fragment in the genome. Using the MEDIPS computational tool (Chavez et al., 2010), we generated methylation profiles of the seven paired NSCLC/healthy lung cases which can be viewed in the genome browsers. This broad view of DNA methylation in NSCLC will provide new opportunities for identification of specific epigenetic markers of NSCLC that could, for instance, be used for early detection of the disease.

The occurrence of Methylated CpGs in the fragments selected for bisulfite sequencing showed a high quantitative correlation with the reads found by MethylCap-seq in the same area. In conjunction with the MethylCap-seq data on the fully methylated and fully unmethylated controls, this demonstrates that MethylCap-seq yields reliable quantitative information on the methylation status of the fragments in a particular DNA sample. This is important considering that the tumor cell content of samples to be analyzed will be variable. Provided that sufficient sequencing depth is achieved, it should be possible to use MethylCap-seq for identification of tumor-specific hypermethylated DNA regions in samples containing only a minority of tumor cells, and possibly in circulating tumor DNA isolated from serum samples. The results obtained with MSP are qualitative in nature, but this PCR-based approach is very sensitive and may therefore be used to detect tumor-specific hypermethylated DNA fragments in serum DNA. As a first step in this direction, we used MSP to investigate the presence of three DMRs in a larger set of paired samples. The results confirmed the MethylCap-seq data and demonstrated that these DMRs are frequently hypermethylated in NSCLC. In particular, the Frag_01 fragment was found to be hypermethylated in 13 out of 16 SCC cases and in 5 of the ADC, suggesting that this might be a useful marker for this type of NSCLC. In addition, for *CDKN2A* amplification with the "M" primer was often observed in healthy tissue samples, although at much lower amplification levels when compared with tumor samples. This could reflect an early event in the methylation status of *CDKN2A* in tumorigenesis.

The mapped reads point to differentially methylated areas across the genome, with some marked preferences for hypermethylation or hypomethylation in certain genomic regions. There is a strong correlation between CpG islands and hypermethylation in tumors, with the most significantly hypermethylated regions associated with promoter regions of genes. As previously reported by many groups (Chung et al., 2011; Esteller, 2007; Liu et al., 2005b; Marsit et al., 2006), gene promoters are a target of methylation as an epigenetic regulatory mechanism. However, our data also reveal a high degree of hypermethylation outside promoter areas at intergenic regions, and at long distances from genes. This indicates that patterns of DNA methylation may play a role not only by silencing promoters of key tumor

suppressor genes, but also by regulating gene expression in a more complex manner through distant regulatory elements such as insulators and enhancers.

We found that satellite regions were particularly hypomethylated in the NSCLC tumors when compared to normal lung tissues. Recently, (Ting et al., 2011) have shown that over expression of satellite repeats is associated with different types of cancer. The observed hypomethylation of these regions in NSCLC would be compatible with their results. We also found that hypomethylated regions are often present in subtelomeric regions, in agreement with the observation that hypomethylation is more frequently located near chromosome ends in lung cancer (Pfeifer & Rauch, 2009). While CpG islands and genes are present in the subtelomeric regions of chromosomes, the impact of subtelomeric hypomethylation on the regulation of these genes is currently unclear.

In an alternative approach, we selected the most significant DMRs in the individual NSCLC cases (14742) and performed unsupervised cluster analysis of the hyper- and hypomethylated regions in the tumor samples. We observed that the samples clustered according to their histological classification. This indicates that there are DMRs that could not only be used as NSCLC markers but also as markers for histological classification of NSCLC tumors. Subsequently we searched for the specific areas that were differentially methylated in all tumors when compared to the paired healthy tissues, and specific for the subtypes. We observed that SCCs displayed more DMRs than ADCs. This observation could be explained by the heterogeneity of ADCs, which would result in fewer DMRs shared by all three ADC samples. The contrary is seen in SCC, which is a more homogeneous type of NSCLC. This characteristic has been observed previously in gene expression profiling studies (Hou et al., 2010; Hou et al., 2011).

Remarkably, the majority (>75%) of the hypermethylated genes present in all 7 samples encoded for transcription factors, while the shared hypomethylated genes were spread among many different functional categories. Transcription factors play a central role in maintaining or modifying cell fate, either in normal processes such as development and cell differentiation, or in cancer. Mechanisms of epigenetic regulation via DNA methylation are not yet completely understood; several factors such as DNA methyltransferases, chromatin remodelling proteins and DNA-binding transcription factors are involved (Chatagnon et al., 2010; Clouaire et al., 2010). Therefore, any perturbation that leads to a decrease in the expression status of those factors may disrupt important pathways for maintenance of the differentiated cellular state. Aberrant methylation can change chromatin structure, rendering DNA binding sites accessible or inaccessible to transcription factors leading to activation or silencing of genes important in cell differentiation and maintenance. It has been recently shown that hypermethylated genes found in lung tumors are associated to cellular morphogenetic differentiation process (Helman et al., 2011). These cellular mechanisms are

orchestrated by transcriptional regulators; therefore changes in the methylation status of their binding sites may disrupt those processes thus contributing to oncogenesis.

Based on the MethylCap-seq data, we generated a list of candidate DMRs and associated genes specific for NSCLC and its subtypes (Table 3). Several of these DMR-associated genes are known to be hypermethylated in NSCLC and other types of cancer, strongly supporting the validity of the data set reported here. For instance, the observed methylation rates of *RASSF1* were in agreement with previously reported results (Pfeifer et al., 2002). In the MetylCap-seq data, hypermethylation of the genomic region where Frag_01 is located was most significant for the SCC samples. This was supported by the bisulfite sequencing and MSP validation experiments, where this fragment showed significantly more hypermethylation in SCCs when compared to ADCs or LCCs. In conclusion, this list of candidate DMR markers can be used to develop sensitive biological markers for NSCLC, which may enable non-invasive diagnosis and early detection of the disease, and potentially allow histological classification. Collectively, we provide a resource containing genome-wide DNA methylation maps of NSCLC and normal lung tissues, and comprehensive lists of DMRs and associated genes in NSCLC.

Methods

Patient samples

Samples were obtained from NSCLC patients (n = 48) who had undergone surgical lung resection at Erasmus MC. Specimens were collected from the tumor and adjacent non-cancerous lung tissue and studied under an anonymous tissue protocol approved by the medical ethical committee of Erasmus University Medical Center. Tissues were snap-frozen within two hours after surgical resection in liquid nitrogen pre-cooled isopentane, and stored at -196 °C or -80°C until DNA extraction.

Histopathological analysis

Patient samples were independently reviewed by two pathologists. The cohort included 17 adenocarcinomas (ADC); 16 squamous cell carcinomas (SCC), 7 large cell carcinomas (LCC), 8 unclassified samples, and paired healthy lung tissues for each tumor sample.

Cell line

The MRC-5 lung fibroblast-like cell line was used as control. Cells were cultured under standard conditions using minimum essential medium supplemented with 10% heat-inactivated fetal bovine serum, 2mM L-glutamine, 1% non-essential amino acids and penicillin/streptomycin. Cells were harvested when they reached 90% confluence.

DNA isolation
Genomic DNA from patient tissues and cultured cells were extracted by overnight treatment incubation with lysis buffer and proteinase K, followed by phenol-chloroform extraction, ethanol precipitation and RNase digestion.

Artificial demethylation and methylation of genomic DNA
We used DNA extracted from the MRC-5 cell line and commercially available Universal unmethylated DNA (UUD; Millipore). Fully unmethylated DNA was obtained by whole-genome amplification using the REPLI-g kit (Qiagen) according to manufacturer protocol, followed by phenol-chloroform extraction. Fully methylated DNA was prepared by treating MRC-5 DNA and UUD with M.SssI enzyme (New England Biolabs) according to the manufacturer's protocol. In short, 10 µg of DNA was incubated for 2 hours at 37 ˚C with 40U of M.SssI and 640 µM of S-adenosylmethionine. DNA was then treated with phenol-chloroform and recovered by ethanol precipitation.

Methyl-DNA capture (MethylCap)
Enrichment of methylated DNA was carried out using the MethylCap technique. DNA samples were sheared by sonication to obtain fragments between 200 and 800 bp. Methylated DNA capture was carried out using an adapted protocol from the MethylMiner Methylated DNA Enrichment kit (Invitrogen). Two adaptations were made. Firstly, 20 µl of beads and 14 µl (7µg) of MBD-biotin protein were used for 5 µg of sheared genomic DNA. Secondly, a single elution fraction was obtained by resuspending the beads in 200 µl of 1X binding/washing buffer containing 2 µl of proteinase K (20mg/ml). The samples were then incubated for 90 minutes at 57 ˚C with 800 rpm agitation. Remaining steps were performed following the manufacturer's protocol scaled up to 5 µg of DNA. Before the samples were used for high-throughput sequencing, methyl-enrichment was tested by qPCR (ICR; data not shown). qPCR was carried out using sybr green, Phire hot start DNA polymerase (Finnzymes), 1 µl of the enriched samples. Primer sequences are available in supplementary table 4. qPCR was performed in triplicate and enrichment levels were calculated as previously described (Karlen et al., 2007).

High-throughput sequencing
HTS was carried out on 18 samples: 7 tumors (3 ADC, 3 SCC and 1 LCC), 7 paired normal tissues and 4 controls (2 artificially fully methylated DNA and 2 fully unmethylated DNA). For each sample, DNA recovered from two independent Methyl Capture experiments was sequenced. The Illumina Genome Analyzer IIx platform was used for both replicates, one sample per lane, according to the manufacturer's protocol. In short: after MethylCap, fragments of the methyl-enriched fraction were end-repaired and ligated to single-read adaptors. Samples were size-selected to an average of 320 bp, PCR-amplified and 36bp sequenced.

The Illumina Casava pipeline was used for base calling, alignment and quality control. Results were mapped against the reference Human UCSChg18 AllChromosomes using eland_extended by Illumina pipeline 1.6.0.

Bioinformatic anlaysis

Detailed information about evaluation and selection of normalization parameters, and identification, characterization and selection of DMRs can be found in supplementary methods.

Bisulfite Sequencing

Bisulfite sequencing was carried out on the same 16 samples used for MethylCap HTS. Detailed information is provided in supplementary methods. In short: bisulfite conversion was carried out using the Epitect kit from Qiagen following the manufacturer's protocol. Samples were amplified using primers specific for the fragments of interest (Supplementary Table 4). Amplified fragments were loaded on 2% agarose gels and extracted from the gel using a NucleoSpin Extract II kit (Machery-Nagel). Fragments were then ligated to pGEM-t easy vector (Promega) and cloned into DH12-β competent cells. Colony PCR was performed on 36 colonies and PCR fragments sequenced. Analysis of methylated and unmethylated CpGs was executed using two online programs: BISMA (http://biochem.jacobs-university.de/BDPC/BISMA) and QUMA (http://quma.cdb.riken.jp).

Methylation-specific PCR (MSP)

To determining the methylation status of the selected regions, we used MSP (Herman et al., 1996) in 96 samples (48 tumors and matching healthy tissues). Samples were bisulfite converted as described for bisulfite sequencing, and amplified using two different set of primers designed for the methylated and unmethylated sequences (Supplementary Table 4). The fully methylated and fully unmethylated DNA samples were used as controls, and a water blank reaction used as control for contamination. After amplification, products were resolved on 2% agarose gels containing ethidium bromide and visualized under UV transillumination.

Statistical analysis

Statistical analysis for bisulfite sequencing was provided by the BISMA program. In short, Fisher's exact test was used for the statistical significance between two bisulfite sequence groups at each CpG, and Mann-Whitney U-test for the statistical significance between two groups of the entire set of CpG sites. Two-tailed p-value of Fisher's exact test was calculated from the 2 x 2 tables at each CpG site. Mann-Whitney U-test was used for the statistical significance of the entire set of CpG sites between the tumor and healthy groups. For determination of significance in the MSP, we used chi-square distribution. Two-tailed p-values were determined by the counts of the amplified bands

for each primer in the group of samples healthy/tumor or SCC/ADC. Groups that showed a p-value < 0.05 were considered significantly different.

References

Agathanggelou A, Honorio S, Macartney DP, Martinez A, Dallol A, Rader J, et al. (2001). Methylation associated inactivation of RASSF1A from region 3p21.3 in lung, breast and ovarian tumours. *Oncogene*, 20, 1509-18.

Bibikova M, Lin Z, Zhou L, Chudin E, Garcia EW, Wu B, et al. (2006). High-throughput DNA methylation profiling using universal bead arrays. *Genome Res*, 16, 383-93.

Bird A. (2002). DNA methylation patterns and epigenetic memory. *Genes Dev*, 16, 6-21.

Bird AP and Wolffe AP. (1999). Methylation-induced repression--belts, braces, and chromatin. *Cell*, 99, 451-4.

Brinkman AB, Simmer F, Ma K, Kaan A, Zhu J and Stunnenberg HG. (2010). Whole-genome DNA methylation profiling using MethylCap-seq. *Methods*, 52, 232-6.

Chatagnon A, Ballestar E, Esteller M and Dante R. (2010). A role for methyl-CpG binding domain protein 2 in the modulation of the estrogen response of pS2/TFF1 gene. *PLoS One*, 5, e9665.

Chavez L, Jozefczuk J, Grimm C, Dietrich J, Timmermann B, Lehrach H, et al. (2010). Computational analysis of genome-wide DNA methylation during the differentiation of human embryonic stem cells along the endodermal lineage. *Genome Res*, 20, 1441-50.

Chung JH, Lee HJ, Kim BH, Cho NY and Kang GH. (2011). DNA methylation profile during multistage progression of pulmonary adenocarcinomas. *Virchows Arch*.

Clark SJ. (2007). Action at a distance: epigenetic silencing of large chromosomal regions in carcinogenesis. *Hum Mol Genet*, 16 Spec No 1, R88-95.

Clouaire T, de Las Heras JI, Merusi C and Stancheva I. (2010). Recruitment of MBD1 to target genes requires sequence-specific interaction of the MBD domain with methylated DNA. *Nucleic Acids Res*, 38, 4620-34.

Dammann R, Li C, Yoon JH, Chin PL, Bates S and Pfeifer GP. (2000). Epigenetic inactivation of a RAS association domain family protein from the lung tumour suppressor locus 3p21.3. *Nat Genet*, 25, 315-9.

Ding L, Mardis ER and al. e. (2010). Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature*, 464, 999-1005.

Esteller M. (2007). Epigenetic gene silencing in cancer: the DNA hypermethylome. *Hum Mol Genet*, 16 Spec No 1, R50-9.

Feng Q, Hawes SE, Stern JE, Wiens L, Lu H, Dong ZM, et al. (2008). DNA methylation in tumor and matched normal tissues from non-small cell lung cancer patients. *Cancer Epidemiol Biomarkers Prev*, 17, 645-54.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5, R80.

Helman E, Naxerova K and Kohane IS. (2011). DNA hypermethylation in lung cancer is targeted at differentiation-associated genes. *Oncogene*.

Herman JG, Graff JR, Myohanen S, Nelkin BD and Baylin SB. (1996). Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. *Proc Natl Acad Sci U S A*, 93, 9821-6.

Hou J, Aerts J, den Hamer B, van Ijcken W, den Bakker M, Riegman P, et al. (2010). Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PLoS One*, 5, e10312.

Hou J, Lambers M, Hamer B, Bakker M, Hoogsteden MC, Grosveld F, et al. (2011). Expression profiling-based subtyping identifies novel NSCLC subgroups and implicates putative resistance to Pemetrexed therapy. *Journal of Thoracic Oncology*, in press.

Hsu HS, Chen TP, Hung CH, Wen CK, Lin RK, Lee HC, et al. (2007). Characterization of a multiple epigenetic marker panel for lung cancer detection and risk assessment in plasma. *Cancer*, 110, 2019-26.

Jemal A, Siegel R, Xu J and Ward E. (2010). Cancer statistics, 2010. *CA Cancer J Clin*, 60, 277-300.

Kameyama K, Huang CL, Liu D, Masuya D, Nakashima T, Sumitomo S, et al. (2003). Reduced ING1b gene expression plays an important role in carcinogenesis of non-small cell lung cancer patients. *Clin Cancer Res*, 9, 4926-34.

Karlen Y, McNair A, Perseguers S, Mazza C and Mermod N. (2007). Statistical significance of quantitative PCR. *BMC Bioinformatics*, 8, 131.

Keshet I, Schlesinger Y, Farkash S, Rand E, Hecht M, Segal E, et al. (2006). Evidence for an instructive mechanism of de novo methylation in cancer cells. *Nat Genet*, 38, 149-53.

Kohonen-Corish MR, Sigglekow ND, Susanto J, Chapuis PH, Bokey EL, Dent OF, et al. (2007). Promoter methylation of the mutated in colorectal cancer gene is a frequent early event in colorectal cancer. *Oncogene*, 26, 4435-41.

Larsen JE, Pavey SJ, Passmore LH, Bowman R, Clarke BE, Hayward NK, et al. (2007). Expression profiling defines a recurrence signature in lung squamous cell carcinoma. *Carcinogenesis*, 28, 760-6.

Leonhardt H, Page AW, Weier HU and Bestor TH. (1992). A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei. *Cell*, 71, 865-73.

Liang P and Pardee AB. (2003). Analysing differential gene expression in cancer. *Nat Rev Cancer*, 3, 869-76.

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462, 315-22.

Little AG. (2006). No nodes is good nodes. *Ann Thorac Surg*, 82, 4-5.

Liu L, Broaddus RR, Yao JC, Xie S, White JA, Wu TT, et al. (2005a). Epigenetic alterations in neuroendocrine tumors: methylation of RAS-association domain family 1, isoform A and p16 genes are associated with metastasis. *Mod Pathol*, 18, 1632-40.

Liu Y, Pang JC, Dong S, Mao B, Poon WS and Ng HK. (2005b). Aberrant CpG island hypermethylation profile is associated with atypical and anaplastic meningiomas. *Hum Pathol*, 36, 416-25.

Marsit CJ, Houseman EA, Christensen BC, Eddy K, Bueno R, Sugarbaker DJ, et al. (2006). Examination of a CpG island methylator phenotype and implications of methylation profiles in solid tumors. *Cancer Res*, 66, 10621-9.

Mayor R, Casadome L, Azuara D, Moreno V, Clark SJ, Capella G, et al. (2009). Long-range epigenetic silencing at 2q14.2 affects most human colorectal cancers and may have application as a non-invasive biomarker of disease. *Br J Cancer*, 100, 1534-9.

Molina JR, Yang P, Cassivi SD, Schild SE and Adjei AA. (2008). Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. *Mayo Clin Proc*, 83, 584-94.

Petty RD, Nicolson MC, Kerr KM, Collie-Duguid E and Murray GI. (2004). Gene expression profiling in non-small cell lung cancer: from molecular mechanisms to clinical application. *Clin Cancer Res*, 10, 3237-48.

Pfeifer GP and Rauch TA. (2009). DNA methylation patterns in lung carcinomas. *Semin Cancer Biol*, 19, 181-7.

Pfeifer GP, Yoon JH, Liu L, Tommasi S, Wilczynski SP and Dammann R. (2002). Methylation of the RASSF1A gene in human cancers. *Biol Chem*, 383, 907-14.

Rauch T, Wang Z, Zhang X, Zhong X, Wu X, Lau SK, et al. (2007). Homeobox gene methylation in lung cancer studied by genome-wide analysis with a microarray-based methylated CpG island recovery assay. *Proc Natl Acad Sci U S A*, 104, 5527-32.

Rauch TA and Pfeifer GP. (2010). DNA methylation profiling using the methylated-CpG island recovery assay (MIRA). *Methods*, 52, 213-7.

Rauch TA, Zhong X, Wu X, Wang M, Kernstine KH, Wang Z, et al. (2008). High-resolution mapping of DNA hypermethylation and hypomethylation in lung cancer. *Proc Natl Acad Sci U S A*, 105, 252-7.

Sano A, Kage H, Sugimoto K, Kitagawa H, Aki N, Goto A, et al. (2007). A second-generation profiling system for quantitative methylation analysis of multiple gene promoters: application to lung cancer. *Oncogene*, 26, 6518-25.

Schmidt B, Liebenberg V, Dietrich D, Schlegel T, Kneip C, Seegebarth A, et al. (2010). SHOX2 DNA methylation is a biomarker for the diagnosis of lung cancer based on bronchial aspirates. *BMC Cancer*, 10, 600.

Shames DS, Girard L, Gao B, Sato M, Lewis CM, Shivapurkar N, et al. (2006). A genome-wide screen for promoter methylation in lung cancer identifies novel methylation markers for multiple malignancies. *PLoS Med*, 3, e486.

Ting DT, Lipson D, Paul S, Brannigan BW, Akhavanfard S, Coffman EJ, et al. (2011). Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. *Science*, 331, 593-6.

Vaissiere T, Hung RJ, Zaridze D, Moukeria A, Cuenin C, Fasolo V, et al. (2009). Quantitative analysis of DNA methylation profiles in lung cancer identifies aberrant DNA methylation of specific genes and its association with gender and cancer risk factors. *Cancer Res*, 69, 243-52.

Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, et al. (2005). Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet*, 37, 853-62.

Zhang Y, Wang R, Song H, Huang G, Yi J, Zheng Y, et al. (2011). Methylation of multiple genes as a candidate biomarker in non-small cell lung cancer. *Cancer Letters*, 303, 21-28.

Zochbauer-Muller S, Fong KM, Virmani AK, Geradts J, Gazdar AF and Minna JD. (2001). Aberrant promoter methylation of multiple genes in non-small cell lung cancers. *Cancer Res*, 61, 249-55.

Supplementary Information

Supplementary files not present here can be found at: https://docs.google.com
Login: thesis.hughes; password: thesishughes_2012

Supplementary Methods

Bioinformatics analysis

Processing of raw mapped reads and all downstream analyses were done using the R statistical computing environment, version 2.12.0 (http://www.r-project.org/) (R_Development_Core_Team, 2010) and specialized software packages within the Bioconductor project (Gentleman et al., 2004) (http://www.bioconductor.org/). Total number and quality of sequenced reads per DNA sample is summarized in Supplementary Table 2. We selected uniquely mapped reads with quality score $\geq$ 15 and generated genome-wide coverage by extending reads up to 200 bp. Raw coverage was visualized in the (http://genome.ucsc.edu/) (Kent et al., 2002) in the context of the human genome (International Human Genome Sequencing Consortium, 2001), March 2006 assembly (hg18).

MethylCap-seq data normalization

For normalization of the MethylCap-seq data we adopted the CpG coupling factor–based normalization method from (Chavez et al., 2010) with certain modifications. It has been shown that MeDIP-derived data shows bias that depends on local CpG densities which is caused by a varying efficiency of antibody binding and immunoprecipitation (Down et al., 2008; Pelizzola et al., 2008). MethylCap experiments that use methyl-binding protein domains (MBD) instead of antibodies follow a similar experimental procedure and thus may encounter the same problem. To compute unbiased methylation signals, raw data has to be normalized with respect to local CpG density. Our normalization approach is based on the concept of coupling factors presented by Down et al. (2008). We combined modeling of dependency between increasing total CpG density and increasing mean MethylCap-seq signal proposed by Chavez et al. (2010) with using an artificially fully methylated reference sample to calculate normalization parameters.

The MEDIPS software package (Chavez et al., 2010) was used to generate raw short reads coverage at 10 bp resolution. Corresponding coupling factors (CF) were calculated by counting number of CpG dinucleotides in varying region widths (+/- 200, 300, 500 and 700 bp) around each 10 bp genomic bin. For the full range of CF values we calculated the mean raw methylation signal of all bins associated with a given value of CF and used it to generate a calibration curve.

We tried several different approaches for modelling dependency between local CpG density and MethylCap-seq signal using varying region widths for coupling factor calculation and fitting two different types of models. First we employed the original method proposed by Chavez et al. (2010) that fits a linear model on the initial portion of the calibration curve, which corresponds to low CF values. This approach is based on the assumption that regions in the genome with low

CpG densities are in general fully methylated. Normalization parameters calculated in this way are represented by slope and intercept that are calculated for each sample separately depending on the corresponding calibration curve. However, such a model is derived from a relatively small range of dependency of signal on CpG density that is in real DNA samples limited to lower CF values. Since we conducted a reference experiment using an artificially fully methylated DNA sample we were able to get the real dependency of signal on CpG density over a much larger range of CF values. Therefore, by fitting a model on a calibration curve derived from the fully methylated DNA sample we include a much larger proportion of the data (~ 70%) and the calculated parameters are more reliable. In addition, we tried to fit a logistic model instead of the linear model since it captures much better the trend of saturation that is expected to occur for regions with high CpG density. Although we see the saturation in our fully methylated sample, it is limited to a relatively low range of CF values after which the mean signal drops again, which we assume is due to lower efficiency of artificial methylation for the regions with very high CpG density and the much lower number of fragments with high CpG density available during sequencing.

To select the best performing normalization method we evaluated the resulting normalized signal in two independent ways. Since we had technical replicates for all our samples, this enabled us to monitor the correlation of normalized signal between replicates. In addition, the bisulfite sequencing experiments for 5 selected genomic regions that we carried out to validate the MethylCap-seq results, enabled us to monitor the correlation of normalized signal and actual percentage of methylated CpG dinucleotides in given regions. Results of the evaluation of different normalization parameters are summarized in Supplementary Figure 1. They show that using parameters obtained by fitting a logistic model to the fully methylated sample for normalizing all other samples preserves correlation between technical replicates and in addition keeps a good correlation to the results of bisulphate sequencing. The logistic model fitted to the calibration curve for the fully methylated sample can be described by the following equation:

$$m_{CF} = \frac{Asym}{1 + e^{(\frac{xmid - CF}{scal})}}$$

where *CF* is the coupling factor, *Asym*, *xmid* and *scal* are estimated parameters and *m* is the predicted value of methylation signal in the 10bp genomic bin having a given value of coupling factor.

The obtained parameters (*Asym*, *xmid*, *scal*) were used to normalize raw methylation signals for each sample as follows:

$$rms_i = \frac{m_i}{\frac{nreads_{sample}}{nreads_{reference}} * \frac{Asym}{1 + e^{(\frac{xmid - CF_i}{scal})}}}$$

where m_i is the raw methylation signal of an individual 10pb genomic bin, CF_i is the corresponding value of coupling factor, $nreads_{reference}$ is the total number of sequenced reads for artificially fully methylated reference sample and $nreads_{sample}$ is the total number of sequenced reads for a given sample. Calculated relative methylation scores (*rms*) can be used for comparison of methylation signal across different samples (e.g. detection of differential methylation) since they are corrected for the total number of reads per sample.

Genome-wide methylation profiles: Identification and characterization of DMRs

Two technical replicates for each sample were merged by calculating mean *rms* of replicates for each 10bp bin. Next we calculated mean *rms* values for genome-wide non-overlapping 500 bp windows for all 14 samples to generate genome-wide methylation profiles. Finally, we tested the regions for differential methylation between corresponding tumor and healthy samples using the MEDIPS software package (Chavez et al., 2010) (Chavez et al., 2010). Distributions of 10bp bins *rms* values within each 500 genomic region were compared using both one-sided Student's t-test and one-sided Wilcoxon rank-sum test. Testing was done separately for hypo- and hypermethylation and p-value threshold = 10^{-18} was used to select significant DMRs.

Global background signal threshold was estimated from the fully unmethylated DNA sample used as a background by selecting 90th percentile of *rpm* (reads per million) values for all 500bp genome-wide windows. Regions with either *rpm* value below the selected threshold or with *rpm*

ratio $\frac{rpm_{sample}}{rpm_{input}} < 1.33$ for both healthy and tumor sample were excluded.

For each sample, regions with p-value $\leq 10^{-18}$ were projected onto the genome and visualized as genome-wide maps of differential methylation. Composition of DMRs was calculated on a set of top 5000 hypomethylated and top 5000 hypermethylated regions with lowest p-values as a percentage of regions covered by selected genomic features. We retrieved annotation for human transcripts from Ensembl (V60) (http://nov2010.archive.ensembl.org/index.html) and annotation for repeats from UCSC Genome Browser database (Kent et al., 2002). Non-overlapping sets of distinct classes of repeats were created and these regions were excluded from subsequent genomic features. All Ensembl transcripts were used to generate non-overlapping sets of promoters (region +/- 1 kb around TSS), exons and introns.

Cluster analysis, identification and characterization of common DMRs and associated genes

We created a non-redundant set of genomic regions by combining DMRs with p-value $\leq 10^{-18}$ from all 7 samples. Feature matrix containing $\log_2(\frac{rms_{tumor}}{rms_{healthy}})$ for all 7 samples at all selected regions was used to calculate Manhattan distance between samples as well as between selected regions. We performed hierarchical clustering on obtained dissimilarity matrices and visualized the result as a heatmap. Clusters of DMRs were definded by cutting clustering dendrogram at height=23 and were projected onto the genome to visualize their genomic localization. To measure the enrichment of DMRs in the subtelomeric area, we calculated for each chromosome arm separately the ratio between proportion of 1Mb region at chromosome end and proportion of the entire chromosome covered by DMRs. In addition, DMRs were overlapped against the set of all Ensembl transcripts to reveal the association to distinct parts of genes (promoters: +/- 1 kb around TSS, three prime end: +/- 1kb around TTS, gene body) and percentages of regions within each cluster associated to distinct gene parts was calculated. We identified DMRs common to all samples (or a subset of samples) by selecting regions with $|\log_2(\frac{rms_{tumor}}{rms_{healthy}})| \geq 1$ (two-fold change in methylation signal) for all corresponding samples.

Bisulfite Sequencing

Primers for bisulfite sequencing were designed using a combination of tools, such as MethPrimer (http://www.urogene.org/methprimer/index1.html), Methyl Primer express software v.1.0 from Applied Biosystems, and manual selection (supplementary table 4). Amplification conditions were as follows: 1X PCR buffer (Invitrogen); 200 µM of dNTP; 1.5 mM of $MgCl_2$; 0.2 µM of primers; 0.2 µl of Platinum Taq (Invitrogen) and 2 µl of the bisulfite treated sample. Amplification was carried out in 50 µl using the same conditions as described above. Colony PCR protocol was as follows: 94 ˚C for 5 min, 30 cycles of 94 ˚C for 15s, 55 ˚C for 30s, 72 ˚C for 30 sec, finalized by 72 ˚C for 5 min. PCR products were recovered using NucleoFast 96 PCR clean-up plates (Machery-Nagel), according to manufacturer instructions, and sequenced using T7 primer and a 3730 DNA Analyzer (Applied-Biosystems).

Methylation-specific PCR (MSP)

Samples were amplified in 50 µl reaction volume containing 1X PCR buffer (Invitrogen); 200 µM of dNTP; 1,5 mM of $MgCl_2$; 0,2 µM of each primer; 0,2 µl of Platinum Taq (Invitrogen) and 1 µl of bisulfite treated sample. PCR was hot-started at 95 ˚C for 5 min, with 30 cycles of 95 ˚C

for 30s, specific primer annealing temperature for 30s and 72˚C for 30 sec, finalized by 72˚C for 5 min. Primer sequences, annealing temperatures and product sizes and described in supplementary table 4.

References

Chavez L, Jozefczuk J, Grimm C, Dietrich J, Timmermann B, Lehrach H, et al. (2010). Computational analysis of genome-wide DNA methylation during the differentiation of human embryonic stem cells along the endodermal lineage. *Genome Res*, 20, 1441-50.

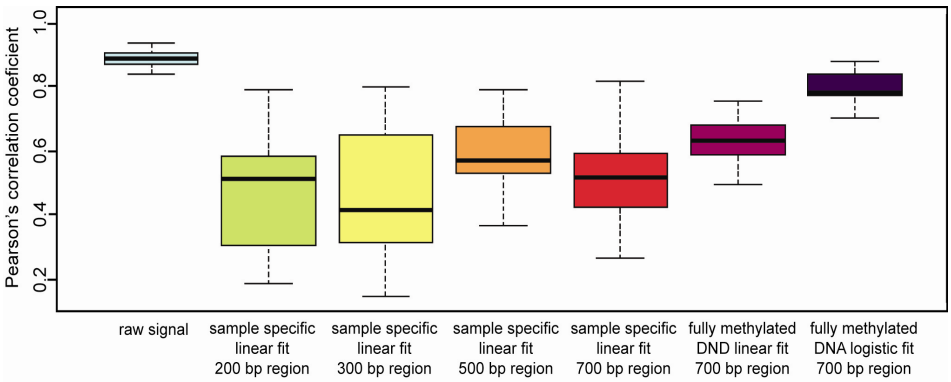
Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, et al. (2008). A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat Biotechnol*, 26, 779-85.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5, R80.

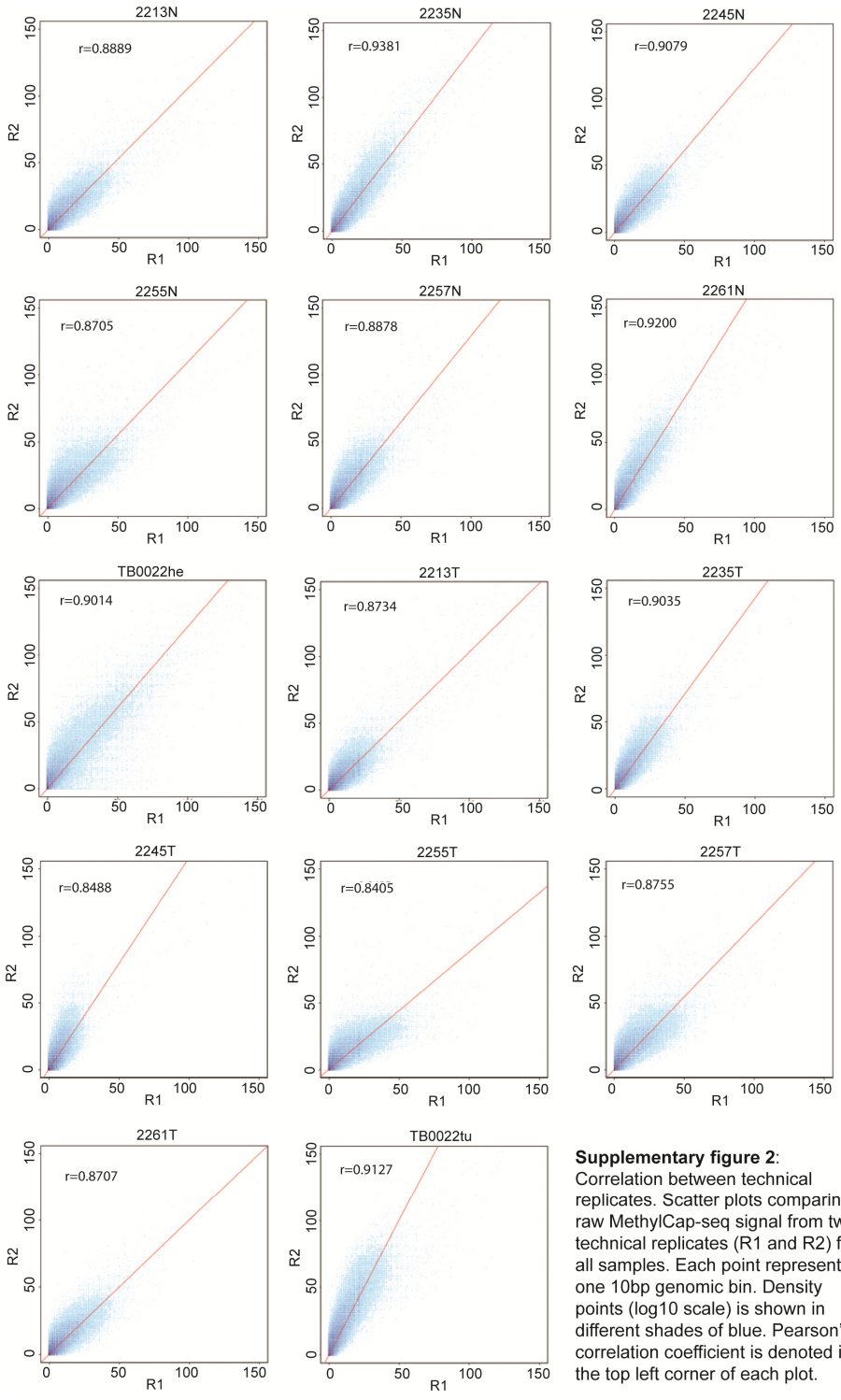
Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. (2002). The human genome browser at UCSC. *Genome Res*, 12, 996-1006.

Pelizzola M, Koga Y, Urban AE, Krauthammer M, Weissman S, Halaban R, et al. (2008). MEDME: an experimental and analytical methodology for the estimation of DNA methylation levels based on microarray derived MeDIP-enrichment. *Genome Res*, 18, 1652-9.

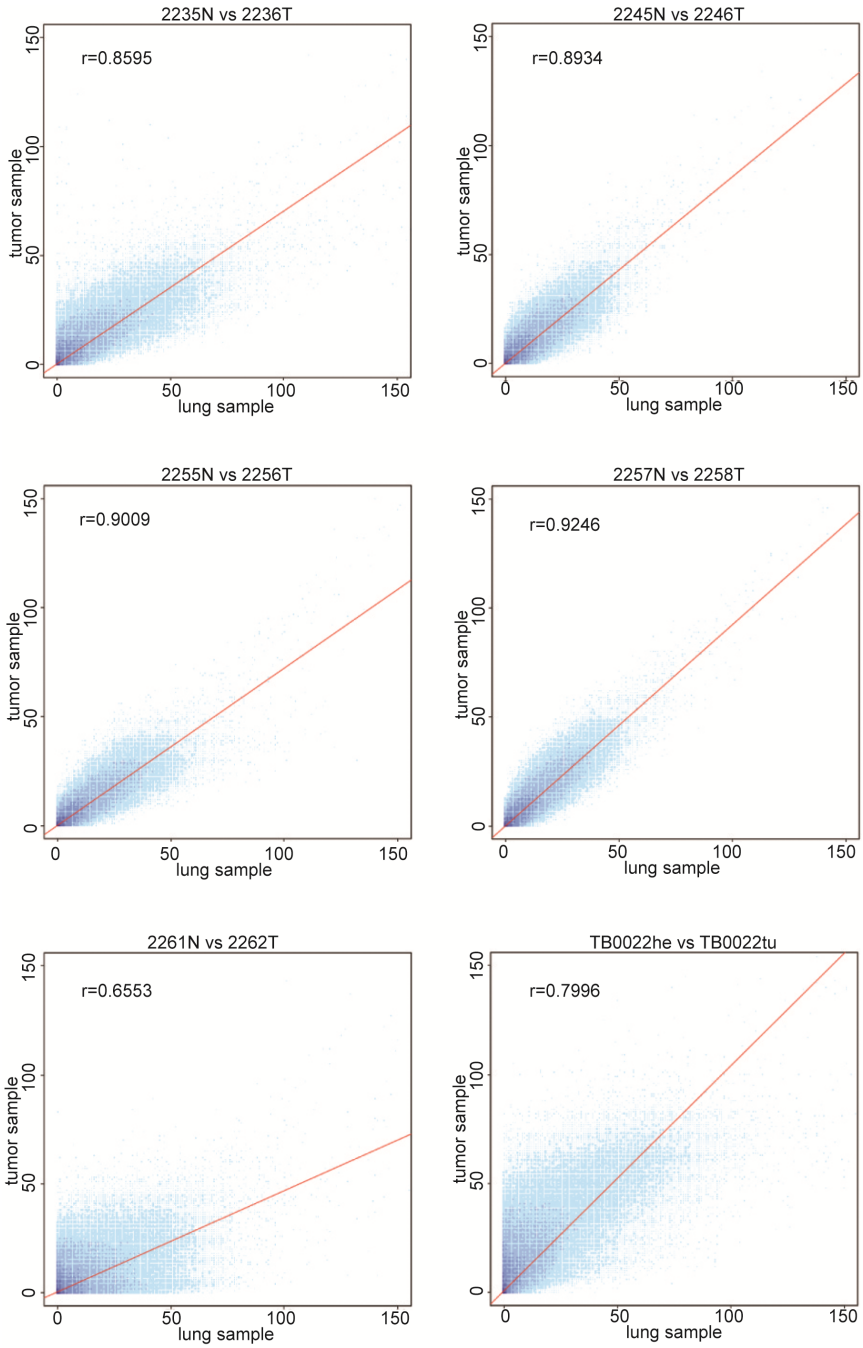
R_Development_Core_Team. (2010). R Foundation for Statistical Computing.



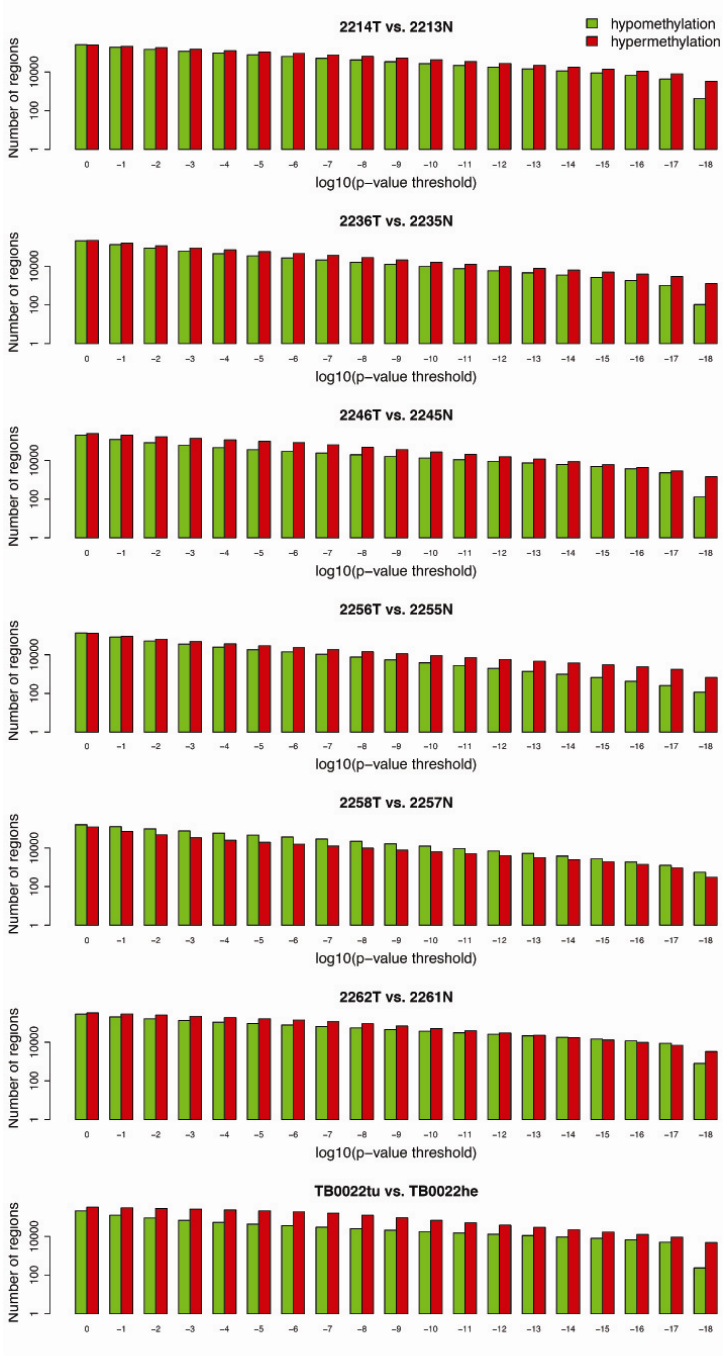
Supplementary figure 1: Selection of normalization method. Distribution of Pearson's correlation coefficients between replicates for different normalization methods.



Supplementary figure 2: Correlation between technical replicates. Scatter plots comparing raw MethylCap-seq signal from two technical replicates (R1 and R2) for all samples. Each point represents one 10bp genomic bin. Density points (log10 scale) is shown in different shades of blue. Pearson's correlation coefficient is denoted in the top left corner of each plot.



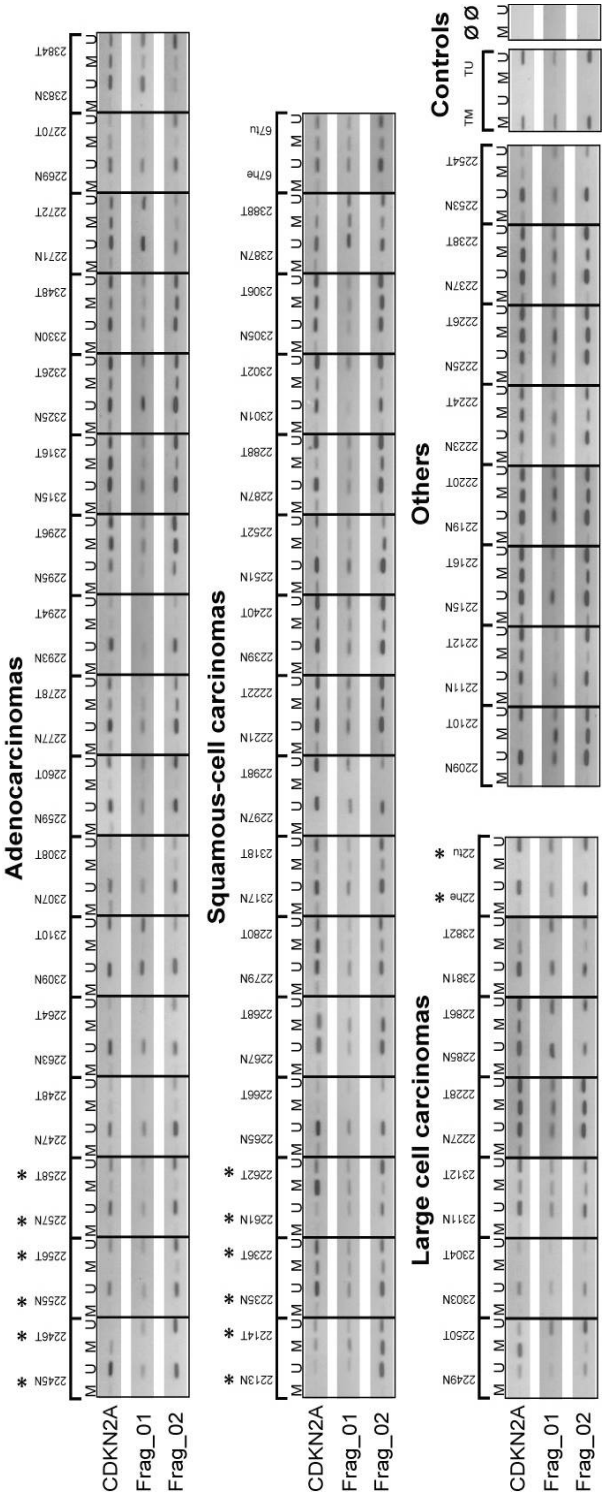
Supplementary figure 3: Correlation between lung and tumor samples. Catter plot comparing raw MethylCap-seq signals from lung samples and tumor samples from 6 samples. Each point represents one 10bp genomic bin. Density of points (log10 scale) is shown in different shades of blue. Pearson's coefficient score is denoted in the top left corner of each plot.



Supplementary Figure 4: Number of significant hypo- and hypermethylated regions for different p-value thresholds. Y-axis shows number of significant regions on log10 scale for decreasing p-value threshold, shown on the x-axis.



Supplementary figure 5: Representation of the log-ratio of relative methylation scores for tumor vs paired lung samples for each samples with DMRs shown along each chromosome. Green color marks hypomethylation in tumor with negatibe log-ration =<-2, and red color marks hypermethylation in tumor with positive log-ratio>=2



Supplementary file 6: MSP screening in NSCLC. Cropped gel images for all 48 samples grouped by histological subtype in the 3 fragments analyzed: 17 ADC, 16 SCC, 7 LCC, 8 others (other types and unclassified NSCLC). TM: Totally methylated control sample; TU: Totally unmethylated control sample; ø: blank/water; U: primer for umethylated CpG; M: primer for methylated CpG; N: normal/healthy; T: tumor; *: samples used in MethylCap-seq.

Supplementary Table 1: All patients/samples used for MSP screening

No.	Sample ID	Tissue type	Case ID	Gender	Age at Dignosis	Ethinic
1	2209N	lung	N/A	N/A	N/A	N/A
	2210T	N/A				
2	2211N	lung	N/A	N/A	N/A	N/A
	2212T	N/A				
3	2213N*	lung	SCC1	M	54,54	1
	2214T*	SCC				
4	2215N	lung	N/A	N/A	N/A	N/A
	2216T	N/A				
5	2219N	lung	N/A	N/A	N/A	N/A
	2220T	N/A				
6	2221N	lung	SCC9	M	73,14	1
	2222T	SCC				
7	2223N	lung	N/A	N/A	N/A	N/A
	2224T	N/A				
8	2225N	lung	N/A	N/A	N/A	N/A
	2226T	N/A				
9	2227N	lung	LCC5	F	65,29	1
	2228T	LCC				
10	2235N*	lung	SCC2	M	73,73	1
	2236T*	SCC				
11	2237N	lung	N/A	N/A	N/A	N/A
	2238T	N/A				
12	2239N	lung	SCC10	M	51,83	1
	2240T	SCC				
13	2245N*	lung	ADC1	F	66,26	1
	2246T*	ADC				
14	2247N	lung	ADC4	M	66,01	2
	2248T	ADC				
15	2249N	lung	LCC2	M	45,98	1
	2250T	LCC				
16	2251N	lung	SCC11	F	67,93	1
	2252T	SCC				
17	2253N	lung	N/A	N/A	N/A	N/A
	2254T	N/A				

continuation

18	2255N*	lung	ADC2	F	54,26	1
	2256T*	ADC				
19	2257N*	lung	ADC3	M	78,96	1
	2258T*	ADC				
20	2259N	lung	ADC8	F	75,08	1
	2260T	ADC				
21	2261N*	lung	SCC3	M	70,03	1
	2262T*	LCC				
22	2263N	lung	ADC5	M	75,12	1
	2264T	ADC				
23	2265N	lung	SCC4	M	63,03	1
	2266T	SCC				
24	2267N	lung	SCC5	M	64,8	1
	2268T	SCC				
25	2269N	lung	ADC16	M	60,27	1
	2270T	ADC				
26	2271N	lung	ADC15	M	72,33	1
	2272T	ADC				
27	2277N	lung	ADC9	M	73,47	1
	2278T	ADC				
28	2279N	lung	SCC6	M	66,06	1
	2280T	SCC				
29	2285N	lung	LCC6	M	77,8	1
	2286T	LCC				
30	2287N	lung	SCC12	M	70,27	1
	2288T	SCC				
31	2293N	lung	ADC10	M	59,66	1
	2294T	ADC				
32	2295N	lung	ADC11	M	62,27	1
	2296T	ADC				
33	2297N	lung	SCC8	M	59,91	1
	2298T	SCC				
34	2301N	lung	SCC13	M	76,71	1
	2302T	SCC				
35	2303N	lung	LCC3	F	69,57	1
	2304T	LCC				

continuation

36	2305N	lung	SCC14	M	62,15	1
	2306T	SCC				
37	2307N	lung	ADC7	F	69,34	1
	2308T	ADC				
38	2309N	lung	ADC6	F	40	1
	2310T	ADC				
39	2311N	lung	LCC4	F	64,15	1
	2312T	LCC				
40	2315N	lung	ADC12	M	52,24	1
	2316T	ADC				
41	2317N	lung	SCC7	M	62,7	1
	2318T	SCC				
42	2325N	lung	ADC13	M	62,64	1
	2326T	ADC				
43	2330N	lung	ADC14	M	63,54	1
	2348T	ADC				
44	2381N	lung	LCC7	N/A	N/A	2
	2382T	LCC				
45	2383N	lung	ADC17	N/A	N/A	2
	2384T	ADC				
46	2387N	lung	SCC15	N/A	N/A	2
	2388T	SCC				
47	22he*	lung	LCC1	F	56,58	1
	22tu*	LCC				
48	67he	lung	SCC16	M	79,89	1
	67tu	SCC				

* Samples used in MethylCap-seq.
ADC: Adenocarcinoma; SCC: Squamous cell carcinoma; LCC: :Large cell carcinoma

Supplementary Table 2: HTS reads from all samples and replicate experiments

Sample	Sequenced reads		Mapped reads (Q>=15)		Unique reads	
	R1	R2	R1	R2	R1	R2
2213N	23520751	43432658	10454508	17176471	9526861	13529253
2214T	22108659	39423791	11539328	19160318	10551768	14187876
2235N	22303747	38361502	10717172	16824208	9619047	13709204
2236T	16854176	38654041	8793685	18607361	8141431	14484869
2245N	20921229	36118546	9367036	17022646	8132802	12726159
2246T	22109493	37776051	10304629	18058216	7333999	14333853
2255N	24104731	45932423	11250236	19361461	10087412	16160441
2256T	20375350	43599897	10332751	19601611	9441639	14148718
2257N	20911413	44361193	10388448	19124272	9622326	15875606
2258T	22485475	45629032	11219874	20254399	10174203	16703815
2261N	21728715	40367488	8087723	17594812	7068828	14692164
2262T	24630635	47033404	11463288	17307368	10178066	13784052
TB0022he	22880893	42675632	9607363	17845094	8104807	13948856
TB0022tu	24671424	108128892	11372693	52157802	10374798	27384505
MRC_unmet	-	43430352	-	21477328	-	18461017
MRC_met	-	45768909	-	24587600	-	19073379
UUD_unmet	-	45270573	-	22839551	-	21081239
UUD_met	-	42716038	-	25148911	-	18606055

For both technical replicates (R1, R2) total number of reads, number of mapped reads with Solexa quality scores >= 15, and number of uniquely mapped reads per sample is shown. Fully methylated and unmethylated reference samples were done in biological replicates with no technical replicates. MRC_unmet = totally unmethylated DNA derived from MRC-5 cell line; MRC_met = totally methylated DNA derived from MRC-5 cell line; UUD_met = totally unmethylated DNA derived from commercial totally unmethylated DNA; UUD_met = totally methylated DNA derived from commercial totally unmethylated DNA.

Supplementary Table 4: Primers used for MethylCap-seq validation

Primers used for Bisulfite sequencing

		Name		Sequence	Ta (°C)	product size
bisulfite sequencing	cloning	CDKN2A	F	5'-TATTTGTGGGTTTGTAGAAGTAGGT-3'	58	348
			R	5'-CCTAATCCAACAACTAATCCACTAC-3'		
		RASSF1	F	5'-GGGGAGTTTGAGTTTATTGAGT-3'	58	297
			R	5'-CTACCCCTTAACTACCCCTTC-3'		
		Frag_01	F	5'-GGTGAATTTTAGGGTTTTGAGAAGT-3'	54	349
			R	5'-ACCCTCACAAAATCCAAAATTC 3'-3'		
		Frag_02	F	5'-GGGAGGGGGTTATTGTAATGTTAAG-3'	58	304
			R	5'-AACTCTCACACAAATTCACTCACTA-3'		
		EN1_prom	F	5'-GATGTGAAATAGATTTAGAGTTGGT -3'	55	399
			R	5'-AAAAATCTAACACAAAAATTACCC -3'		
	Seq.	MF-20i	F	5'-GTAAAACGACGGCCAG-3'		
		M13-reverse	R	5'-CAGGAAACAGCTATGACC-3'		
MSP		M_CDKN2A	F	5'-AGCGTTTTAGGTTTATTCGGC-3'	56	201
			R	5'-GCAAACTAACTAACTCACTCCGC-3'		
		U_CDKN2A	F	5'-TAAGTGTTTTAGGTTTATTTGGTGG-3'	56	203
			R	5'-ACAAACTAACTAACTCACTCCACA-3'		
		M_Frag_01	F	5'-GGAGGTTTTATTATTGTCGC-3'	56	119
			R	5'-AAACGAACGCAAACTCTC-3'		
		U_Frag_01	F	5'-ATTGGAGGTTTTATTATTGTTGT-3'	56	119
			R	5'-AAACAAACACAAACTCTCAAC-3'		
		M_Frag_02	F	5'-GGGAACGAGTGTGTAGTTTC-3'	56	192
			R	5'-ACTATTCACGTAAACACGCA-3'		
		U_Frag_2	F	5'-GAGGGGAATGAGTGTGTAGTTTT-3'	56	192
			R	5'-AAAACTATTCACATAAACACACA-3'		
qPCR		H19-ICR	F	5'-TCCTATTCCCAGATGACCCC-3'	60	149
			R	5'-AGTTGTGGAATCGGAAGTGG-3'		
		RASSF1	F	5'-CCGAGGACTAGGTCCGTTAC-3'	60	81
			R	5'-TACGAGTGGAGTGCGACAAG-3		
		APRT	F	5'-GCCTTGACTCGCACTTTTGT-3	60	85
			R	5'-TAGGCGCCATCGATTTTAAG-3'		

Ta = Annealing temperature used for primers. Seq.: sequencing primers



3

Comprehensive genome-wide methylation analysis of NSCLC reveals novel methylated regions for use as epigenetic tumor markers

Rejane Hughes Carvalho[1,2], Jun Hou[3], Vanja Haberle[4], Joachim Aerts[5], Frank Grosveld[1,2,3,6], Boris Lenhard[4] and Sjaak Philipsen[1,2,3]

[1] Department of Cell Biology, [2] Centre for Cancer Genomics, [3] Dutch Consortium for Systems Biology, [5] Department of Lung Diseases, [6] Center for Biomedical Genetics, Erasmus MC, Rotterdam, The Netherlands
[4] Uni BCCS, University of Bergen, Norway

(Manuscript in preparation)

Abstract

DNA methylation is part of the epigenetic regulatory mechanism present in all normal cells. It is tissue specific and stably maintained throughout development, but abnormally changed in cancer. Non-small cell lung carcinoma (NSCLC) is the most deadly type of cancer, involving different tumor subtypes. This heterogeneity is a challenge for correct diagnosis and patient treatment. The stability and specificity of DNA methylation makes it a very suitable marker for epigenetic phenotyping of tumors. To identify candidate markers for use in NSCLC diagnosis, we used genome-wide methylation maps that we had previously generated by MethylCap and next generation sequencing, and listed the most significant differentially methylated regions (DMRs). For screening a larger number of samples we selected 25 DMRs from the list of most differentially methylated regions that we had previously identified by MethylCap-seq. The methylation status of these DMRs was investigated in 61 tumors and matching lung tissues by methylation-specific PCR. We found 12 novel DMRs that show significant difference between tumor and lung tissues. We also identified three novel DMRs for each of the two most common NSCLC subtypes. We propose a panel of five DMRs, composed of novel and known markers, which exhibit 100% specificity and 94% sensitivity to distinguish tumors from lung tissues. The novel markers will aid the development of a highly specific epigenetic panel for accurate identification NSCLC tumors.

Introduction

DNA methylation is an epigenetic modification present in the genome and an essential regulatory mechanism involved in crucial processes for normal development (Suzuki and Bird, 2008; Takebayashi et al., 2007). It is an important player in chromatin structure, where wide-spread methylation in the DNA leads to the formation of heterochromatin (Li, 2002). Another process involving DNA methylation is silencing of viral and repetitive elements that have been incorporated into the genome over time (Esteller, 2007; Miura et al., 2001). DNA methylation patterns are tissue-specific and these patterns can be stably inherited (Bird, 2002) .

It is known that tumor cells often display methylation patterns different from their normal counterparts (Sharma et al., 2010). Therefore, many studies have focused on finding and quantifying these differences (Belinsky et al., 2003; Bird, 1986; Eads et al., 2000; Herman et al., 1996; Usadel et al., 2002; Zochbauer-Muller et al., 2001). To date, many genes are known to be aberrantly methylated in virtually all types of cancer. Epigenetic inactivation of tumor suppressor genes (TSG) by DNA hypermethylation is believed to be an early and frequent event in oncogenesis (Belinsky et al., 1998; Huang et al., 2011). TSGs, such as RASSF1, CDKN2A, DAPK, APC and p14 are aberrantly methylated in a variety of cancers (Calmon et al., 2007; Daniunaite et al., 2011; Jeong et al., 2006; Richiardi et al., 2009; Verdoodt et al., 2011), including Non-small cell

lung carcinoma (NSCLC) (Belinsky, 2004; Belinsky et al., 1998; Grote et al., 2008; Kim et al., 2003; Niklinska et al., 2009; Zhang et al., 2011).

NSCLC is the cancer with the highest incidence and mortality rates (Ferlay et al., 2010). It is also a very heterogeneous malignancy, involving different tumor subtypes. According to the World Health Organization (WHO), NSCLC can be broadly divided into three major groups: adenocarcinoma (ADC), squamous cell carcinoma (SCC) and large cell carcinoma (LCC). Within these three groups, and including other less frequent types, NSCLC presents more than 30 histopathological subtypes (Travis et al., 2004). This heterogeneity forms a great problem for diagnosis, which is still currently based on histological classification aided by immunohistochemical staining (Travis et al., 2011). This degree of variability presents a challenge for unambiguous histological classification of NSCLC cases. Genetic and epigenetic studies show that tumors belonging to the same traditional histological classification present different molecular signatures (Travis et al., 2011). This indicates that, although the tumors are part of the same group, they may present different aetiologies, behave differently in growth, predisposition to develop metastasis and also respond differently to treatment.

The amount of data on methylated genes in NSCLC have been accumulating in the past years (Heller et al., 2010; Palmisano et al., 2000; Toyooka et al., 2001; Zochbauer-Muller et al., 2001), however not only the identified methylated loci differs between studies, but the level of methylation is often divergent. Moreover, there is a lack of methylation markers able to discriminate between different histological subtypes.

We have recently performed genome-wide methylation profiling of NSCLCs, where we identified differentially methyalted regions (DMRs) in tumors when compared to matching lung lung. Among the DMRs identified there were several regions which appeared to be tumor- and subtype specific. Here, we used those DMRs to screen 61 NSCLC tumors and matching lung lung by Methylation-specific PCR (MSP). We investigated the potential of 25 DMRs to distinguish tumor from lung lung with high specificity and sensitivity, and to discriminate between the two most common NSCLC subtypes (ADC and SCC).

Results

Selection of DMRs for MSP

For screening a larger number of samples (n = 61), we selected 25 DMRs (Table 1) from the list of most differentially methylated regions that we had previously identified by MethylCap-seq (Chapter 2). Among these 25 DMRs, 10 were present in all 7 tumors used for MethylCap-seq, 5 were specific for the 3 SCC samples, and another 5 were specific for the 3 ADC samples. The remaining 5 DMRs were used for validation in the previous study and therefore included here. Selection was based on 5 criteria: 1) Regions with highest $\log_2$ ratio across all sample of the interest group (all tumors, SCC or ADC). 2) Length of the region spanned by the DMR which presented as significant

(p-value ≤ E-18). Because relative methylation scores were calculated based on 500 bp bins, each DMR is represented by a 500 bp window. Regions that showed more than one serially significant DMR were preferred 3) Region of the gene where methylation was observed: promoter, 3'UTR or gene body; 4) Likelihood of the gene to play a role in tumorigenesis and 5) Novelty of the region.

Table1: DMRs selected from the genome-wide methylation profiles

	DMR	log2 ratio of nms tumor vs. lung sample							gene region	HGNC ID	Gene Class
		2214T	2236T	2246T	2256T	2258T	2262T	22tu			
Tumor DMRs	BARX1	6,96	5,10	4,46	4,20	4,22	4,36	3,99	body	955	TR
	DMRTA2	2,99	3,16	2,15	2,78	1,92	4,44	4,32	body	13908	TR
	FGF12	3,52	6,67	2,94	3,17	1,49	3,95	4,21	promoter	3668	Growth factor
	GATA3	2,75	4,13	2,97	2,82	2,12	3,10	3,01	promoter	4172	TR
	GRIK2	5,34	1,63	1,22	1,74	1,75	3,91	3,31	promoter	4580	Ion channel
	ONECUT2	3,73	3,84	2,84	3,04	1,98	5,86	3,92	body	8139	TR
	PARP15	2,86	2,91	3,94	2,50	2,51	2,77	2,48	promoter	26876	TR
	PRDM14	4,23	3,81	3,65	5,36	1,48	3,77	2,43	promoter	14001	TR
	RAX	5,22	2,92	2,31	2,81	1,77	4,18	5,48	promoter	18662	TR
	TBX15	3,54	2,36	2,48	1,91	1,60	3,26	1,27	body	11594	TR
SCC DMRs	ZIC4	7,69	2,38	0,00	0,00	-0,15	6,58	0,00	promoter	20393	TR*
	DOT1L	3,61	6,32	0,00	0,00	0,00	4,65	0,00	1rst intron	24948	Chromatin regulator
	LIMK1	4,30	5,51	0,00	0,00	0,00	4,44	-0,98	body	6613	Cell organization
	ANKRD1B	6,57	4,49	0,00	0,00	0,00	3,74	0,00	3'UTR	26363	Unknown
	HIST1H3C	3,44	4,45	0,00	0,00	0,00	6,99	0,00	promoter	4768	Histone
ADC DMRs	MSC	0,00	0,00	2,98	5,58	4,66	0,00	0,00	promoter	7321	TR
	GAS1	0,00	0,00	2,06	2,79	4,94	0,00	0,00	promoter	4165	Cell cycle
	HOXA1	0,00	0,00	3,96	2,88	3,86	0,00	0,00	promoter	5099	TR
	SEPT9	0,00	0,00	5,35	3,02	1,44	0,00	0,00	body	7323	Enzyme
	FAM78B	0,00	0,00	5,02	4,30	2,96	0,00	0,00	1rst intron	13495	Unknown
	CDKN2A	2,73	2,73	0,75	-1,16	1,78	5,72	0,00	3'UTR	1787	TS
	RASSF1	NA	NA	NA	NA	NA	NA	NA	promoter	9882	TS*
	EN1	NA	NA	NA	NA	NA	0,00	0,00	promoter	3342	TR
	Frag_01	5,09	4,21	0,00	0,00	0,00	2,64	0,00			
	Frag_02	NA	NA	0	NA	NA	NA	0,00			

Rms: relative methylation score; HGNC ID: HUGO gene nomenclature committee. NA: not available; TR: trscription regulator; TS: tumor suppressor; * putative.

Figure 1 shows 8 of the 25 DMRs for one paired sample (lung vs tumor), where multiple DMRs are observed in each fragment. Among the 25 selected DMRs, one is located within a genomic region about 10kb upstream the EN1 gene (chr2:119328097-1193284, referred to as Frag_01). Because we observed the presence of a highly differentially methylated genomic region located at 2q14.2 and spanning more then 25Kb (chr2:119,307,647-119,333,875), another two regions within this area were selected: Frag_02 (chr2:119331343-119331692) and the promoter of the EN1 gene (Figure 1F). RASSF1 was selected since it is frequently reported as methylated in NSCLC. RASSF1, Frag_02 and EN1 promoter did not appear in the list of DMRs since their p-values are -14,71, -17,63 and -17,58 respectively, falling outside the cutoff p-value <-18.

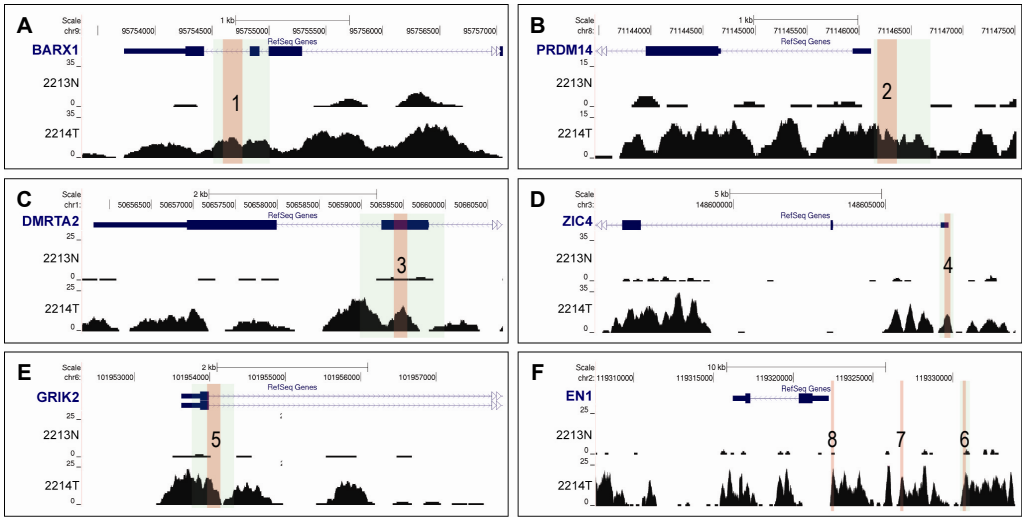


Figure 1: Methylation reads for selected DMRs. UCSC genome browser representation of the sequenced methylated reads in 6 DMRs for a paired sample: lung sample (2231N) in the top panel of each figure and tumor sample (2214T) in the bottom panel of each figure. Light green: region represented as significantly differentially methylated; light red: region selected for MSP. Y axis: number of reads ranging from 0 to the maximum number of reads in the highest peak in the selected window per sample. **A**: DMRs in the BARX1 (1) gene spanning more than 3000 bp, with 500 bp significant bin and 172 bp MSP fragment. **B**: DMRs in the PRDM14 (2) gene spanning from 5'UTR to first exon (promoter region) with a 1000 bp bin and 194 bp MSP fragment. **C**: Patterns of DMRs in the DMRTA2 (3) gene with a 1000 bp significant bin and 155 MSP fragment. **D**: Spread of DMRs in the ZIC4 (4) gene in the promoter and first exon, showing a 500 bp significant bin in the promoter and 209 MSP fragment. **E**: DMRs in the promoter region of the GRIK2 (5) gene, with a 500 bp bin and 179 bp MSP fragment. **F**: Patterns of methylation in the EN1 gene showing DMRs spanning more then 25 kbp, and 500 bp bin about 10 kbp ustream the gene and 3 MSP fragments in 10 kbp upstream (6), 6 kbp upstream (7) and EN1 promoter (8) with 119 bp, 192 bp and 214 bp respectively.

Methylation profiles by MSP

We performed MSP for determining the methylation status for all 25 DMRs in the 61 primary tumors and matching lung tissues (Figure 2 and Supplementary Figure 1.A1-A25). For this, we used two sets of primers for each DMR: one set (M) amplifies

methylated DNA; the other (U) requires unmethylated DNA for amplification. We counted the numbers of samples that showed amplification for each of the primers, considering any amplification as "1" and no amplification as "0" (supplementary Table 1). We then calculated the percentages of samples that were amplified with the methylated or unmethylated primers (Table 2).

We found that PARP15, CDKN2A, HIS1H3C, ANKRD1B, DOT1L, GAS1 and SEPT9 presented no significant difference between lung sample and tumor considering a p-value <0.001. PARP15 and CDKN2A were selected as tumor DMRs and showed high sensitivity but low specificity, while DOT1L, GAS1, SEPT9, HIST3B and ANKRD1B were selected as subtype DMRs. GAS1 and DOT1L showed a moderate specificity and sensitivity for differentiating between normal and tumor, and SEPT9 low sensitivity but very high specificity. We observed that lung samples were methylated for CDKN2A and PARP15 in more than 90% of the cases, although the levels of methylation were much lower compared to the matched tumors (Figure 2B). This was also observed for DOT1L and GAS1 although at a much lower level since the sensitivity of these DMRs was much lower. HISTH3C and ANKRD1B were removed from further analysis because both were methylated in more than 90% of the lung samples, with no difference observed in degree of amplification (Figure 2C). Although SEPT9 had very low sensitivity, there was no amplification in normal samples.

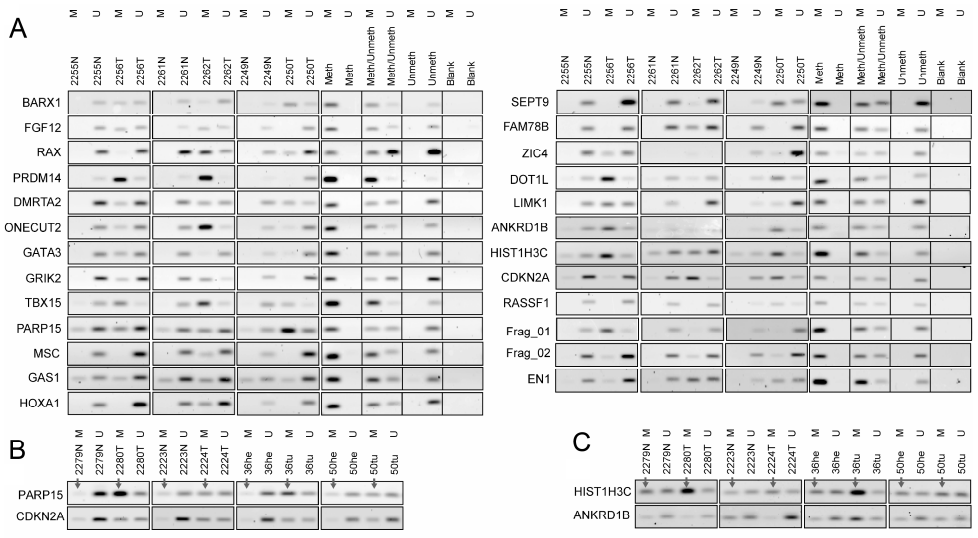


Figure 2: Representation of MSP screening. **A**: Cropped gel image of 3 paired samples and controls for all DMRs used. Meth: Totally methylated control; Unmeth: totally umethylated control; M: primer for methylated CpG; U: primer for unmethylated CpG. **B**: Lung samples in PARP15 and CDKN2A show less intense amplification than tumors samples. **C**: Intensities of lung samples in HIST1H3B and ANKRD1B are not significantly different from tumors. Arrows points to methylated primer in lung and tumor.

Table 2: Methylation status of Lung vs Tumor for all DMRs

	Total (n)	Methylation status (M primer)				p-value
		Lung		Tumor		
		n	%	n	%	
BARX1	56	2	3.57	46	82.14	8.82E-19
FGF12	61	1	1.64	44	72.13	1.16E-17
PRDM14	53	5	9.43	43	81.13	2.92E-14
RAX	60	1	1.67	49	81.67	2.25E-21
DMRTA2	61	2	3.28	44	72.13	2.09E-16
ONECUT2	59	3	5.08	49	83.05	3.87E-19
PARP15	59	55	93.22	59	100.00	0.1186
GATA3	56	2	3.57	44	78.57	2.63E-17
GRIK2	60	7	11.67	54	90.00	4.13E-19
TBX15	60	12	20.00	57	95.00	3.85E-18
CDKN2A	60	55	91.67	60	100.00	0.05731
RASSF1	56	0	0.00	23	41.07	1.37E-08
ZIC4	57	0	0.00	40	70.18	2.96E-17
DOTL1	60	38	63.33	53	88.33	0.0024
LIMK1	56	1	1.79	28	50.00	1.54E-09
HIS1H3C	59	56	94.92	58	98.31	0.6185
ANKRD1B	51	46	90.20	48	94.12	0.7152
Frag_01	52	0	0.00	33	63.46	1.10E-13
Frag_02	55	1	1.82	39	70.91	2.03E-15
EN1	56	15	26.79	46	82.14	3.77E-07
MSC	58	0	0.00	16	27.59	9.20E-06
GAS1	59	24	40.68	37	62.71	0.0266
HOXA1	57	0	0.00	13	22.81	0.0001
SEPT9	59	0	0.00	7	11.86	0.0129
FAM78B	61	0	0.00	22	36.07	4.70E-08

p-value calculated by Fisher exact test considering a confidence interval of 95% and two sided p-value for p[O>=E|O<=E].

For the remaining 10 DMRs, we observed that the difference between methylation status of lung samples and tumor samples were highly significant. We found that the percentage of methylated tumor samples for RASSF1 in our cohort reflects previous reports (Pfeifer et al., 2002; Wang et al., 2011). All other DMRs tested had a high percentage of methylated tumors, varying from 100% to 72%. Regarding the DMRs selected for histological subtypes (Table 3), one of the ADCs DMRs (FAM78B) was highly significant ($p<0.001$) to discriminate between SCC and ADC, while other two ADC DMRs (MSC, GAS1) and a SCC DMR (DOT1L) showed a moderate significance ($0.001 < p < 0.01$). We observed that the DMRs selected for

histological subtype presented a lower potential to discriminate between the two histological subtypes than between lung and tumor samples.

Table 3: Methylation status of SCC vs ADC for histological subtypes DMRs

	Methylation status tumor						p-value
	SCC			ADC			
	Total (n)	Meth (n)	%	Total (n)	Meth (n)	%	
ZIC4	27	23	85,19	16	16	56,25	0.0679
DOT1L	27	27	100,00	19	20	78,95	0.0101
LIMK1	25	17	68,00	18	18	33,33	0.0333
Frag_01	26	21	80,77	14	14	50,00	0.0705
Frag_02	24	20	83,33	18	11	61,11	0.1586
MSC	27	4	14,81	18	18	61,11	0.0029
GAS1	25	14	56,00	20	20	90,00	0.0197
HOXA1	25	3	12,00	19	19	42,11	0.0353
SEP9	27	2	7,41	18	18	22,22	0.1989
FAM78B	27	5	18,52	20	20	70,00	0.0007

Semi-quantitative MSP: intensities and percentages of methylated DNA per sample

Because we observed that amplification levels with methylated primers were often lower in lung samples than in their paired tumors, we decided to measure the band intensity of each sample and determine whether this difference was significant or not. We scanned the gels using a Typhoon 9410 scanner and used ImageQuant software to determine the intensity of each band. For this we selected each individual band (sample) and the equivalent area above and bellow each one (individual background). We then asked the software to calculate the pixels intensities within each selected area. We first subtracted the sample intensities of their individual backgrounds. Intensities of unmethylated and methylated primers were then normalized using a correction coefficient calculated from each control sample (Supplementary Table 2 and Supplementary File 1).

Distributions of the intensities of each DMR with the methylation primer are shown in Figure 3. Although a large variation between tumors was observed, the difference in methylation intensities was highly significant between lung and tumor for all DMRs (Figure 3A), including CDKN2A and PARP15, which were not significant for the band counts, since about 90% of lung samples showed amplification with the methylated primer. By intensities, those two fragments are highly significant to differentiate between lung and tumor. For the histological DMRs (Figure 3B), we found that the intensities of amplified bands follow in general the same line as for the

band counts. FAM78B and GAS1 appear still significant to differentiate ADC from SCC, while DOT1L intensities show more significant to distinguish SCC versus ADC (Figure 3C). The other DMRs were not representative for either ADC or SCC.

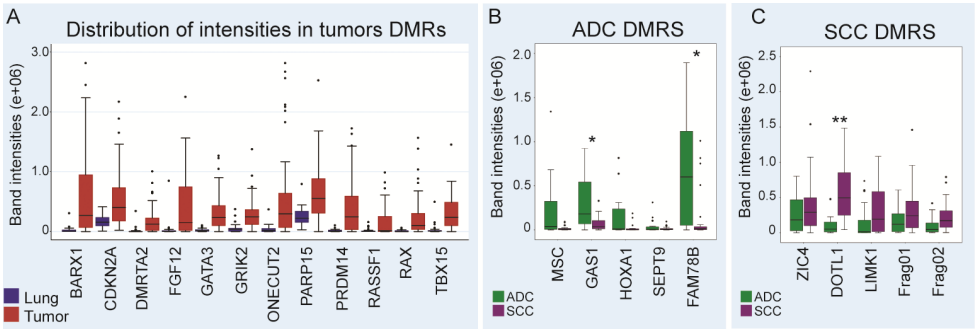


Figure 3: Intensities for methylated bands. **A**: Distribution of DMRs selected for tumor; blue: lung sample (N); red: tumor sample (T); p-value<0.00001 for all DMRs. **B**: Intensities distribution of ADC DMRs in ADC vs SCC tumors. **C:** Intensities distribution of SCC DMRs in ADC vs SCC tumors: ADC tumors: red; SCC tumors: blue. *p-value<0.001; **p-value<0.0001.

Considering significance was found in all tumor DMRs, we then used the band intensities for methylated and unmethylated primers to calculate the level of DNA methylation (percentages) in each sample (Figure 4). For this we established M+U = 100%. We then divided the intensity of the methylated primer by the sum of the intensities with the methylated and unmethylated primers (M / (M+U)) (Supplementary Table 3). We found that the percentage of methylation in tumors was also significantly higher than lung samples for all DMRs (Figure 4A). Because the percentages of DNA methylation are reflected by the intensities of the bands, we also observe here a large

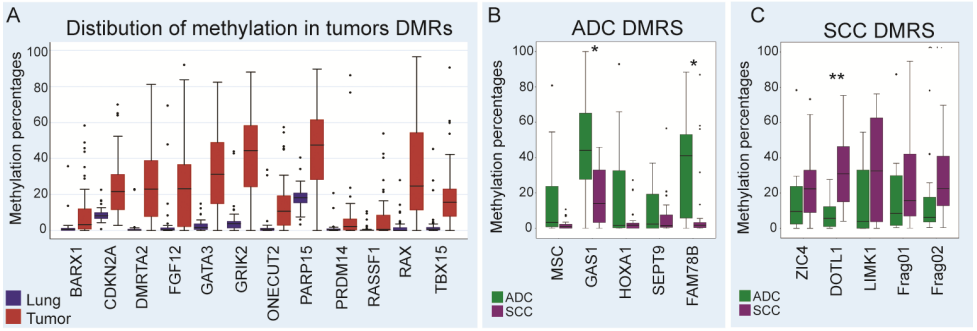


Figure 4: Percentages of DNA methylation. **A**: Distribution of DMRs selected for tumor; lung sample (N): blue; tumor sample (T): red; p-value<0,0001 except for PRDM14 (p=0.0016). **B**: Percentages distribution of ADC DMRs in ADC vs SCC tumors. **C**: Distribution of SCC DMRs in ADC vs SCC tumors *p-value<0.001; **p-value<0.0001.

variation in the tumor samples. However the median difference of DNA methylation between tumor and lung samples is higher using percentages, since it takes in account both methylated and unmethylated DNA content per sample and the unmethylated content in lung samples were higher than in the tumors. PRDM14 was less significant and RASSF1 had a much smaller significance level than for intensities. All other DMRS followed the same model as in the band intensities.

Tumor markers

Next, we wished to determine which group of DMRs would give the best tumor prediction. Since differences were observed between results for the standard MSP results (band counts) and the calculated level of DNA methylation per samples (DNA-methylation percentage), regarding the significance of the individual DMRs to distinguish between tumor and lung tissue, we used both methods, where two panels were generated. Using the DNA-methylation percentages, we calculated tumor specificity and sensitivity for each DMR (Supplementary Figure 2) by logistic regression to estimate the best panel of markers to differentiate between tumor and lung samples. We found a 4 marker panel composed of BARX1, GRIK2, MSC and SEPT9 which together had a 100% tumor sensitivity and 95% tumor specificity. By standard MSP approach, we found another panel of 5 DMRs (FGF12, GATA3, RAX, TBX15 and RASSF1) that presented 100% specific for tumors and 94% sensitivity for detection of tumors (Table 4). Using a test set with 63 samples and three different methods (PAM, SVM and nearest centroid) we could correctly predict 57 to 60 of the 63 cases depending on the predicting method used.

Table 4: Tumor markers panel

		Logistic regression	PAM	SVM	Nearest Centroid
5 markers (FGF12, GATA3, RAX, TBX15, RASSF1)	AUC	1,000	0,955	0,955	0,955
	Tumor sensitivity	0,940	0,940	0,935	0,935
	Tumor specificity	1,000	0,960	0,929	0,929
Train set (n=59)	Correct Prediction (%)		95,00	93,00	93,00
Test set (n=63)	Correct Prediction (%)		92,00	90,50	95,00

AUC: Area under the curve; PAM: Partitioning Around Medoids; SVM: Support Vector Machine.

Discussion

Many methods have been developed to investigate site specific methylation, but the simplest and widest used is MSP. In this study we assessed the methylation status of 25 DMRs in 61 NSCLC patients, comparing methylation differences in tumors and lung

tissues by MSP. We selected genes and genomic regions based on genome-wide methylation profiles we have performed in seven of those 61 patients (Chapter 2), and focused on novel and infrequently reported genes. As reference we used CDKN2A and RASSF1 which have been frequently reported as methylated in NSCLC.

When using band counts we found that the methylation status of CDKN2A and PARP15 were not significant to differentiate between lung and tumor. They presented high sensitivity but no specificity for tumors, however we observed a much lower methylation level in lung samples than in their paired tumors. CDKN2A hypermethylation in known as an early event in carcinogenesis (Belinsky, 2004; Belinsky et al., 1998), therefore tissues which appear histological normal may already present early cellular transformations that will eventually lead to oncogenesis. Although methylation rates in the paired lung and tumor tissues were not significant for those two genes, when we measured the percentages of DNA methylation in lung and tumor samples we observed a highly significant difference between the paired tissues. Interestingly, we observed a highly similar pattern in the PARP15 gene. Poly (ADP-ribose) polymerase (PARP) is a family of transcription regulators playing an important role in cellular processes such as DNA repair and apoptosis. PARP1 is the most studied protein of this family, and its role in different types of cancer has been investigated (Domagala et al., 2011; Gangopadhyay et al., 2011; Sterling et al., 2006). PARP1 promoter hypermethylation has been reported to be involved in leukemogenesis (Gao et al., 2010). Little is known about the other proteins of the same family, including PARP15 and its role in tumorgenesis or its methylation status in NSCLC. In our genome-wide profiles (chapter 2) we found that many genes of the PARP family (PARP1, PARP4, PARP6 and PARP8) were hypermethylated in at least one of the tumor samples, and PARP15 appeared hypermethylated in all seven tumors investigated. Considering the importance of this family of proteins in regulating key pathways involved in oncogenesis (Gaymes et al., 2009; Javle and Curtin, 2011; Le Page et al., 2003; Moroni et al., 2009; Smulson et al., 2000), our findings may point to the direction that, as for CDKN2A, PARP15 hypermethylation may also be an early even in tumorgenesis. However, further investigation is needed to proof this hypothesis. In addition to CDKN2A and PARP15, we found GRIK2, GATA3, DMRTA2, RAX, Frag01, Frag02, EN1, TBX15, FGF12, LIMK1, ONECUT2, ZIC4 and BARX1 among the DMRs with highest sensitivity and specificity and highly significant with both MSP counts and percentages of DNA methylation.

Since we focused on genes which were poorly or not at all reported as methylated in cancer and/or NSCLC, not much is known about them or the role that methylation of those genes plays in tumorigenesis. We note that many of the genes are members of the large homeobox family of transcription factors, or genes known to have some involvement in neural development.

In this study we found that GRIK2 was hypermethylated in 90% of the tumors. GRIK2 is one of the genes belonging to the glutamate receptor family, which are

ligand-activated ion channels. The most studied role of this gene is neurotransmission regulation (Bah et al., 2004; Delorme et al., 2004; Motazacker et al., 2007; Zeng et al., 2006). Little is known about its role in oncogenesis, but several studies have observed differential expression of neuronal receptors in cancer. Recently a group observed promoter hypermethylation of GRIK2 gene in gastric cancer and they observed that rescue of GRIK2 expression in gastric cancer cells decreased cell migration (Wu et al., 2010a). This may be an evidence for the role of neuronal receptors and genes in cellular processes involved in tumorgenesis. We found in this study other genes commonly known to be involved in neural development: DMRTA2, ZIC4, RAX and EN1. DMRTA2 (Doublesex- and mab-3-related transcription factor 5) is a transcription factor involved in neuronal development in zebrafish (Yoshizawa et al., 2011) and involved in specification of ventral-medial neural progenitor cell fate in mammalian embryonic stem cells. We have found no studies reporting DMRTA2 gene or its methylation in cancer. We found that DMRTA is methylated in 72% of the tumors and in only 2 of the 61 lung samples. Considering the percentages of DNA methylation, it presented the highest individual tumor sensitivity/specificity of all fragments tested. ZIC4 is another poorly characterized gene, which involvement in central nervous system has been studied. ZIC4 is a member of the zic finger proteins family. Genetic alterations in this gene cause congenital cerebellar malformation (Grinberg et al., 2004). It is assumed that this gene plays an important role during embryonic development. A recent report has shown that ZIC4 and other members of the same family can suppress β-catenin-mediated transcriptional activation. β-catenin is a key regulator in early development, part of the Wnt pathway, which is well-known to be abnormally activated in cancer. Promoter hypermethylation of ZIC1, another member of this family, was observed in gastric and colorectal cancer (Gan et al., 2011; Wang et al., 2009), we have not found a previous report investigating methylation of ZIC4 in NSCLC or other types of cancer. ZIC4 methylation status was investigated in 57 lung samples and paired tumors, and we found no methylation in the normal lung tissues while 70% of the tumors were methylated for this gene. For RAX, we found methylation in 82% of the tumor samples while only 1 of 60 lung samples was methylated. RAX is a homeobox-containing transcription factor gene, suggested to be a member of the RNA-induced silencing complex. It is also thought to interact with double-stranded RNA-dependent protein kinase (PKR) and activate it. RAX/PKR is implicated to be involved in response to cellular stress where activation of PKR induces eIF2alpha phosphorylation, leading to apoptosis (Bennett et al., 2006). In a recent report, RAX has been found methylated in human astrocitomas (Wu et al., 2010b).

Frag01 and Frag02 are genomic regions located on chromosome 2 q14.2. We have found that a large region containing this genomic sector is highly differentially methylated in NSCLC tumors. This region span more then 20 kb under which the EN1 gene is located. In the work described in chapter 2 we studied three fragments in this region and they were significant to differentiate lung from their paired tumor tissue by

MSP and bisulfite sequencing; besides Frag01 were also significant to differentiate between ADC and SCC. In this study, Frag01 is still highly specific with none of the lung samples presenting methylation for this gene and about 65% of the tumors being methylated. However in this study we do not repeat the same finding for histological subtypes. Here Frag01 did not show a significant rate probably due to the small sample numbers where only 14 of the amplified tumors for this fragment belonged to the ADC subtype, substantially decreasing the power of statistical analysis. This genomic region has been previously observed as hypermethylated in tumors. This finding was first detected in 2004 and published in a Chinese journal (Liu et al., 2004). Later other groups observed the same in prostate (Devaney et al., 2011) and colorectal cancer where they implicated it in long-range epigenetic silencing and suggested it as non-invasive biomarker for this type of tumor (Mayor et al., 2009).

A study in colorectal cancer has just been published where they find promoter hypermethylation of BARX1 in 56% of the patients (Kober et al., 2011). In our study this gene is methylated in 82% of the tumors and in only 2 of the lung tissues. BARX1 is another homeobox transcription regulator implicated to indirectly suppress Wnt cascade. BARX1 plays its role in Wnt pathway by activating Sfrp1 and Sfrp2 which are Wnt inhibitors (Kim et al., 2005). The Wnt pathway is activated in different types of cancer, and it is involved in abnormal cell growth.

GATA3 is a member of the GATA transcription factor family, known to play critical role in hematopoietic development. The most described role of this gene involves regulation of T-cell development. However GATA3 has also been reported to play important function in epithelial formation (Kaufman et al., 2003; Kouros-Mehr et al., 2006), and as a positively regulator of cyclin D1 expression in neuroblastoma cells, participating in the pathway responsible for maintenance of undifferentiated cell state (Molenaar et al., 2010). The loss of GATA3 expression is associated with poor prognosis in breast cancer (Yoon et al., 2010).

TBX15 is another transcription factor gene. It belongs to the T-box family of genes and encodes for one of the conserved family of proteins which contain a T-box DNA-binding domain. Genes from this family regulate a variety of developmental processes. Mutations in TBX15 were found to associate with Cousin syndrome (Lausch et al., 2008), and a recent report showed that differential methylation of a distal promoter of TBX15 associates with pathological placentas (Chelbi et al., 2011). We observe TBX15 methylation in 95% of the tumors, with AUC ROC curve of 0.922, indicating high specificity and sensitivity for methylation in tumors.

FGF12 is a member of the fibroblast growth factor family. Members of this family have been reported to play a role in nuclear localization signaling pathways, and implicated to be involved in a diversity of cellular processes including the ones related to tumorgenesis, such as cell growth and invasion. However not much is yet known about the role of this specific gene either in normal cellular function or tumorgenesis.

FGF12 hypermethylation was observed in Myeloid Leukemia cell line (Gebhard et al., 2006) and it has been suggested as a candidate marker for SCC, where the region containing this gene showed different levels of methylation in SCCs and ADCs (Kang et al., 2009).

LIMK1 was selected as a SCC DMR. However it was more efficient to discriminate between tumor and normal samples with 50% of the tumors and only 1.7% of lung samples being methylated for this gene. LIMK1 is a serine/threonine kinase member of the LIM family of proteins, known for their cysteine-rich structures with 2 zinc finger domains. LIMK1 is involved in cytoskeletal organization and cellular filament dynamics by regulating actin polymerization via its binding factor cofilin (Gorovoy et al., 2005). LIMK1 has been reported to be a downstream target of another cyclin kinase inhibitor protein: CDKN1C (Vlachos and Joseph, 2009). Although there are some studies which investigate the role of LIMK1 expression in tumor progression and invasion (Davila et al., 2007; McConnell et al., 2011; Vlecken and Bagowski, 2009). One report have have directly investigated LIMK1 methylation status finding higher levels of methylated LIMK1 in micro-dissected invasive and intraductal carcinoma in breast cancer (Dietrich et al., 2009).

ONECUT2 is a transcription factor member gene containing a homeodomain. Its most studied role is in hepatocyte differentiation, though its mechanism is not well characterized, and involvement of ONECUT2 in tumorgenesis is unknown. ONECUT2 has been reported as hypermethylated in B-cell lymphoma, where it was found methylated in 85% of the tumors (Pike et al., 2008). In accordance with this previous report, we found methylation of this gene in 83% of our tumors. Another study investigating methylation in lung cancer cells lines found ONECUT2, among several other homeodomain genes, as abnormally methylated (Rauch et al., 2006).

In contrast to the tumor DMRs, the selected DMRs for histological subtype did not present the same performance to discriminate tumor subtype. We found that 2 of the ADC DMRs (GAS1 and FAM78B), and one of the SCC DMRs (DOTL1) were significant to differentiate between the two histological subtypes. No panel could be build using those genes that would give a good prediction rate. Although those genes failed as histological markers, based on our genome-wide methylation profiles we believe it may be possible to define a panel of methylated genes specific for the two larger NSCLC subtypes (ADC and SCC). To build a panel, identification of more DMRs that can significantly differentiate between the two subtypes will be needed. The subtype DMRs tested may have failed due to the small cohort. We had 27 SCC and 20 ADC, from which there were tumors from both subtypes that were not methylated or failed to amplify with the methylated primer. This reduced the sample size and considerably decreased the statistical power for those DMRs. Therefore a larger cohort may be needed to produce more powerful statistical results. Moreover, testing a larger number of DMRs may be needed to find the ones with the highest significance. In this study we have tested five DMRs for each subtype. However from those profiles we

produce a comprehensive list with DMRs which were specific for each subtype, where 151 genes and 93 genomic regions are specific for SCC, and 18 genes and 6 genomic regions for ADC. Among them may be present genes that play specific role in the make up of each histological subtye. Testing more DMRs in a larger cohort will be needed to find the most suitable markers, and consequently building a panel that may enable correct subtype prediction.

As we observed this divergence between standard MSP analysis and assessment of amplification levels in their potential to discriminate tumors, we conclude that determination of methylation status of tumors versus normal tissues is not an all or non measurement. Determining the levels of DNA methylation is important to properly reflect the potential of markers to correctly differentiate between normal lung and tumors, which will be problematic when dealing with biopsies which are heterogeneous in nature. Nonetheless, our results indicate that a panel of markers might correctly discriminate biopsies containing tumor cells from normal lung.

In this study we have shown novel methylated genes in NSCLC that are significant to distinguish between tumor and lung samples using either MSP counts or DNA methylation percentages, presenting also high specificity and sensitivity. These markers outperform those previously reported for NSCLC. Furthermore we present a panel of 5 methylated genes that have together 94% sensitivity and 96% specificity, which could correctly predict tumor tissue in >90% of the cases.

Methods

Patient samples and Histopathological analysis

Samples were obtained from 62 patients diagnosed with NSCLC who had undergone surgical lung resection at Erasmus MC. Tissues from the tumor and adjacent non-cancerous lung were collected and snap-frozen in liquid nitrogen pre-cooled isopentane within two hours after surgical resection. Samples were stored at -196 °C or -80°C until DNA extraction. Patient samples were independently reviewed by two pathologists. The cohort included 20 adenocarcinomas (ADC); 27 squamous cell carcinomas (SCC), 10 large cell carcinomas (LCC) and 4 unclassified samples, and their paired lung tissues. Specimens were collected and studied under an anonymous tissue protocol approved by the medical ethical committee

Cell line

The MRC-5 lung fibroblast-like cell line was used. Cells were cultured under standard conditions. In short: minimum essential medium supplemented with 10% heat-inactivated fetal bovine serum, 2mM L-glutamine, 1% non-essential amino acids and penicillin/streptomycin; incubated at 37 ˚C and 5% CO_2. Cells were harvested when they reached 90% confluence.

DNA isolation

Genomic DNA from patient tissues and cultured cells were extracted by overnight incubation with lysis buffer and proteinase K, followed by phenol-chloroform extraction, ethanol precipitation and RNase digestion.

Artificial demethylation and methylation of genomic DNA

We used DNA extracted from the MRC-5 cell line and commercially available Universal unmethylated DNA (UUD; Millipore). Fully unmethylated DNA was obtained by whole-genome amplification using the REPLI-g kit (Qiagen) according to manufacturer protocol, followed by phenol-chloroform extraction. Fully methylated DNA was prepared by treating MRC-5 DNA and UUD with M.SssI enzyme (New England Biolabs) according to the manufacturer's protocol. In short, 10 µg of DNA was incubated for 2 hours at 37 ˚C with 40U of M.SssI and 640 µM of S-adenosylmethionine. DNA was then treated with phenol-chloroform and recovered by ethanol precipitation.

Methylation-specific PCR (MSP)

We used MSP (Herman et al., 1996) to determine the methylation status of the 25 selected regions in 122 samples (61 tumors and matching lung tissues). Samples were bisulfite converted using the Epitect kit from Qiagen following the manufacturer's protocol. Samples were amplified using two different set of primers designed for the methylated and unmethylated sequences (primers and amplification conditions are specified in supplementary table 4). The fully methylated and fully unmethylated DNA samples were used as controls, and a water blank reaction used as control for contamination. After amplification, products were resolved on 2% agarose gels containing ethidium bromide.

Band counts and intensities

Gels were scanned using Typhoon Trio Scanner or Typhoon 9410 from GE Health care, with settings as follow: filter in 610 BP 30; green (532 nm); PMT varying between 580 V and 610 V; focal plane in +3 mm; pixel size: 200. All samples from the same fragment were scanned using the same setting conditions. Amplification intensities of the bands were calculated using analysis tool box of ImageQuantTL software. Volume report for each band was obtained by drawing a rectangle around the band and using the software for calculating pixels intensities in each rectangle (I). Background was obtained manually by drawing half of the rectangle above (bg1) and half below each band (bg2). We obtained the intensity of each band by extracting the band volume from the background volume (Final Intensity). Values for all DMRs in all samples can be found in supplementary file 2)

Normalization of intensities

We used the fully methylated and fully unmethylated control samples for determination of maximum and minimum intensity in each fragment. Amplification was done in triplicate. Normalization was calculated by the amplification ratio between the fully methylated and fully unmethylated samples intensities. We also generated a correction coefficient for the primers. For this, we amplified dilution mixes of the two control samples (100/0 – 50/50 – 0/100) with each primer for each fragment. Coefficient was calculated by the ratio between methylated primer and unmethylated primer in each point of the dilution mix. Intensities of the bands were adjusted for each fragment and corrected for difference in primer efficiency.

Statistical analysis

We used PASW 18.0 statistical software for analysis. We applied Fisher exact test considering a 95% confidence interval (CI)and two sided p-value (p[O>=E|O<=E]) for calculating the significance in methylation status by band counts in each DMR in the group of samples lung vs tumor or SCC/ADC. For MSP intensities and percentages we used paired t-test (CI 95%) for significance between lung and tumor in each DMR, and Mann-Whitney for SCC versus ADC DMRs. ANOVA was used for comparison between lung and tumor in pooled DMRs. Logistic regression with 95% CI was used for sensitivity and specificity on estimating the best panel of markers.

References

Bah, J., Quach, H., Ebstein, R.P., Segman, R.H., Melke, J., Jamain, S., Rietschel, M., Modai, I., Kanas, K., Karni, O.*, et al.* (2004). Maternal transmission disequilibrium of the glutamate receptor GRIK2 in schizophrenia. Neuroreport *15*, 1987-1991.

Belinsky, S.A. (2004). Gene-promoter hypermethylation as a biomarker in lung cancer. Nat Rev Cancer *4*, 707-717.

Belinsky, S.A., Klinge, D.M., Stidley, C.A., Issa, J.P., Herman, J.G., March, T.H., and Baylin, S.B. (2003). Inhibition of DNA methylation and histone deacetylation prevents murine lung cancer. Cancer Res *63*, 7089-7093.

Belinsky, S.A., Nikula, K.J., Palmisano, W.A., Michels, R., Saccomanno, G., Gabrielson, E., Baylin, S.B., and Herman, J.G. (1998). Aberrant methylation of p16(INK4a) is an early event in lung cancer and a potential biomarker for early diagnosis. Proc Natl Acad Sci U S A *95*, 11891-11896.

Bennett, R.L., Blalock, W.L., Abtahi, D.M., Pan, Y., Moyer, S.A., and May, W.S. (2006). RAX, the PKR activator, sensitizes cells to inflammatory cytokines, serum withdrawal, chemotherapy, and viral infection. Blood *108*, 821-829.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. Genes Dev *16*, 6-21.

Bird, A.P. (1986). CpG-rich islands and the function of DNA methylation. Nature *321*, 209-213.

Calmon, M.F., Colombo, J., Carvalho, F., Souza, F.P., Filho, J.F., Fukuyama, E.E., Camargo, A.A., Caballero, O.L., Tajara, E.H., Cordeiro, J.A.*, et al.* (2007). Methylation profile of genes CDKN2A (p14 and p16), DAPK1, CDH1, and ADAM23 in head and neck cancer. Cancer Genet Cytogenet *173*, 31-37.

Chelbi, S.T., Doridot, L., Mondon, F., Dussour, C., Rebourcet, R., Busato, F., Gascoin-Lachambre, G., Barbaux, S., Rigourd, V., Mignot, T.M.*, et al.* (2011). Combination of promoter hypomethylation and PDX1 overexpression leads to TBX15 decrease in vascular IUGR placentas. Epigenetics *6*, 247-255.

Daniunaite, K., Berezniakovas, A., Jankevicius, F., Laurinavicius, A., Lazutka, J.R., and Jarmalaite, S. (2011). Frequent methylation of RASSF1 and RARB in urine sediments from patients with early stage prostate cancer. Medicina (Kaunas) *47*, 147-153.

Davila, M., Jhala, D., Ghosh, D., Grizzle, W.E., and Chakrabarti, R. (2007). Expression of LIM kinase 1 is associated with reversible G1/S phase arrest, chromosomal instability and prostate cancer. Mol Cancer *6*, 40.

Delorme, R., Krebs, M.O., Chabane, N., Roy, I., Millet, B., Mouren-Simeoni, M.C., Maier, W., Bourgeron, T., and Leboyer, M. (2004). Frequency and transmission of glutamate receptors GRIK2 and GRIK3 polymorphisms in patients with obsessive compulsive disorder. Neuroreport *15*, 699-702.

Devaney, J., Stirzaker, C., Qu, W., Song, J.Z., Statham, A.L., Patterson, K.I., Horvath, L.G., Tabor, B., Coolen, M.W., Hulf, T.*, et al.* (2011). Epigenetic deregulation across chromosome 2q14.2 differentiates normal from prostate cancer and provides a regional panel of novel DNA methylation cancer biomarkers. Cancer Epidemiol Biomarkers Prev *20*, 148-159.

Dietrich, D., Lesche, R., Tetzner, R., Krispin, M., Dietrich, J., Haedicke, W., Schuster, M., and Kristiansen, G. (2009). Analysis of DNA methylation of multiple genes in microdissected cells from formalin-fixed and paraffin-embedded tissues. J Histochem Cytochem *57*, 477-489.

Domagala, P., Huzarski, T., Lubinski, J., Gugala, K., and Domagala, W. (2011). PARP-1 expression in breast cancer including BRCA1-associated, triple negative and basal-like tumors: possible implications for PARP-1 inhibitor therapy. Breast Cancer Res Treat *127*, 861-869.

Eads, C.A., Danenberg, K.D., Kawakami, K., Saltz, L.B., Blake, C., Shibata, D., Danenberg, P.V., and Laird, P.W. (2000). MethyLight: a high-throughput assay to measure DNA methylation. Nucleic Acids Res *28*, E32.

Esteller, M. (2007). Cancer epigenomics: DNA methylomes and histone-modification maps. Nat Rev Genet *8*, 286-298.

Ferlay, J., Shin, H.R., Bray, F., Forman, D., Mathers, C., and Parkin, D.M. (2010). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int J Cancer *127*, 2893-2917.

Gan, L., Chen, S., Zhong, J., Wang, X., Lam, E.K., Liu, X., Zhang, J., Zhou, T., Yu, J., Si, J.*, et al.* (2011). ZIC1 is downregulated through promoter hypermethylation, and functions as a tumor suppressor gene in colorectal cancer. PLoS One *6*, e16916.

Gangopadhyay, N.N., Luketich, J.D., Opest, A., Visus, C., Meyer, E.M., Landreneau, R., and Schuchert, M.J. (2011). Inhibition of poly(ADP-ribose) polymerase (PARP) induces apoptosis in lung cancer cell lines. Cancer Invest *29*, 608-616.

Gao, A., Zuo, X., Liu, Q., Lu, X., Guo, W., and Tian, L. (2010). Methylation of PARP-1 promoter involved in the regulation of benzene-induced decrease of PARP-1 mRNA expression. Toxicol Lett *195*, 114-118.

Gaymes, T.J., Shall, S., MacPherson, L.J., Twine, N.A., Lea, N.C., Farzaneh, F., and Mufti, G.J. (2009). Inhibitors of poly ADP-ribose polymerase (PARP) induce apoptosis of myeloid leukemic cells: potential for therapy of myeloid leukemia and myelodysplastic syndromes. Haematologica *94*, 638-646.

Gebhard, C., Schwarzfischer, L., Pham, T.H., Schilling, E., Klug, M., Andreesen, R., and Rehli, M. (2006). Genome-wide profiling of CpG methylation identifies novel targets of aberrant hypermethylation in myeloid leukemia. Cancer Res *66*, 6118-6128.

Gorovoy, M., Niu, J., Bernard, O., Profirovic, J., Minshall, R., Neamu, R., and Voyno-Yasenetskaya, T. (2005). LIM kinase 1 coordinates microtubule stability and actin polymerization in human endothelial cells. J Biol Chem *280*, 26533-26542.

Grinberg, I., Northrup, H., Ardinger, H., Prasad, C., Dobyns, W.B., and Millen, K.J. (2004). Heterozygous deletion of the linked genes ZIC1 and ZIC4 is involved in Dandy-Walker malformation. Nat Genet *36*, 1053-1055.

Grote, D., Boualia, S.K., Souabni, A., Merkel, C., Chi, X., Costantini, F., Carroll, T., and Bouchard, M. (2008). Gata3 acts downstream of beta-catenin signaling to prevent ectopic metanephric kidney induction. PLoS Genet *4*, e1000316.

Heller, G., Zielinski, C.C., and Zochbauer-Muller, S. (2010). Lung cancer: from single-gene methylation to methylome profiling. Cancer Metastasis Rev *29*, 95-107.

Herman, J.G., Graff, J.R., Myohanen, S., Nelkin, B.D., and Baylin, S.B. (1996). Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. Proc Natl Acad Sci U S A *93*, 9821-9826.

Huang, L.W., Pan, H.S., Lin, Y.H., Seow, K.M., Chen, H.J., and Hwang, J.L. (2011). P16 methylation is an early event in cervical carcinogenesis. Int J Gynecol Cancer *21*, 452-456.

Javle, M., and Curtin, N.J. (2011). The role of PARP in DNA repair and its therapeutic exploitation. Br J Cancer *105*, 1114-1122.

Jeong, D.H., Youm, M.Y., Kim, Y.N., Lee, K.B., Sung, M.S., Yoon, H.K., and Kim, K.T. (2006). Promoter methylation of p16, DAPK, CDH1, and TIMP-3 genes in cervical cancer: correlation with clinicopathologic characteristics. Int J Gynecol Cancer *16*, 1234-1240.

Kang, J.U., Koo, S.H., Kwon, K.C., Park, J.W., and Kim, J.M. (2009). Identification of novel candidate target genes, including EPHB3, MASP1 and SST at 3q26.2-q29 in squamous cell carcinoma of the lung. BMC Cancer *9*, 237.

Kaufman, C.K., Zhou, P., Pasolli, H.A., Rendl, M., Bolotin, D., Lim, K.C., Dai, X., Alegre, M.L., and Fuchs, E. (2003). GATA-3: an unexpected regulator of cell lineage determination in skin. Genes Dev *17*, 2108-2122.

Kim, B.M., Buchner, G., Miletich, I., Sharpe, P.T., and Shivdasani, R.A. (2005). The stomach mesenchymal transcription factor Barx1 specifies gastric epithelial identity through inhibition of transient Wnt signaling. Dev Cell *8*, 611-622.

Kim, D.H., Kim, J.S., Ji, Y.I., Shim, Y.M., Kim, H., Han, J., and Park, J. (2003). Hypermethylation of RASSF1A promoter is associated with the age at starting smoking and a poor prognosis in primary non-small cell lung cancer. Cancer Res *63*, 3743-3746.

Kober, P., Bujko, M., Oledzki, J., Tysarowski, A., and Siedlecki, J.A. (2011). Methyl-CpG binding column-based identification of nine genes hypermethylated in colorectal cancer. Mol Carcinog *50*, 846-856.

Kouros-Mehr, H., Slorach, E.M., Sternlicht, M.D., and Werb, Z. (2006). GATA-3 maintains the differentiation of the luminal cell fate in the mammary gland. Cell *127*, 1041-1055.

Lausch, E., Hermanns, P., Farin, H.F., Alanay, Y., Unger, S., Nikkel, S., Steinwender, C., Scherer, G., Spranger, J., Zabel, B.*, et al.* (2008). TBX15 mutations cause craniofacial dysmorphism, hypoplasia of scapula and pelvis, and short stature in Cousin syndrome. Am J Hum Genet *83*, 649-655.

Le Page, F., Schreiber, V., Dherin, C., De Murcia, G., and Boiteux, S. (2003). Poly(ADP-ribose) polymerase-1 (PARP-1) is required in murine cell lines for base excision repair of oxidative DNA damage in the absence of DNA polymerase beta. J Biol Chem *278*, 18471-18477.

Li, E. (2002). Chromatin modification and epigenetic reprogramming in mammalian development. Nat Rev Genet *3*, 662-673.

Liu, J.Y., An, Q., Zhang, J.J., Lei, W.D., Cheng, S.J., and Gao, Y.N. (2004). [Screening of hypermethylated DNA fragments in tumor tissue derived from patients with lung cancer]. Yi Chuan Xue Bao *31*, 389-394.

Mayor, R., Casadome, L., Azuara, D., Moreno, V., Clark, S.J., Capella, G., and Peinado, M.A. (2009). Long-range epigenetic silencing at 2q14.2 affects most human colorectal cancers and may have application as a non-invasive biomarker of disease. Br J Cancer *100*, 1534-1539.

McConnell, B.V., Koto, K., and Gutierrez-Hartmann, A. (2011). Nuclear and cytoplasmic LIMK1 enhances human breast cancer progression. Mol Cancer *10*, 75.

Miura, A., Yonebayashi, S., Watanabe, K., Toyama, T., Shimada, H., and Kakutani, T. (2001). Mobilization of transposons by a mutation abolishing full DNA methylation in Arabidopsis. Nature *411*, 212-214.

Molenaar, J.J., Ebus, M.E., Koster, J., Santo, E., Geerts, D., Versteeg, R., and Caron, H.N. (2010). Cyclin D1 is a direct transcriptional target of GATA3 in neuroblastoma tumor cells. Oncogene *29*, 2739-2745.

Moroni, F., Formentini, L., Gerace, E., Camaioni, E., Pellegrini-Giampietro, D.E., Chiarugi, A., and Pellicciari, R. (2009). Selective PARP-2 inhibitors increase apoptosis in hippocampal slices but protect cortical cells in models of post-ischaemic brain damage. Br J Pharmacol *157*, 854-862.

Motazacker, M.M., Rost, B.R., Hucho, T., Garshasbi, M., Kahrizi, K., Ullmann, R., Abedini, S.S., Nieh, S.E., Amini, S.H., Goswami, C.*, et al.* (2007). A defect in the ionotropic glutamate receptor 6 gene (GRIK2) is associated with autosomal recessive mental retardation. Am J Hum Genet *81*, 792-798.

Niklinska, W., Naumnik, W., Sulewska, A., Kozlowski, M., Pankiewicz, W., and Milewski, R. (2009). Prognostic significance of DAPK and RASSF1A promoter hypermethylation in non-small cell lung cancer (NSCLC). Folia Histochem Cytobiol *47*, 275-280.

Palmisano, W.A., Divine, K.K., Saccomanno, G., Gilliland, F.D., Baylin, S.B., Herman, J.G., and Belinsky, S.A. (2000). Predicting lung cancer by detecting aberrant promoter methylation in sputum. Cancer Res *60*, 5954-5958.

Pfeifer, G.P., Yoon, J.H., Liu, L., Tommasi, S., Wilczynski, S.P., and Dammann, R. (2002). Methylation of the RASSF1A gene in human cancers. Biol Chem *383*, 907-914.

Pike, B.L., Greiner, T.C., Wang, X., Weisenburger, D.D., Hsu, Y.H., Renaud, G., Wolfsberg, T.G., Kim, M., Weisenberger, D.J., Siegmund, K.D., *et al.* (2008). DNA methylation profiles in diffuse large B-cell lymphoma and their relationship to gene expression status. Leukemia *22*, 1035-1043.

Rauch, T., Li, H., Wu, X., and Pfeifer, G.P. (2006). MIRA-assisted microarray analysis, a new technology for the determination of DNA methylation patterns, identifies frequent methylation of homeodomain-containing genes in lung cancer cells. Cancer Res *66*, 7939-7947.

Richiardi, L., Vizzini, L., Merletti, F., and Barone-Adesi, F. (2009). Cardiovascular benefits of smoking regulations: The effect of decreased exposure to passive smoking. Prev Med *48*, 167-172.

Sharma, S., Kelly, T.K., and Jones, P.A. (2010). Epigenetics in cancer. Carcinogenesis *31*, 27-36.

Smulson, M.E., Simbulan-Rosenthal, C.M., Boulares, A.H., Yakovlev, A., Stoica, B., Iyer, S., Luo, R., Haddad, B., Wang, Z.Q., Pang, T., *et al.* (2000). Roles of poly(ADP-ribosyl)ation and PARP in apoptosis, DNA repair, genomic stability and functions of p53 and E2F-1. Adv Enzyme Regul *40*, 183-215.

Sterling, J.A., Wu, L., and Banerji, S.S. (2006). PARP regulates TGF-beta receptor type II expression in estrogen receptor-positive breast cancer cell lines. Anticancer Res *26*, 1893-1901.

Suzuki, M.M., and Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet *9*, 465-476.

Takebayashi, S., Tamura, T., Matsuoka, C., and Okano, M. (2007). Major and essential role for the DNA methylation mark in mouse embryogenesis and stable association of DNMT1 with newly replicated regions. Mol Cell Biol *27*, 8243-8258.

Toyooka, S., Toyooka, K.O., Maruyama, R., Virmani, A.K., Girard, L., Miyajima, K., Harada, K., Ariyoshi, Y., Takahashi, T., Sugio, K., *et al.* (2001). DNA methylation profiles of lung tumors. Mol Cancer Ther *1*, 61-67.

Travis, W.D., Brambilla, E., Noguchi, M., Nicholson, A.G., Geisinger, K.R., Yatabe, Y., Beer, D.G., Powell, C.A., Riely, G.J., Van Schil, P.E., *et al.* (2011). International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. J Thorac Oncol *6*, 244-285.

Travis, W.D., World Health Organization., International Agency for Research on Cancer., International Association for the Study of Lung Cancer., and International Academy of Pathology. (2004). Pathology and genetics of tumours of the lung, pleura, thymus and heart (Lyon

Oxford, IARC Press

Oxford University Press (distributor)).

Usadel, H., Brabender, J., Danenberg, K.D., Jeronimo, C., Harden, S., Engles, J., Danenberg, P.V., Yang, S., and Sidransky, D. (2002). Quantitative adenomatous polyposis coli promoter methylation analysis in tumor tissue, serum, and plasma DNA of patients with lung cancer. Cancer Res *62*, 371-375.

Verdoodt, B., Sommerer, F., Palisaar, R.J., Noldus, J., Vogt, M., Nambiar, S., Tannapfel, A., Mirmohammadsadegh, A., and Neid, M. (2011). Inverse association of p16(INK4a) and p14(ARF) methylation of the CDKN2a locus in different Gleason scores of prostate cancer. Prostate Cancer Prostatic Dis *14*, 295-301.

Vlachos, P., and Joseph, B. (2009). The Cdk inhibitor p57(Kip2) controls LIM-kinase 1 activity and regulates actin cytoskeleton dynamics. Oncogene *28*, 4175-4188.

Vlecken, D.H., and Bagowski, C.P. (2009). LIMK1 and LIMK2 are important for metastatic behavior and tumor cell-induced angiogenesis of pancreatic cancer cells. Zebrafish *6*, 433-439.

Wang, J., Wang, B., Chen, X., and Bi, J. (2011). The prognostic value of RASSF1A promoter hypermethylation in non-small cell lung carcinoma: a systematic review and meta-analysis. Carcinogenesis *32*, 411-416.

Wang, L.J., Jin, H.C., Wang, X., Lam, E.K., Zhang, J.B., Liu, X., Chan, F.K., Si, J.M., and Sung, J.J. (2009). ZIC1 is downregulated through promoter hypermethylation in gastric cancer. Biochem Biophys Res Commun *379*, 959-963.

Wu, C.S., Lu, Y.J., Li, H.P., Hsueh, C., Lu, C.Y., Leu, Y.W., Liu, H.P., Lin, K.H., Hui-Ming Huang, T., and Chang, Y.S. (2010a). Glutamate receptor, ionotropic, kainate 2 silencing by DNA hypermethylation possesses tumor suppressor function in gastric cancer. Int J Cancer *126*, 2542-2552.

Wu, X., Rauch, T.A., Zhong, X., Bennett, W.P., Latif, F., Krex, D., and Pfeifer, G.P. (2010b). CpG island hypermethylation in human astrocytomas. Cancer Res *70*, 2718-2727.

Yoon, N.K., Maresh, E.L., Shen, D., Elshimali, Y., Apple, S., Horvath, S., Mah, V., Bose, S., Chia, D., Chang, H.R., *et al.* (2010). Higher levels of GATA3 predict better survival in women with breast cancer. Hum Pathol *41*, 1794-1801.

Yoshizawa, A., Nakahara, Y., Izawa, T., Ishitani, T., Tsutsumi, M., Kuroiwa, A., Itoh, M., and Kikuchi, Y. (2011). Zebrafish Dmrta2 regulates neurogenesis in the telencephalon. Genes Cells *16*, 1097-1109.

Zeng, W., Gillis, T., Hakky, M., Djousse, L., Myers, R.H., MacDonald, M.E., and Gusella, J.F. (2006). Genetic analysis of the GRIK2 modifier effect in Huntington's disease. BMC Neurosci *7*, 62.

Zhang, Y., Wang, R., Song, H., Huang, G., Yi, J., Zheng, Y., Wang, J., and Chen, L. (2011). Methylation of multiple genes as a candidate biomarker in non-small cell lung cancer. Cancer Lett *303*, 21-28.

Zochbauer-Muller, S., Fong, K.M., Virmani, A.K., Geradts, J., Gazdar, A.F., and Minna, J.D. (2001). Aberrant promoter methylation of multiple genes in non-small cell lung cancers. Cancer Res *61*, 249-255.

Supplementary Information

Supplementary files not present here can be found at: https://docs.google.com
Login: thesis.hughes; password: thesishughes_2012

Supplementary Table 1: Primer sets used for MSP

primer name	primer sequence	bp	CpG	C	Product size	T (ºC)	cycles
M_BARX1-F	CGTATTAGGTATCGAGGGAATC	22	3	5	172	54	29
M_BARX1-R	TAAAAACGTAATACCAAAATCGAA	24	2	5			
U_BARX1-F	GTGTGTATTAGGTATTGAGGGAATT	25	3	5	172		
U_BARX1-R	TAAAAACATAATACCAAAATCAAAAAA	27	2	5			
M_PRDM14-F	CGACGTTTTCGAGGTTTC	18	4	2	194	54	29
M_PRDM14-R	AATACGATCTTTAAACGCTCGATA	24	3	6			
U_PRDM14-F	TTTTGATGTTTTTGAGGTTTT	21	4	2	194		
U_PRDM14-R	TCAAATACAATCTTTAAACACTCAATA	27	3	6			
M_RAX-F	TTAGTTTTTCGCGGTTGTATAC	22	3	7	147	56	28
M_RAX-R	CCGACGAAACGAAACTATC	19	3	3			
U_RAX-F	GGTTAGTTTTTTGTGGTTGTATAT	24	3	7	147		
U_RAX-R	CACCAACAAAACAAAACTATC	21	3	3			
M_DMRTA2-F	GATTTCGTAGGCGGGTTTC	19	3	4	155	58	28
M_DMRTA2-R	GCCAACGTATCATAACGACG	20	4	5			
U_DMRTA2-F	AAGATTTTGTAGGTGGGTTTTG	22	3	4	158		
U_DMRTA2-R	CACCAACATATCATAACAACACA	23	4	5			
M_ONECUT2-F	TGATTTTAAATGTACGGATTCGTTC	25	3	4	175	58	28
M_OECUT2-R	CGAAAATATATACTCGAACCCGC	23	3	5			
U_ONCUT2-F	ATTTTAAATGTATGGATTTGTTTGG	25	3	4	174		
U_ONCUT2-R	CCAAAAATATATACTCAAACCCACA	25	3	5			

continuation

M_PARP15-F	TAAATTCGAGGTTTTAATTTTTGACG	26	2	6	100	58	28
M_PARP15-R	AACTCGCTAAAACTACAATTACGCC	25	2	7			
U_PARP15-F	AATTTGAGGTTTTAATTTTTGATGG	25	2	5	100		
U_PARP15-R	TCAACTCACTAAAACTACAATTACACC	27	2	7			
M_GATA3-F	TGTTTAACGGGTAGTATTCGGATAC	25	3	7	129	56	28
M_GATA3-R	AAACGTAATTACCTTAACCGTCGAT	25	3	6			
U_GATA3-F	TTTAATGGGTAGTATTTGGATATGT	25	3	8	127		
U_GATA3-R	AAACATAATTACCTTAACCATCAAT	25	3	6			
M_GRIK2-F	5' ACGATGGAGGTACGGTTAGTC 3'	21	3	3	179	58	28
M_GRIK2-R	5' ACTAACGAAAACGAACAAACG 3'	21	3	7			
U_GRIK2-F	5' ATGATGATGGAGGTATGGTTAGTT 3'	24	3	3	179		
U_GRIK2-R	5' CACTAACAAAAACAAACAAACACA 3'	24	3	7			
M_TBX15-F	TAGTTTTTTTTGCGGCGAAAC	21	3	4	152	56	28
M_TBX15-R	AATTTAAACTCCGAATAAAATCCGAA	26	2	7			
U_TBX15-F	TAGTTTTTTTTGTGGTGAAATGG	23	3	4	151		
U_TBX15-R	ATTTAAACTCCAAATAAAATCCAAA	25	2	7			
M_ZIC4-F	GTGGATTTAGGCGGTATTTGTC	22	2	5	209	58	28
M_ZIC4-R	GTAACCCAAAACTTCCCTCG	20	2	4			
U_ZIC4-F	GGTGGATTTAGGTGGTATTTGTTGT	25	2	5	213		
U_ZIC4-R	ATCATAACCCAAAACTTCCCTCAC	24	2	5			
M_DOTL1-F	GTTTACGTTCGGAAATGAGTAGTTC	25	3	4	132	54	29
M_DOTL1-R	ACCATACGAAACTTAAAAAAAACGAT	26	2	9			
U_DOTL1-F	TTATGTTTGGAAATGAGTAGTTTGG	25	2	4	134		
U_DOTL1-R	AAATACCATACAAAACTTAAAAAAAACA	28	2	8			
M_LIMK1-F	TTTGCGAGGGTAGGATTTAAAGTC	24	2	5	168	58	29
M_LIMK1-R	AAAAACCGACAATTCACCGAA	21	2	8			
U_LIMK1-F	TTGTGAGGGTAGGATTTAAAGTTGG	25	2	5	170		
U_LIMK1-R	ACAAAAAACCAACAATTCACCAAA	24	2	9			
M_HIST1H3C-F	TAAAGTAGTTCGTAAGAGCGTTTCG	25	3	5	101	58	27
M_HIST1H3C-R	TCTAATAACGACGAATTTCGCG	22	4	4			
U_HIST1H3C-F	AAAGTAGTTTGTAAGAGTGTTTTGG	25	3	5	102		
U_HIST1H3C-R	CTTCTAATAACAACAAATTTCACACA	26	4	4			
M_ANKRD1B-F	TTTAGGGCGGAGTTAGGC	18	2	4	102	56	29
M_ANKRD1B-R	TTAATAAACAAAAATACGAACCGAC	25	2	6			
U_ANKRD1B-F	TTTTAGGGTGGAGTTAGGTGG	21	2	4	105		
U_ANKRD1B-R	AATTAATAAACAAAAATACAAACCAAC	27	2	8			

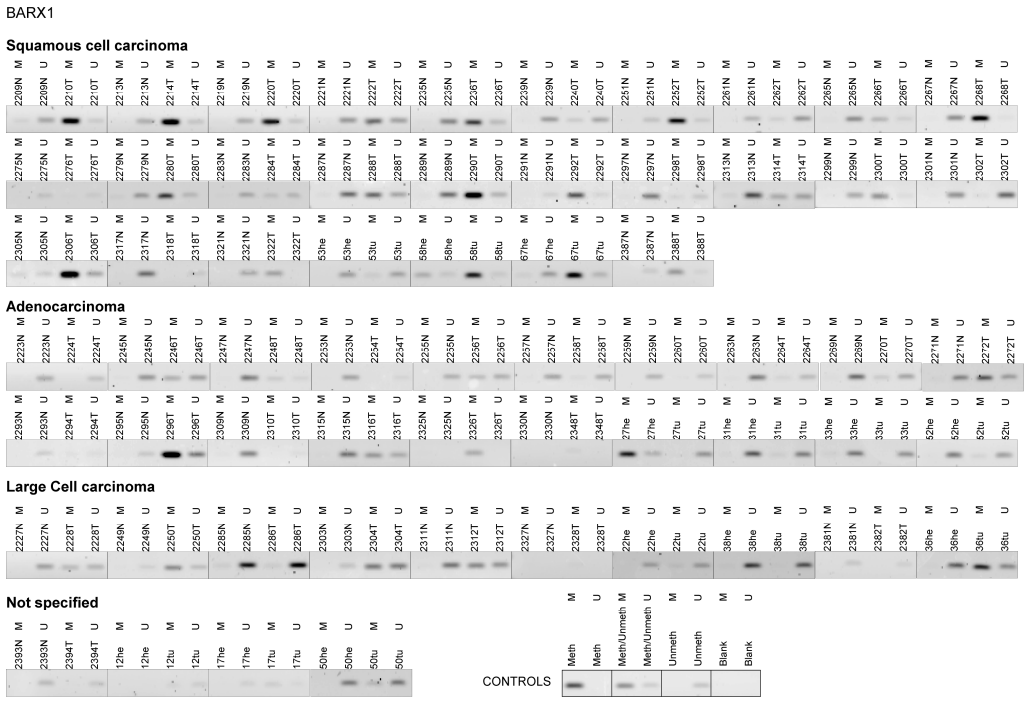
continuation

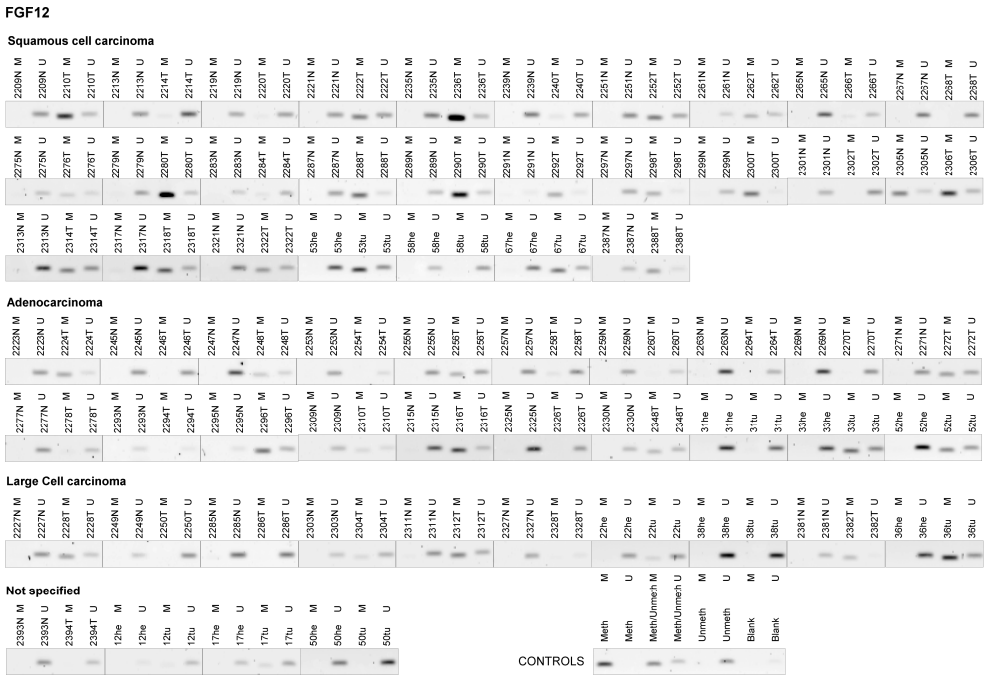
M_MSC-F	TTAAAAATTTAGGTCGGGAAGC	22	2	5	112	58	28
M_MSC-R	AAACTTAAACGCCAACTCGTA	21	2	5			
U_MSC-F	AGATTAAAAATTTAGGTTGGGAAGT	25	2	5	121		
U_MSC-R	AATTCTAAACTTAAACACCAACTCA	25	2	6			
M_GAS1-F	CGGTGCGTTGTTCGTTATC	19	3	3	112	58	29
M_GAS1-R	ATAACCCGCCTATACTTCGACG	22	4	5			
U_GAS1-F	TTTTGGTGTGTTGTTTGTTATTGT	24	3	6	116		
U_GAS1-R	CATAACCCACCTATACTTCAACACC	25	4	5			
M_HOXA1-F	GTTTTCGTAGTGATGGATTATC	22	2	5	100	56	28
M_HOXA1-R	AAATCGACCTTTACAATCG	19	2	3			
U_HOXA1-F	GGTTTTTGTAGTGATGGATTATT	23	2	5	100		
U_HOXA1-R	CAAATCAACCTTTACAATCA	20	2	3			
M_SEP9-F	GGTTATGGTGGCGGTGTC	18	2	3	125	58	29
M_SEP9-R	GAAAATCTAATCCCCGCGTA	20	3	3			
U_SEP9-F	ATGGTTATGGTGGTGGTGTTG	21	2	3	134		
U_SEP9-R	TAAAACCAAAAATCTAATCCCCACA	25	3	5			
M_FAM78B-F	TTGTGGTTTTAGACGGAAACGT	22	2	3	152	54	30
M_FAM78B-R	CCTTTAAAAAAACCGAAACGAA	22	2	7			
U_FAM78B-F	TTGTGGTTTTAGATGGAAATGT	22	2	3	152		
U_FAM78B-R	CCTTTAAAAAAACCAAAACAAA	22	2	7			
M_FGF12-F	TTCGTTTGTCGTTTTTGTCG	20	3	7	93	56	28
M_FGF12-R	TAAACGCCGAATTTTCGAT	19	3	4			
U_FGF12-F	ATTTTTTTGTTTGTTGTTTTTGTTG	25	3	10	93		
U_FGF12-R	ACCTTATAAACACCAAATTTTCAAT	25	3	5			
M_CDKN2A-F	AGCGTTTTAGGTTTATTCGGC	21	3	7	201	56	28
M_CDKN2A-R	GCAAACTAACTAACTCACTCCGC	23	2	5			
U_CDKN2A-F	TAAGTGTTTTAGGTTTATTTGGTGG	25	3	8	203		
U_CDKN2A-R	ACAAACTAACTAACTCACTCCACA	24	2	5			
M_RASSF1-F	AGCGAAGTACGGGTTTAATC	20	3	5	220	52	29
M_RASSF1-R	GCACCACGTATACGTAACG	19	3	4			
U_RASSF1-F	GTTAGTGAAGTATGGGTTTAATT	23	3	5	220		
U_RASSF1-R	ACACCACATATACATAACAAAC	22	3	4			
M_Frag01-F	GGAGGTTTTATTATTGTCGC	20	2	4	119	56	28
M_Frag01-R	AAACGAACGCAAACTCTC	18	2	4			
U_Frag01-F	ATTGGAGGTTTTATTATTGTTGT	23	2	4	119		
U_Frag01-R	AAACAAACACAAACTCTCAAC	21	2	4			

continuation

M_Frag02-F	GGGAACGAGTGTGTAGTTTC	20	2	3	192	56	28
M_Frag02-R	ACTATTCACGTAAACACGCA	20	2	2			
U_Frag02-F	GAGGGGAATGAGTGTGTAGTTTT	23	2	3	192		
U_Frag02-R	AAAACTATTCACATAAACACACA	23	2	2			
M_EN1-F	AGTTGTTATTATTTTTTCGTTTGTCG	26	2	8	214	54	29
M_EN1-R	CGTTTACTAACGCCTTTCTCG	21	3	2			
U_EN1-F	TTTAGTTGTTATTATTTTTTTGTTTGTT	28	2	10	219		
U_EN1-R	CTCATTTACTAACACCTTTCTCAAC	25	3	3			

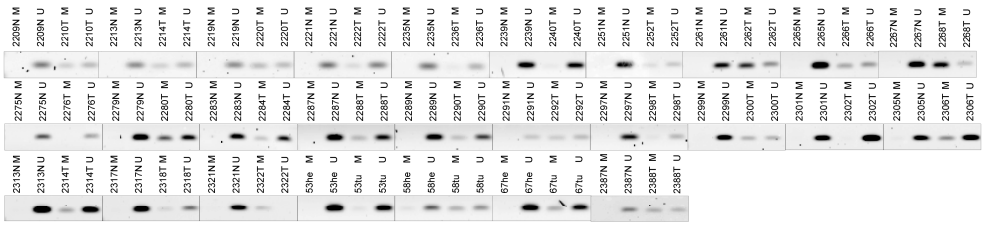
Supplementary Figure 1.1 to 1.25: Gels from all 25 DMRs investigated in this study




FGF12
Squamous cell carcinoma
Adenocarcinoma
Large Cell carcinoma
Not specified
CONTROLS
PRDM14
Squamous cell carcinoma
Adenocarcinoma
Large Cell carcinoma
Not specified
CONTROLS

RAX

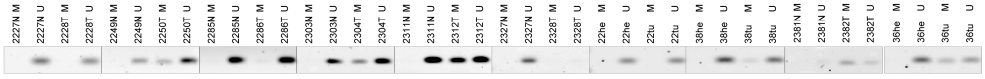
Squamous cell carcinoma



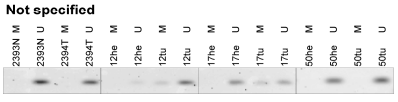
Adenocarcinoma

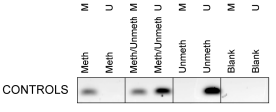


Large Cell carcinoma



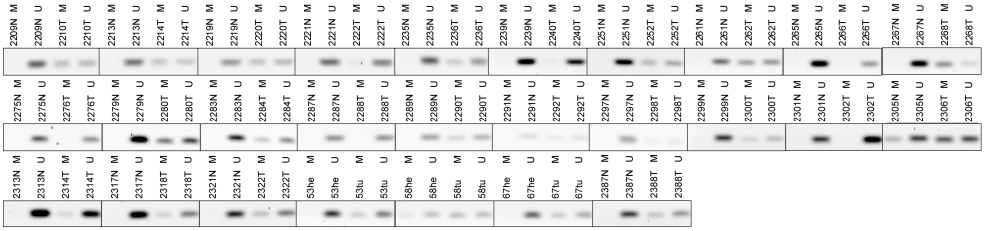
Not specified



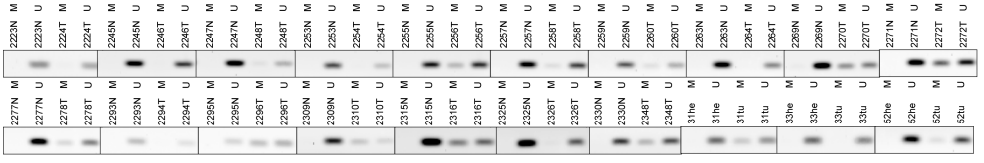


DMRTA2

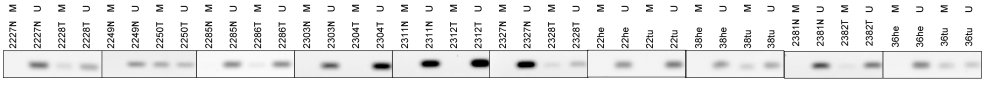
Squamous cell carcinoma



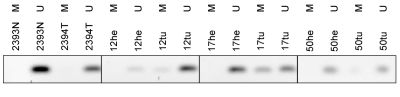
Adenocarcinoma

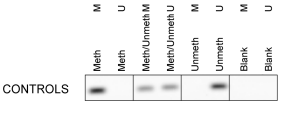


Large Cell carcinoma



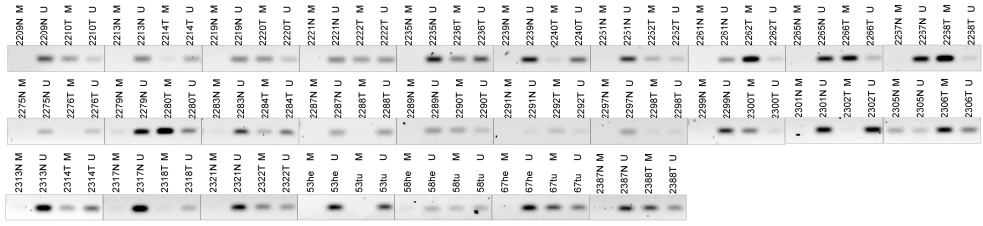
Not specified





ONECUT2

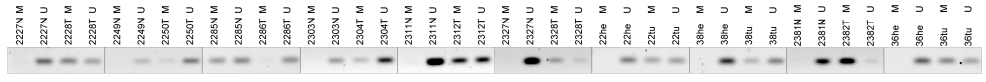
Squamous cell carcinoma



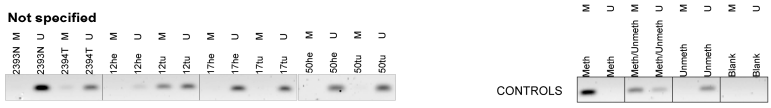
Adenocarcinoma



Large Cell carcinoma

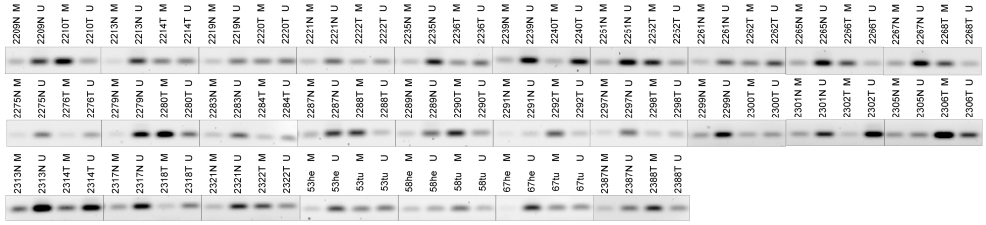


Not specified



PARP15

Squamous cell carcinoma



Adenocarcinoma

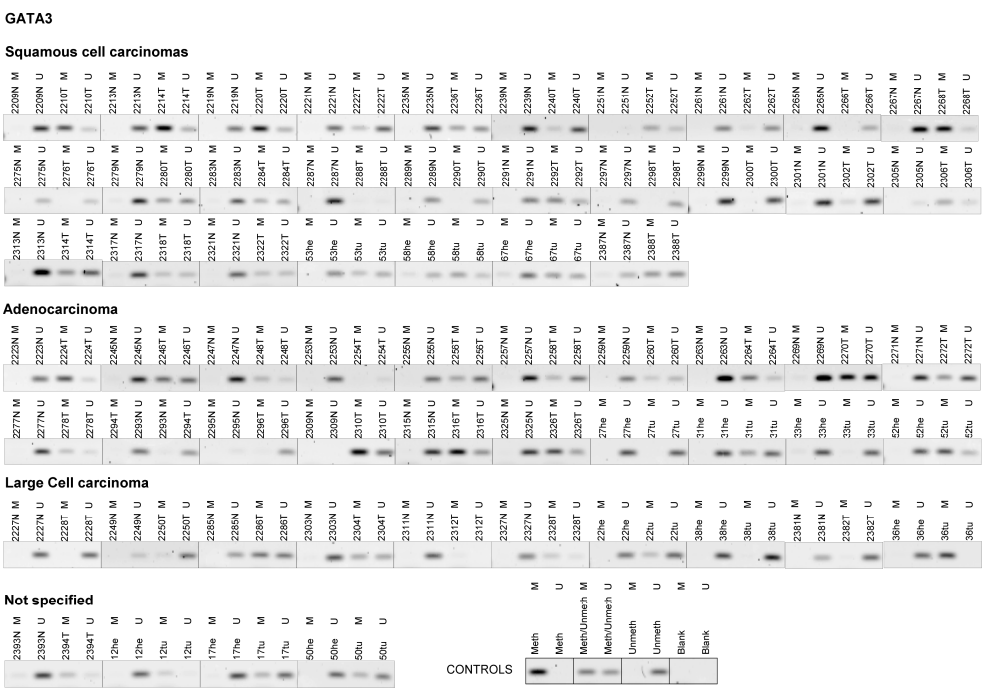


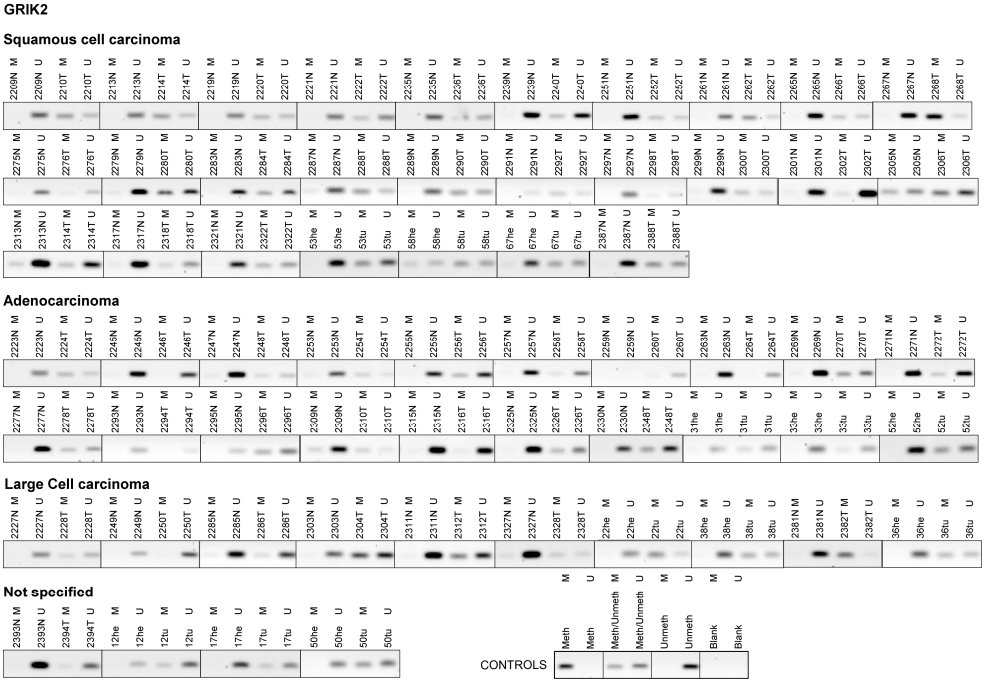
Large Cell carcinoma



Not specified

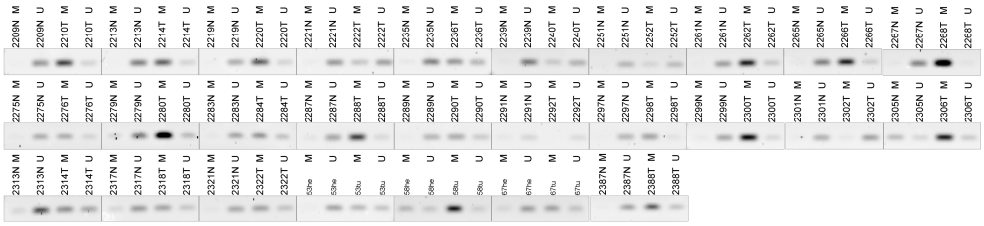



GATA3
Squamous cell carcinomas
Adenocarcinoma
Large Cell carcinoma
Not specified
CONTROLS
Meth
Meth
Meth/Unmeth
Meth/Unmeth
Unmeth
Unmeth
Blank
Blank

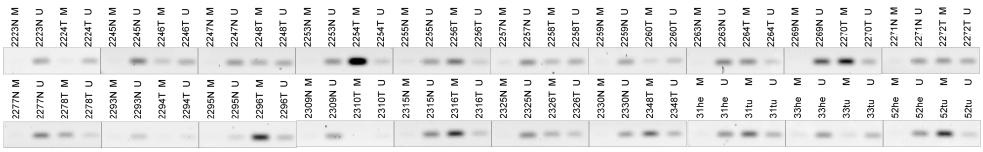

GRIK2
Squamous cell carcinoma
Adenocarcinoma
Large Cell carcinoma
Not specified
CONTROLS
Meth
Meth
Meth/Unmeth
Meth/Unmeth
Unmeth
Unmeth
Blank
Blank

TBX15

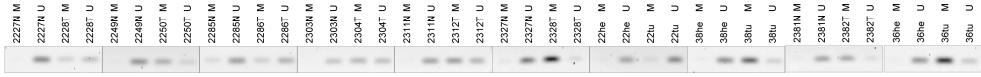
Squamous cell carcinoma



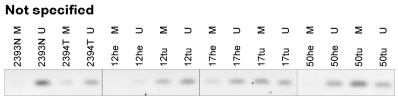
Adenocarcinoma

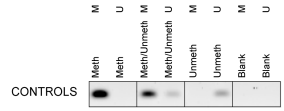


Large Cell carcinoma



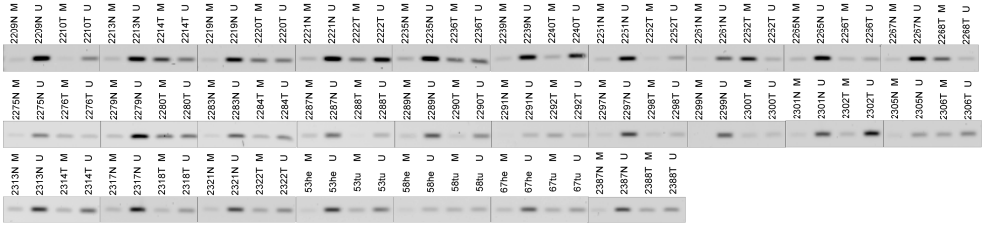
Not specified



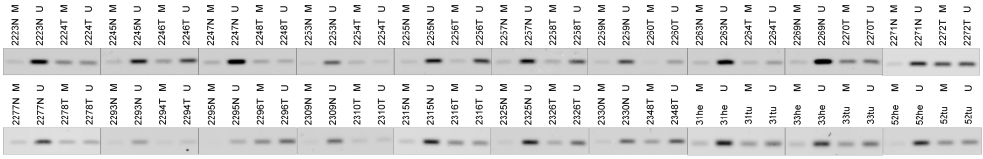


CDKN2A

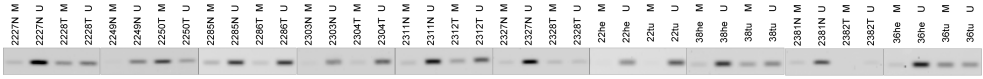
Squamous cell carcinoma



Adenocarcinoma

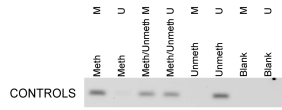


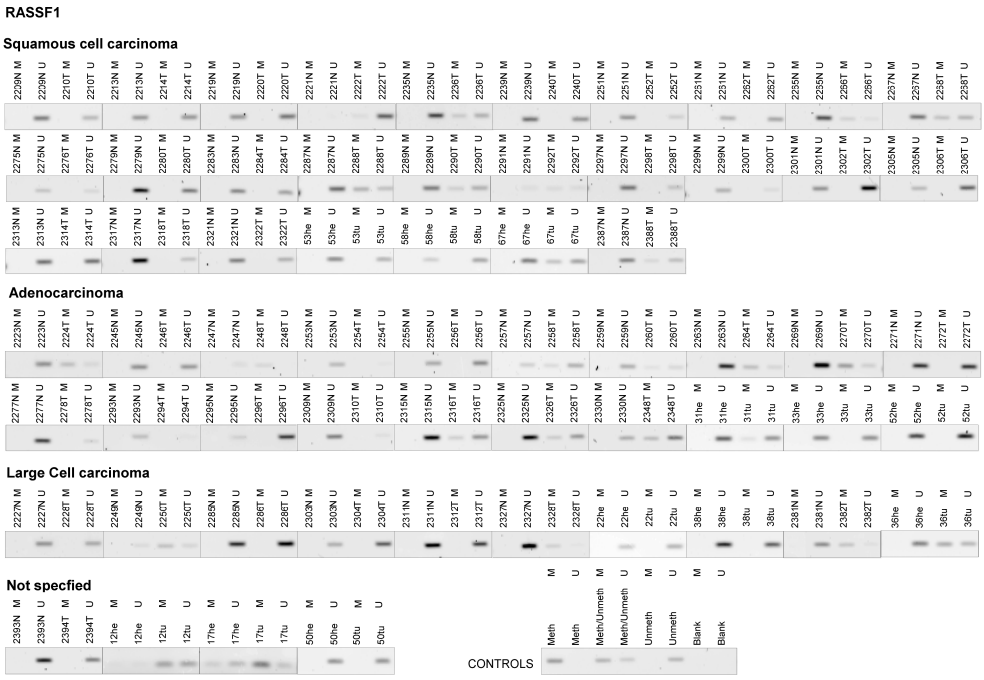
Large Cell carcinoma

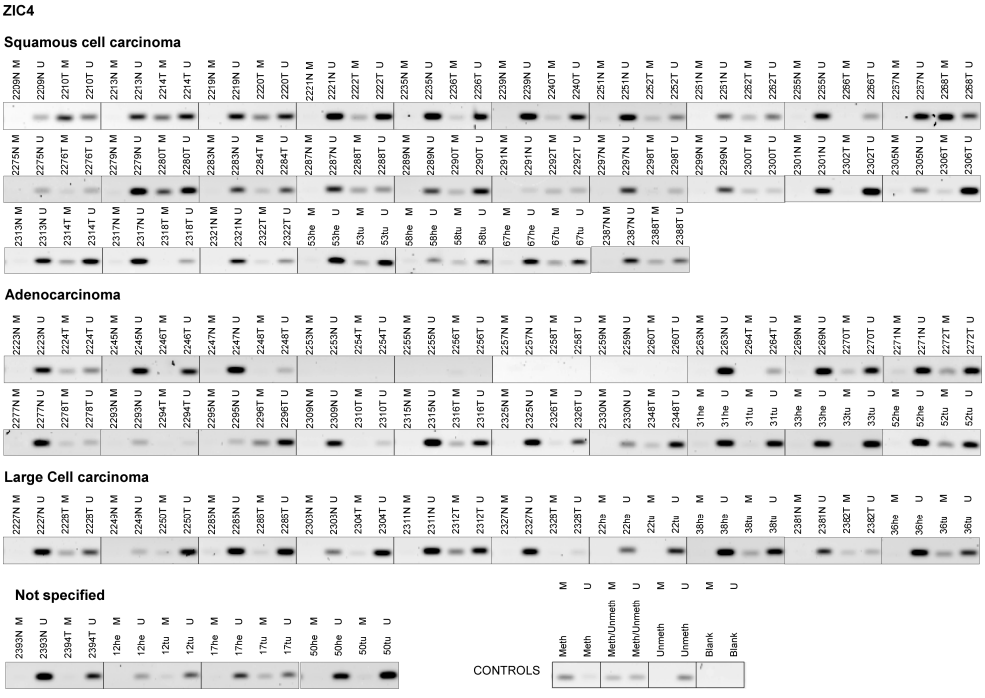


Not specified



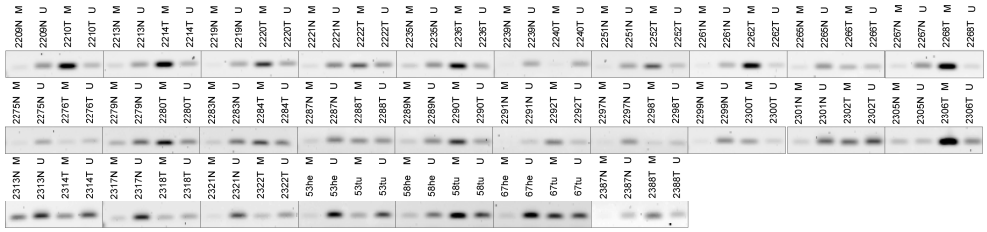



RASSF1
Squamous cell carcinoma
Adenocarcinoma
Large Cell carcinoma
Not specfied
CONTROLS

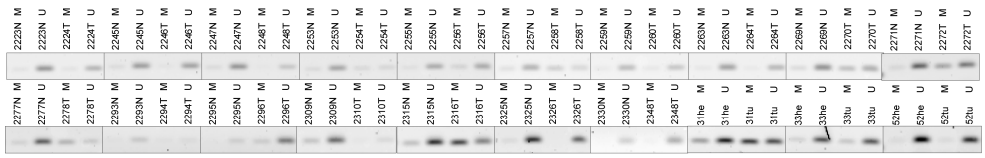

ZIC4
Squamous cell carcinoma
Adenocarcinoma
Large Cell carcinoma
Not specified
CONTROLS

DOT1L

Squamous cell carcinoma



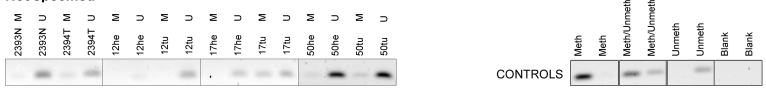
Adenocarcinoma



Large Cell carcinoma

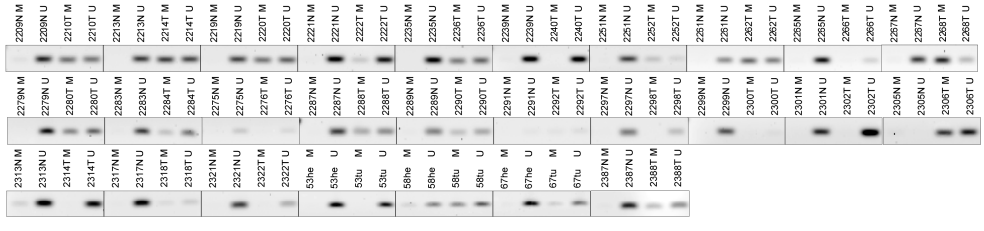


Not specified

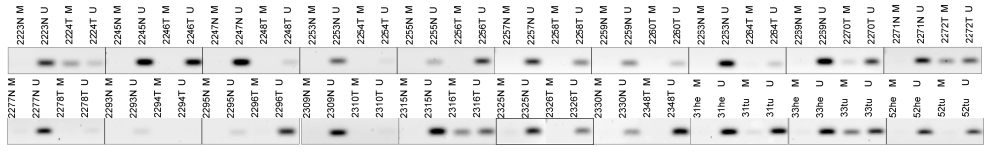


LIMK1

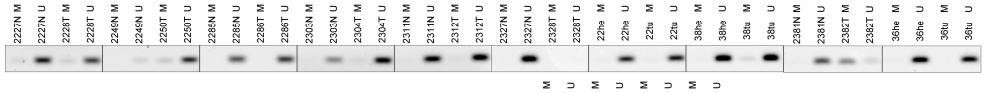
Squamous cell carcinoma



Adenocarcinoma



Large Cell carcinoma

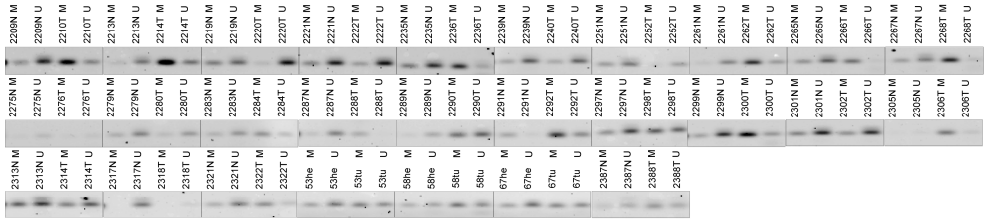


Not specified

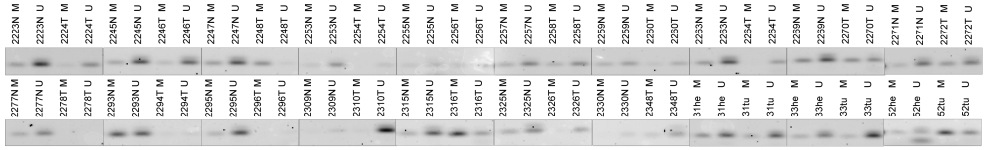


ANKRD1B

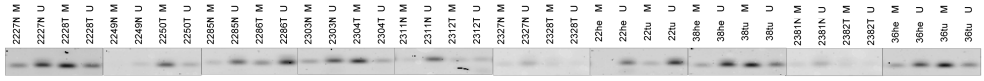
Squamous cell carcinoma



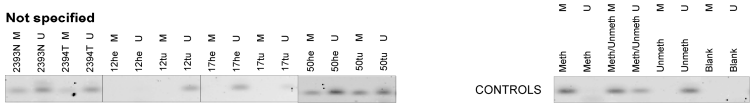
Adenocarcinoma



Large Cell carcinoma

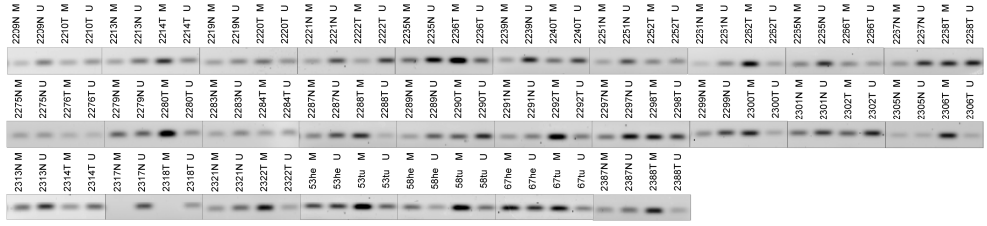


Not specified

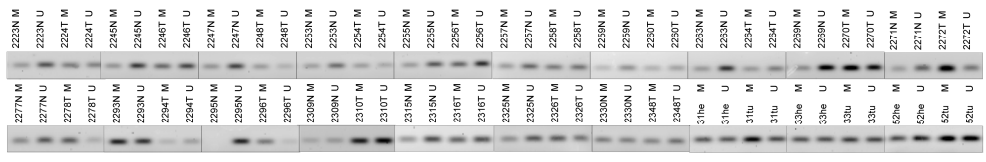


HIST1H3C

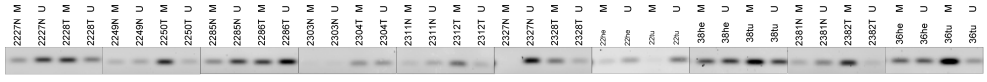
Squamous cell carcinoma



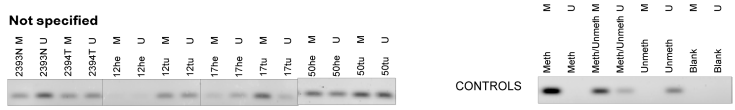
Adenocarcinoma



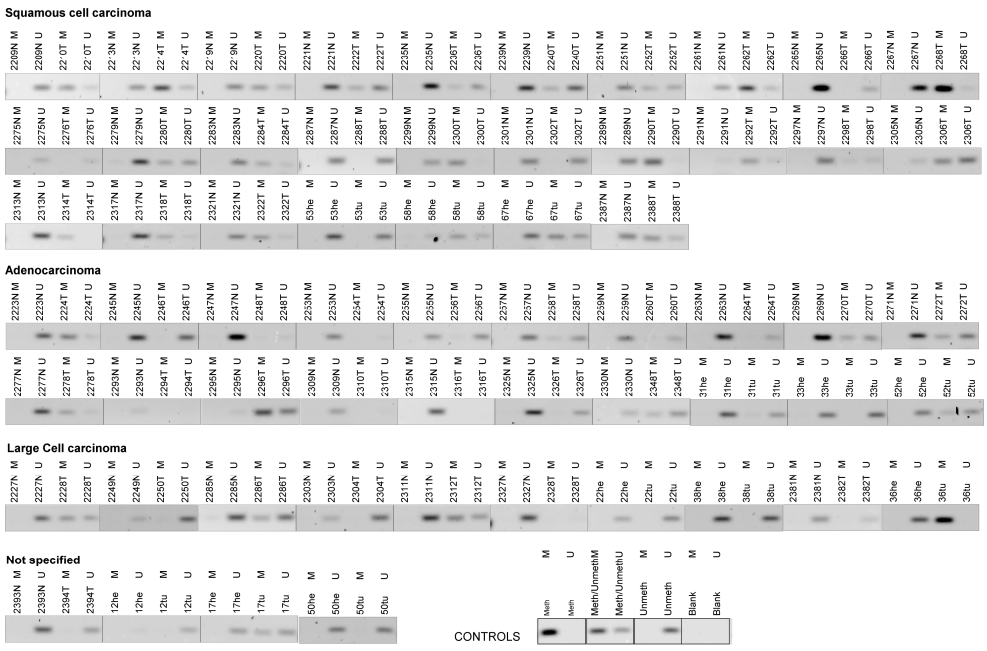
Large Cell carcinoma



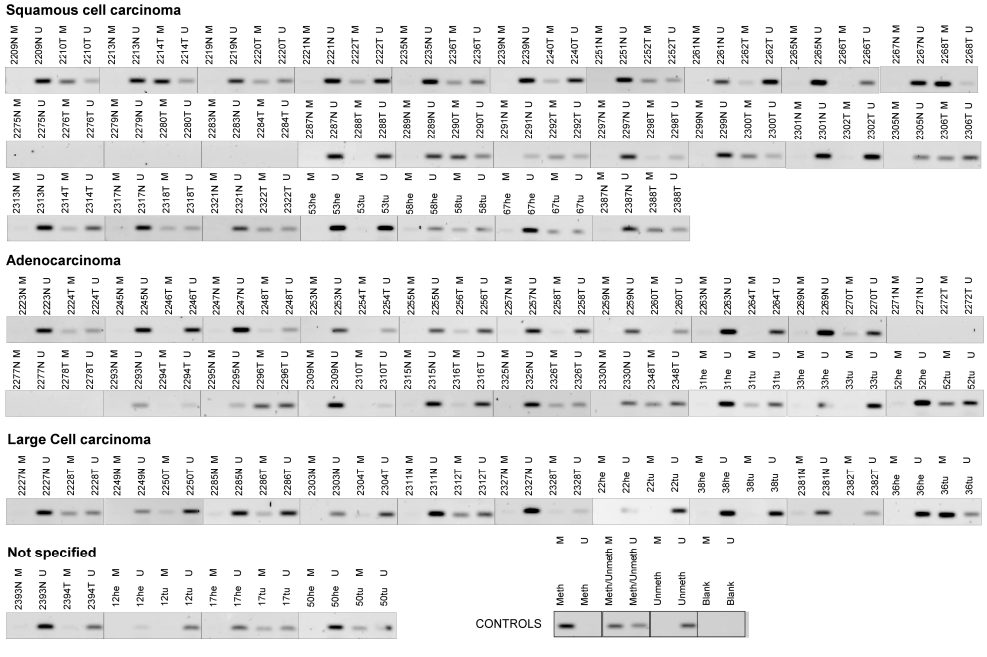
Not specified

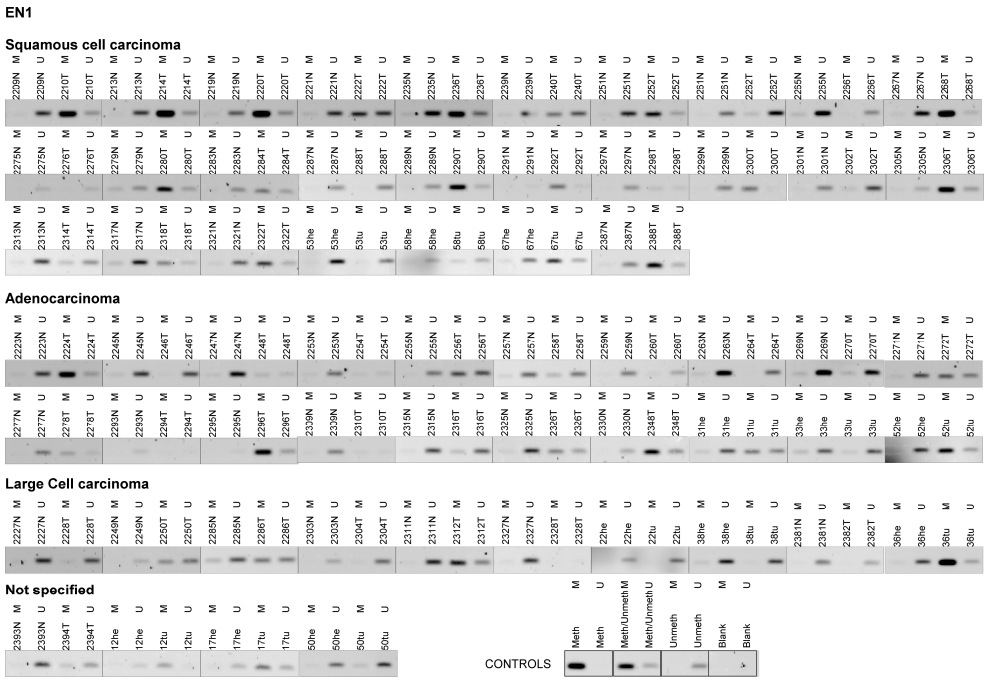


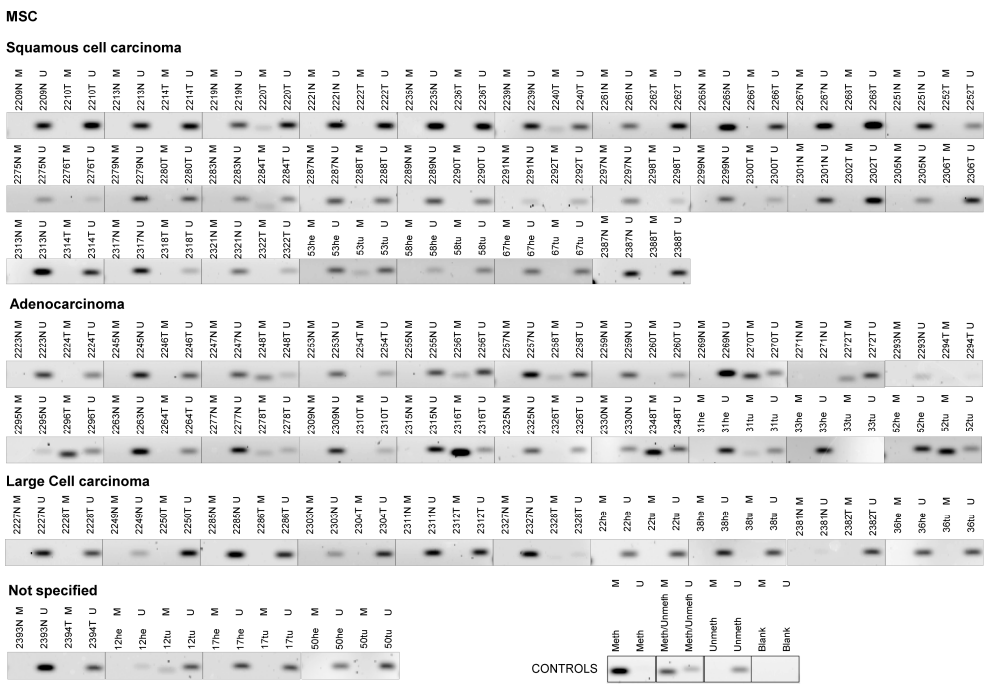
Frag01



Frag02

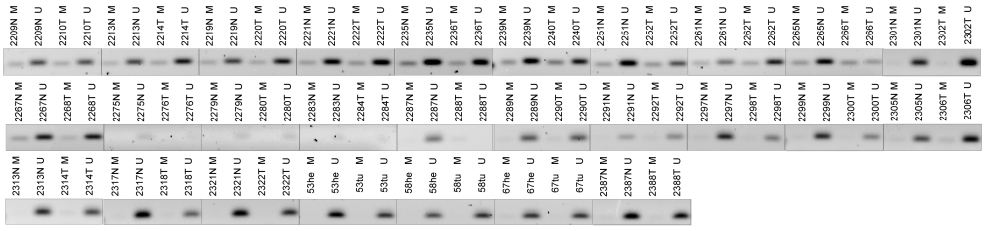



EN1
Squamous cell carcinoma
Adenocarcinoma
Large Cell carcinoma
Not specified
CONTROLS
Meth
Meth
Meth/Unmeth
Meth/Unmeth
Unmeth
Unmeth
Blank
Blank

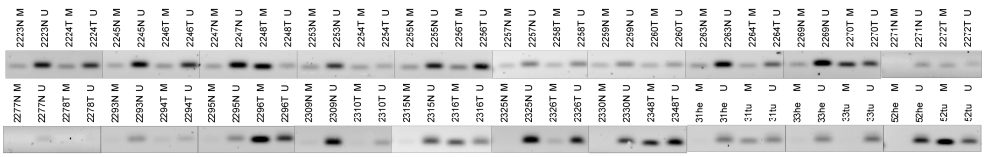

MSC
Squamous cell carcinoma
Adenocarcinoma
Large Cell carcinoma
Not specified
CONTROLS
Meth
Meth
Meth/Unmeth
Meth/Unmeth
Unmeth
Unmeth
Blank
Blank

GAS1

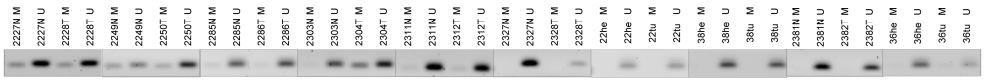
Squamous cell carcinoma



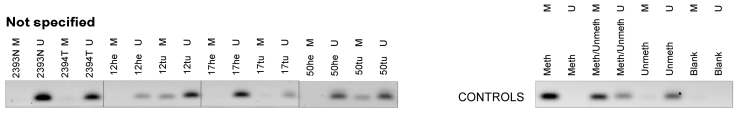
Adenocarcinoma



Large Cell carcinoma

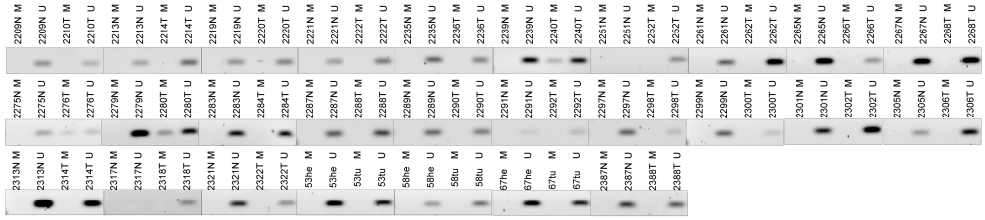


Not specified

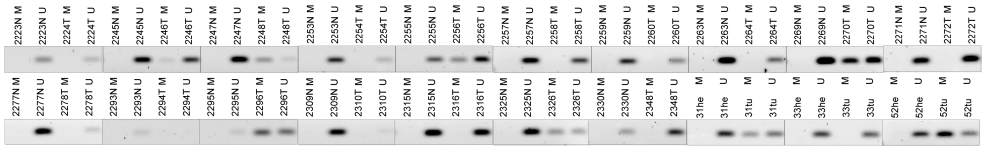


HOXA1

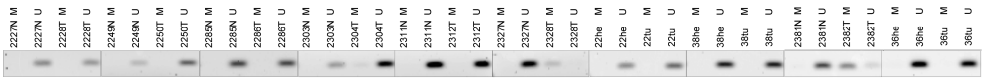
Squamous cell carcinoma



Adenocarcinoma

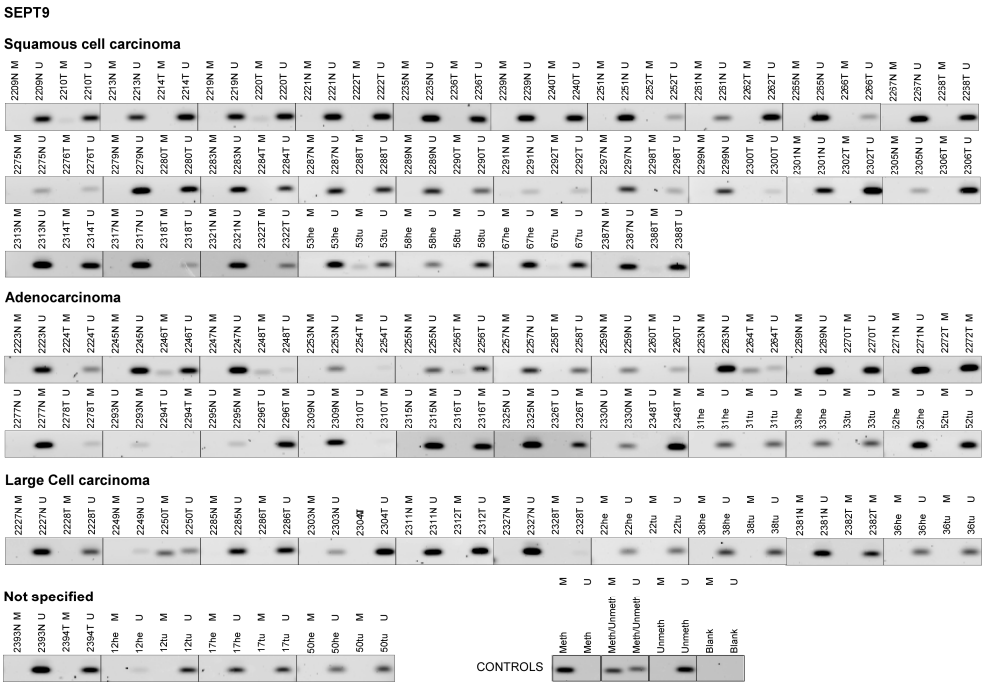


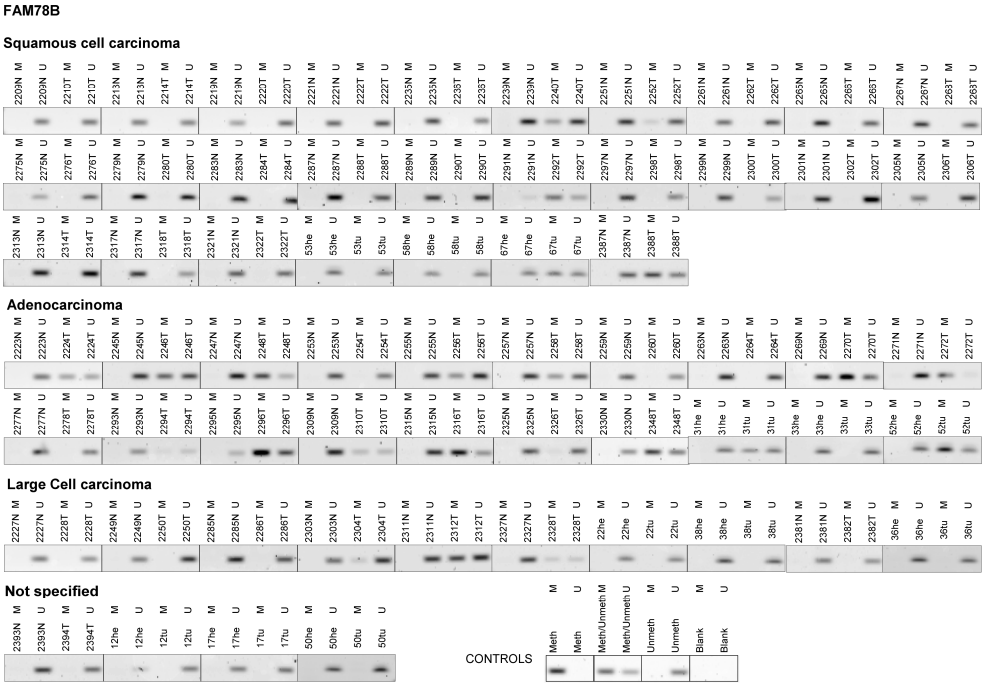
Large Cell carcinoma

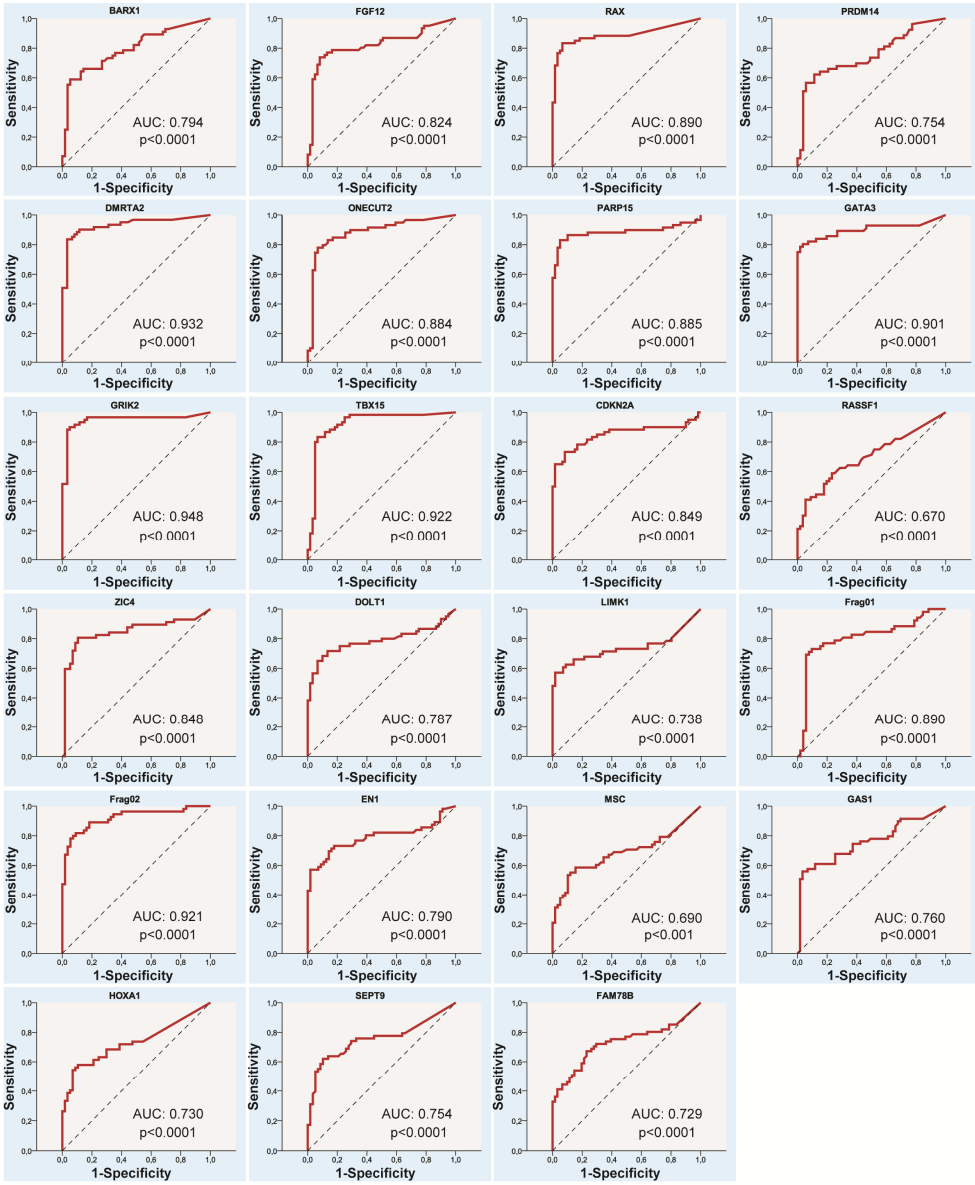


Not specified




SEPT9
Squamous cell carcinoma
Adenocarcinoma
Large Cell carcinoma
Not specified
CONTROLS
Meth
Meth
Meth/Unmeth
Meth/Unmeth
Unmeth
Unmeth
Blank
Blank


FAM78B
Squamous cell carcinoma
Adenocarcinoma
Large Cell carcinoma
Not specified
CONTROLS
Meth
Meth
Meth/Unmeth
Meth/Unmeth
Unmeth
Unmeth
Blank
Blank



Supplementary figure 2: Roc curves for all DMRs calculated on the percentages of methylated DNA from tumor versus lung sample. Y axis: sensitivity expressed in ratio; X axis: ratio of 1-specificity.



4

General Discussion

Discussion and prospects

Past, present and future

The genome era started in 1953 when the structure of DNA was unveiled by Francis Crick and James Watson. A few years later, one of the most famous concepts of molecular biology was conceived by the same Francis Crick, when he stated that transmission of genetic information goes from DNA to RNA and to protein. In the 80's the gene sequencing run started powered by the idea that organisms were the result from the sum of their genes. It all culminated with the sequence of the complete genome of James Watson himself fifty-four years later. In recent years, the genomes of many organisms have been unveiled and genetic discoveries are still being made at a high pace. Some old concepts have crumbled, others adapted and many new concepts have risen. In the past years the genetic/genome era has given more space to the epigenetic/epigenome age.

It is almost a paradox that the existence of 5mC was discovered even before the structure of DNA itself (Hotchkiss 1948), and that the knowledge that 5mC is an integral part of DNA came only a few years later (Gold et al 1963). Compared to genetic mechanisms, much less is known about how this epigenetic mark works in controlling cellular processes. One explanation for this is that epigenetics in general, and DNA methylation in particular, is a more plastic, dynamic and complex process than genetics. The genome of all cells in the body is the same and independent of the function of the cells. The epigenome is different. For example, although DNA methylation is a stable mark that can be faithfully transmitted through many cell divisions, it is influenced by environment factors and it is also tissue-specific.

Different developmental stages of an organism and different aging stages of a tissue provide examples of how DNA methylation patters can vary with time (Borgel et al 2010, Isagawa et al 2011, Koch and Wagner 2011). Studies in monozygotic twins (MZ) showed that the rate of genetic variance does not alter among MZ twins who were raised together or apart. In contrast, the variance in DNA methylation patterns is higher between these who lived in different environments (Bell and Spector 2011). Tissue specificity of DNA methylation patterns was observed long ago (Frei 1971, Waalwijk and Flavell 1978), and DNA methylation differences between human normal and tumor tissues were first described almost 30 years ago (Feinberg and Vogelstein 1983).

Mapping the genome took many years of combined effort from several groups. Because of its variability, mapping the epigenome is a much more difficult task. Many groups all over the world are getting together to join efforts in this direction. The Epigenome Network of Excellence, now the EpiGeneSys (EpiGeneSys 2011), the NIH Roadmap Epigenomics Program (Bernstein et al 2010), and more recently an international consortium was build (IHEC: International Human Epigenome

Consortium) aiming at mapping 1000 epigenome (IHEC 2010). Their plan is to start with 250 cell types at different developmental stages, from different organisms and different diseases.

At the end of this Discussion you can find an epigenetic timeline where I highlight the most important discoveries in epigenetics, with emphasis on DNA methylation.

The discovery that DNA methylation differs between tumor and normal cells led in the past three decades to a great effort by many groups on profiling methylation patters in tumors versus normal tissues.

In normal cells DNA methylation is involved in regulating differentiation and cell fate by two basic means. 1) Working as a signaling mechanism which affects positively or negatively gene transcription. Promoter hypermethylation is the best known mechanism of gene regulation, resulting in gene silencing. In contrast, methylation of the gene body is believed to positively affect gene transcription, working as a signal facilitating transcription by improving elongation efficiency and preventing recognition of false initiation sites by the transcription machinery (Hellman and Chess 2007, Zilberman et al 2007). 2) DNA methylation can physically interfere with accessibility of transcription factors (Wiench et al 2011, Xu et al 2010) or serve as an anchor for proteins involved in chromatin remodeling (Nakao 2001).

Each of these mechanisms has been observed to be altered in tumors. Tumorigenesis is a complex process involving combination of several types of genetic and epigenetics alterations. To better understand it, single gene analysis or localized gene investigation approaches will not suffice. Genome-wide methylation profiles (GWMP) are needed to give more insight in how methylation is involved in the biology of tumorigenesis. By itself GWMP may not help in understanding what these methylation differences mean, how the different aspects of DNA methylation influence and regulate gene expression, and how they can be applied in therapy. However GWMP of normal and tumor cells will shed light on the global mechanism of DNA methylation, consequently helping on identifying new candidate genes or genomic regions that undergo aberrant methylation at early neoplastic stages.

In the work described in this thesis we performed genome-wide methylation screening on seven NSCLC tumors and their matching lung tissues. NSCLC is a complex and heterogeneous malignancy which makes it a good model for defining whether GWMP can provide good basis for epigenetic markers discovery. In addition, the gene expression profiles from the same samples were available

Truly global methylation profiles of tumors are scarce and to date no genome-wide profile of NSCLC has been provided with the coverage depth and stringent methods applied in this work. Presently, one of the major issues in GWMP is the lack of standardized methodologies for either generating the methylated sequence reads

(method chosen for isolating and sequencing methylated DNA) or for the data analysis (method chosen for correcting and normalizing the reads). Therefore when data released by GWMP studies are compared most DMRs reported do not overlap. With these considerations in mind, we chose to produce our profiles using the MethylCap-seq methodology, which enriches genomic DNA samples for methylated DNA. The difference in our approach lies in the application of two novel procedures. Firstly we performed two independent experiments for each profile, two independent enrichment procedures and sequence runs for each of the tumors and lung tissues. This ensured that our profiles could be replicated. Secondly, we profiled fully methylated and fully unmethylated control DNA samples to help with normalization of the data, and for determining methylation background and saturation levels. Moreover we chose for a stringent analysis and cutoff level, ensuring that the reported profiles would be composed essentially of true DMRs, though in detriment of losing other possible DMRs. Collectively, this resulted in the production of a very reliable data set, which was validated by two independent methodologies.

We found almost 14500 DMRs distributed in genes (including promoter, gene body and 3'UTR) and genomic regions (including gene shore and gene deserts). Many of our findings are in accordance with previous published data. However we also find poorly described genes, regions which were observed only in other types of tumors, and totally novel DMRs. We observe a tendency for hypermethylation of CpG island in promoter regions of genes, which is compatible with studies indicating that is the most common mechanism of gene silencing by hypermethylation. However we also find hypermethylation of non-CpG island promoters. Non-CpG island promoters receive little attention in methylation studies. In a recent publication, (Han et al 2011) described that non-CpG island promoters are also targets of silencing by hypermethylation, similar to genes whose promoters fall in CpG islands.

Within the DMRS we identified we found regions which were common to all tumors (tumor methylation profile) and that were unique for each of the two most common NSCLC subtypes (subtype profiles). Our most interesting findings regarding these profiles are the DMRs present in gene shores, gene desert areas and poorly or not described genes and/or novel hypermethylated genes in tumors in general and NSCLC in particular. Gene shores and gene desert areas are seldom investigated due to the nature of the methodologies used until recently which focused mostly on genes. We have identified a gene shore region in chromosome 2q14.2 which has been implicated as a biomarker for colorectal cancer (Mayor et al 2009) and more recently for prostate cancer (Devaney et al 2011). Hypermethylation of this region had not yet been described in NSCLC. Susceptibility for genetic gain of this region has been reported for squamous cell carcinomas (Kang et al 2009). Interestingly hypermethylation spanning 1000 bp in this region was present in our squamous-cell carcinoma methylation profile. The closest gene to this DMR is EN1, situated 10 kbp downstream. EN1 is a transcription factor implicated in differentiation during early development,

and a member of the Wnt pathway as a regulator of β-catenin (Bachar-Dahan et al 2006). The role of hypermethylation of these specific regions in NSCLC tumorgenesis remains to be explained since EN1 is expressed in both NSCLC and lung tissues. However, long-range gene regulation is a known phenomenom (Splinter et al 2006) and long-range epigenetic silencing involving this region has also being described for colorectal tumors (Karpinski et al 2008, Mayor et al 2009).

By performing unsupervised cluster analysis in all DMRs we observed that the tumors grouped to the corresponding histological subtypes, indicating that our methylation profiles can provide a subtype-specific methylation signature. The heterogeneity of NSCLC poses a challenge for methylation profiles, where a large divergence is observed within the subtypes. This finding is in accordance with a recent report claiming that greater variation is observed within the same tumor tissue than within the normal cells, and that methylation profile of tumor cells are more similar among different tumors than to those found in the cells they originate from (Hansen et al 2011), indicating that there are particularities in each cancer but, at the same time the existence of a common epigenetic signature underlining the tumorgenesis mechanism.

A striking observation regards the number and type of genes hypermethylated in the tumors and in each of the subtype profiles. ADC is the most heterogeneous of the NSCLC subtypes and its methylation gene profile is similar to that of the tumors, with the majority of the hypermethylated genes being transcription regulators. In contrast, SCC presented a more extensive list of DMRs which fits the idea that

SCC is a more homogeneous type, therefore more regions are expected to be common between SCC samples. Moreover the SCC profile presented many genes associated with cytoskeleton organization. This is an intriguing observation since SCCs have peculiar cytological architectural features, such as abnormal production of keratin, presence of filament bundles and intercellular bridges, which are all cellular structures in which cytoskeleton organization genes are involved. This suggests the presence of a CpG Island methylator phenotype in this subtype of NSCLC, similar to that reported for colorectal and gastric cancers (Chen et al 2011, Samowitz 2007).

For testing the methylation profiles generated we decided to select candidate regions for screening a larger cohort of tumors and paired lung samples by MSP, which allows rapid investigation of the methylation status of selected genes or regions by PCR. Since we observed genes that had not been previously identified as hypermethylated in NSCLC and others which has not been reported in any type of cancer, we decided to focus on these genes, and include other known hypermethylated genes such as CDKN2A and RASSF1 as controls. For the tumor profiles we confirmed 23 out of 25 tested DMRs with this large scale test. Two fragments failed in the analysis since they amplified equally well in tumor and lung DNA samples. We observed that the percentage of tumors methylated for RASSF1 in our cohort corresponds to the rates observed in other studies (Honorio et al 2003, Yanagawa et al

2007, Yu et al 2008). For CDKN2A we observed a large percentage of the lung tissue DNA samples amplifying for methylation. The lung tissue DNA samples showed a much lower level of amplification than observed for the tumor DNA samples, supporting the hypothesis that CDKN2A hypermethylation is an early event in tumorgenesis (Huang et al 2011, Sterlacci et al 2011). All other DMRs were significantly differentially methylated between tumor and lung tissue DNA samples. Strikingly PARP15 followed the same pattern of amplification as CDKN2A. Little is known about the biological role of PARP15; other proteins of the PARP family are involved in DNA repair and apoptosis. PARP15 methylation is suggested to be related to leukemia (Gao et al 2010), suggesting a tumor suppressor role for the protein.

For some genes, such as BARX1, FGF12, RAX, DMRTA2, ONECUT2, GATA3 and a genomic region Frag02 we found more than 70% of the tumor samples methylated while only between 1% and 5% of the lung samples were methylated. Furthermore, the ZIC4 gene and the DMR located in 2q14 (named Frag01) were fully tumor specific, showing no amplification in lung samples. However, their sensitivity was lower than that of some of the other DMRS. Many of the selected DMRs are coinciding with genes encoding transcriptional regulators whose roles were first reported in neuronal development. Recently more reports are appearing emphasizing the role of neuro-developmental genes in tumorigenesis (Jung et al 2011, Kreisler et al 2010, Shin et al 2008). The methylation status of the genes reported in our study is higher than that reported for most candidate tumor markers found in literature (Ji et al 2011, Shames et al 2006, Zhang et al 2011).

Among these 25 DMRS, five were selected from the SCC profile and other five from the ADC profile. Using each of the SCC or ADC profiles we could not combine the DMRs to generate markers to predict either subtype. This may be due to several reasons: 1) The subtype cohort tested was fairly small with only 14 to 20 ADC tumors and 27 SCC; 2) The difference in numbers between the two subtype groups influenced the power of the statistical tests; 3) The small number of DMRs selected in each subtype profile. Probably the combination of the above issues was decisive for the impossibility to generate a subtype panel.

Nevertheless we could develop panels of candidate tumor markers to distinguish between tumor and lung tissue. This was achieved using two different approaches: 1) by calculating the percentage of methylated DNA in the amplified bands of each lung and tumor tissue and 2) by the standard MSP analysis (counting the number of lung and tumor samples with amplification for methylated genes). We generated two panels of markers. From the first approach a 4 marker panel was composed (BARX1, GRIK2, MSC and SEPT9) which had a 100% tumor sensitivity and 95% tumor specificity. From the standard MSP analysis we generated a panel of 5 markers (FGF12, GATA3, RAX, TBX15 and RASSF1) which has together 94% sensitivity and 96% specificity. Using a test sample set we could correctly predict tumor tissue in >90% of the cases.

We are currently correlating the gene expression profiles of these tumors with the methylation profiles. A preliminary analysis did not reveal a high correlation between the two sets. Few of the genes associated with the DMRs from the methylation profiles (tumor, ADC or SCC) diplayed differential expression in the microarray analysis (Figure 1). Only a few studies have attempted to correlate gene expression with methylation profiles, and those reporting a correlation were performed using the same methodology for producing the DMR profiles, CpG island amplification followed by CpG island microarray (Vincent et al 2011). Nevertheless, there are also studies which do not observe correlation between methylation levels and gene expression (Shearstone et al 2011). Our profiles were generated using two different methodologies, with different normalization methods, different statistical parameters and stringency. Further investigation and analysis of the data will be needed to bring both data sets to the same level where a more legitimate comparison can be produced.

In addition to the implications of methylation in tumor biology, and independent of how methylation affects gene expression, GWMP is an important tool for detecting markers which can provide valuable information on complex, heterogeneous and deadly diseases such as NSCLC. The diagnosis influences prognosis and treatment, and markers for early detection are believed to be crucial for improving patient care and survival. We have shown that for NSCLC there are more DMRs in the genome than the conventionally tested regions, and we have demonstrated that these genes, in addition to those that are commonly tested in cancer studies, are differentially methylated and can be used to discriminate tumor tissues. We propose that based on the comprehensive lists of DMRs originating from this study it may be possible to develop DMR signatures for correct identification of NSCLC subtypes. Furthermore, the DMRs are an exceptional source of potential biomarkers to be tested in serum and sputum, providing a method for minimally invasive diagnosis. Moreover, with the rapid advances in biotechnology, it will be possible in the near future to develop highly sensitive methods to detect these biomarkers in small amounts of DNA, allowing investigation of microdissected samples, therefore providing a more precise diagnosis of highly heterogeneous tumor types.

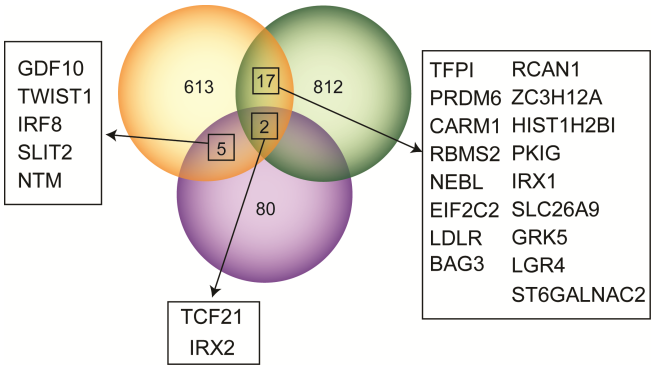
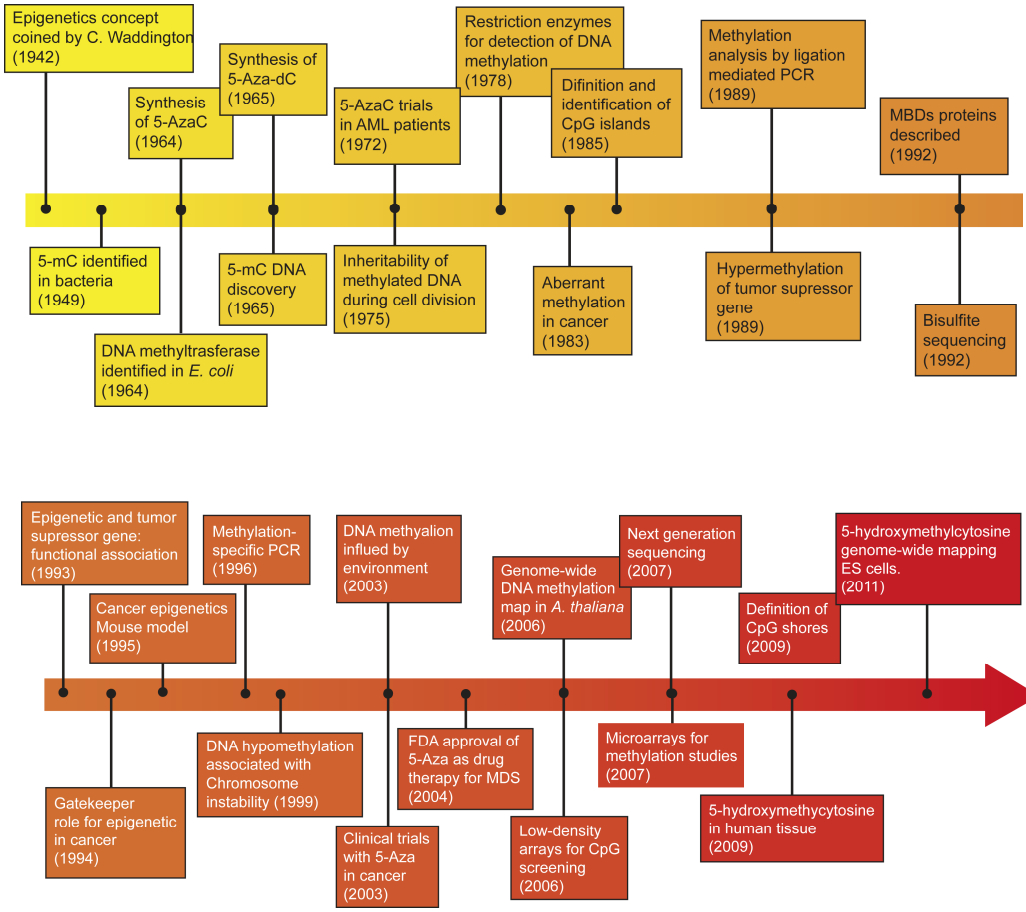


Figure 1: Preliminary results of cross-analysis between NSCLC expression profiles and methylation profiles. Analysis was done with the seven tumor/lung tissues used to generate the methylation profiles. Yellow circle: genes 1.5 fold up and down regulated in the tumors; green circle: genes hypermethylated in at least all squamous cell carcinomas; purple circle: genes hypermethylated in at least all adenocarcinomas. Amount of genes found to correlate between both profiles are featured within rectangles (Analysis and graphic from Mirjam van den Hout – Dept. Biomics, ErasmusMC).

Timeline of the hallmarks in DNA methylation discovery



References

Bachar-Dahan L, Goltzmann J, Yaniv A, Gazit A (2006). Engrailed-1 negatively regulates beta-catenin transcriptional activity by destabilizing beta-catenin via a glycogen synthase kinase-3beta-independent pathway. *Mol Biol Cell* **17:** 2572-2580.

Bell JT, Spector TD (2011). A twin approach to unraveling epigenetics. *Trends Genet* **27:** 116-125.

Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A *et al* (2010). The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28:** 1045-1048.

Borgel J, Guibert S, Li Y, Chiba H, Schubeler D, Sasaki H *et al* (2010). Targets and dynamics of promoter DNA methylation during early mouse development. *Nat Genet* **42:** 1093-1100.

Chen HY, Zhu BH, Zhang CH, Yang DJ, Peng JJ, Chen JH *et al* (2011). High CpG island methylator phenotype is associated with lymph node metastasis and prognosis in gastric cancer. *Cancer Sci*.

Devaney J, Stirzaker C, Qu W, Song JZ, Statham AL, Patterson KI *et al* (2011). Epigenetic deregulation across chromosome 2q14.2 differentiates normal from prostate cancer and provides a regional panel of novel DNA methylation cancer biomarkers. *Cancer Epidemiol Biomarkers Prev* **20:** 148-159.

EpiGeneSys (2011). EpiGeneSys Network of Excellence:http://www.epigenesys.eu/ (accession date: 02/12/2011).

Feinberg AP, Vogelstein B (1983). Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature* **301:** 89-92.

Frei JV (1971). Tissue-dependent differences in DNA methylation products of mice treated with methyl-labelled methylnitrosourea. *Int J Cancer* **7:** 436-442.

Gao A, Zuo X, Liu Q, Lu X, Guo W, Tian L (2010). Methylation of PARP-1 promoter involved in the regulation of benzene-induced decrease of PARP-1 mRNA expression. *Toxicol Lett* **195:** 114-118.

Gold M, Hurwitz J, Anders M (1963). The enzymatic methylation of RNA and DNA. *Biochem Biophys Res Commun* **11:** 107-114.

Han H, Cortez CC, Yang X, Nichols PW, Jones PA, Liang G (2011). DNA methylation directly silences genes with non-CpG island promoters and establishes a nucleosome occupied promoter. *Hum Mol Genet* **20:** 4299-4310.

Hansen KD, Timp W, Bravo HC, Sabunciyan S, Langmead B, McDonald OG *et al* (2011). Increased methylation variation in epigenetic domains across cancer types. *Nat Genet* **43:** 768-775.

Hellman A, Chess A (2007). Gene body-specific methylation on the active X chromosome. *Science* **315:** 1141-1143.

Honorio S, Agathanggelou A, Schuermann M, Pankow W, Viacava P, Maher ER *et al* (2003). Detection of RASSF1A aberrant promoter hypermethylation in sputum from chronic smokers and ductal carcinoma in situ from breast cancer patients. *Oncogene* **22:** 147-150.

Hotchkiss RD (1948). The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *J Biol Chem* **175:** 315-332.

Huang LW, Pan HS, Lin YH, Seow KM, Chen HJ, Hwang JL (2011). P16 methylation is an early event in cervical carcinogenesis. *Int J Gynecol Cancer* **21:** 452-456.

IHEC (2010). The International Human Epigenome Consortium:http://www.ihec-epigenomes.org/ (accession date: 02/12/2011).

Isagawa T, Nagae G, Shiraki N, Fujita T, Sato N, Ishikawa S *et al* (2011). DNA methylation profiling of embryonic stem cell differentiation into the three germ layers. *PLoS One* **6:** e26052.

Ji M, Guan H, Shi B, Hou P (2011). Concomitant hypermethylation of multiple genes in non-small cell lung cancer (NSCLC). *Folia Histochem Cytobiol* **49:** 132-141.

Jung S, Jeong D, Kim J, Yi L, Koo K, Lee J *et al* (2011). Epigenetic regulation of the potential tumor suppressor gene, hLHX6.1, in human cervical cancer. *Int J Oncol* **38:** 859-869.

Kang JU, Koo SH, Kwon KC, Park JW, Kim JM (2009). Identification of novel candidate target genes, including EPHB3, MASP1 and SST at 3q26.2-q29 in squamous cell carcinoma of the lung. *BMC Cancer* **9:** 237.

Karpinski P, Ramsey D, Grzebieniak Z, Sasiadek MM, Blin N (2008). The CpG island methylator phenotype correlates with long-range epigenetic silencing in colorectal cancer. *Mol Cancer Res* **6:** 585-591.

Koch CM, Wagner W (2011). Epigenetic-aging-signature to determine age in different tissues. *Aging (Albany NY)* **3:** 1018-1027.

Kreisler A, Strissel PL, Strick R, Neumann SB, Schumacher U, Becker CM (2010). Regulation of the NRSF/REST gene by methylation and CREB affects the cellular phenotype of small-cell lung cancer. *Oncogene* **29:** 5828-5838.

Mayor R, Casadome L, Azuara D, Moreno V, Clark SJ, Capella G *et al* (2009). Long-range epigenetic silencing at 2q14.2 affects most human colorectal cancers and may have application as a non-invasive biomarker of disease. *Br J Cancer* **100:** 1534-1539.

Nakao M (2001). Epigenetics: interaction of DNA methylation and chromatin. *Gene* **278:** 25-31.

Samowitz WS (2007). The CpG island methylator phenotype in colorectal cancer. *J Mol Diagn* **9:** 281-283.

Shames DS, Girard L, Gao B, Sato M, Lewis CM, Shivapurkar N *et al* (2006). A genome-wide screen for promoter methylation in lung cancer identifies novel methylation markers for multiple malignancies. *PLoS Med* **3:** e486.

Shearstone JR, Pop R, Bock C, Boyle P, Meissner A, Socolovsky M (2011). Global DNA demethylation during mouse erythropoiesis in vivo. *Science* **334:** 799-802.

Shin SS, Martino JJ, Chen S (2008). Metabotropic glutamate receptors (mGlus) and cellular transformation. *Neuropharmacology* **55:** 396-402.

Splinter E, Heath H, Kooren J, Palstra RJ, Klous P, Grosveld F *et al* (2006). CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes Dev* **20:** 2349-2354.

Sterlacci W, Tzankov A, Veits L, Zelger B, Bihl MP, Foerster A *et al* (2011). A comprehensive analysis of p16 expression, gene status, and promoter hypermethylation in surgically resected non-small cell lung carcinomas. *J Thorac Oncol* **6:** 1649-1657.

Vincent A, Omura N, Hong SM, Jaffe A, Eshleman J, Goggins M (2011). Genome-wide analysis of promoter methylation associated with gene expression profile in pancreatic adenocarcinoma. *Clin Cancer Res* **17:** 4341-4354.

Waalwijk C, Flavell RA (1978). DNA methylation at a CCGG sequence in the large intron of the rabbit beta-globin gene: tissue-specific variations. *Nucleic Acids Res* **5:** 4631-4634.

Wiench M, John S, Baek S, Johnson TA, Sung MH, Escobar T *et al* (2011). DNA methylation status predicts cell type-specific enhancer activity. *EMBO J* **30:** 3028-3039.

Xu J, Huo D, Chen Y, Nwachukwu C, Collins C, Rowell J *et al* (2010). CpG island methylation affects accessibility of the proximal BRCA1 promoter to transcription factors. *Breast Cancer Res Treat* **120:** 593-601.

Yanagawa N, Tamura G, Oizumi H, Kanauchi N, Endoh M, Sadahiro M *et al* (2007). Promoter hypermethylation of RASSF1A and RUNX3 genes as an independent prognostic prediction marker in surgically resected non-small cell lung cancers. *Lung Cancer* **58:** 131-138.

Yu ZH, Wang YC, Chen LB, Song Y, Liu C, Xia XY *et al* (2008). [Analysis of RASSF1A promoter hypermethylation in serum DNA of non-small cell lung cancer]. *Zhonghua Zhong Liu Za Zhi* **30:** 284-287.

Zhang Y, Wang R, Song H, Huang G, Yi J, Zheng Y *et al* (2011). Methylation of multiple genes as a candidate biomarker in non-small cell lung cancer. *Cancer Lett* **303:** 21-28.

Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S (2007). Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nat Genet* **39:** 61-69.



5

Addendum

Summary
Samenvatting
Curriculum vitae
PhD portfolio
Acknowledgements

SUMMARY

DNA methylation is a chemical modification of the DNA that interferes with gene regulation without causing any change in the gene sequence. Therefore, it is called an **epi**genetic modification. This epigenetic modification is involved in diverse cellular processes, such as imprinting and X-chromosome inactivation. Changes in the patterns of these markers are associated with development of different types of disease. Although the exact role of DNA methylation in gene regulation is not yet fully understood, in the past years genome-wide methylation profiling has become an important tool for identification of genes which are aberrantly methylated in complex diseases such as non-small cell lung carcinoma (NSCLC). NSCLC comprises a number of subtypes making it a heterogeneous disease. These subtypes present different biology and respond differently to treatment. Therefore correct diagnosis is important for good assessment of the disease, prognosis prediction and choice of treatment. Because DNA methylation is an epigenetic modification stably transmitted during cell division, it is a good candidate for the development of diagnostic and prognostic markers. In addition, DNA methylation is in principle a reversible mark and therefore a potential target for drug development.

In the first part of **Chapter 1**, I provide a review of NSCLC, focusing on the main characteristics which allow understanding of its complexity, the genetic and epigenetic factors involved in the origin of the disease and the recent advances in medicine regarding diagnostics and treatment. In the second part of this introductory chapter, I explain what DNA methylation is, what is known about its mechanism of action, and I discuss the most common methodologies to study it.

In this thesis we focused on profiling DNA methylation in NSCLC.

In **Chapter 2** we profiled seven NSCLC and their adjacent lung tissues. Using an approach named MethylCap-seq, where methylated DNA is precipitated by protein affinity and then sequenced by next-generation sequencing we have provided a comprehensive map of differentially methylated regions (DMRs) in the tumors compared to their control lung counterparts. To proof that this methodology reliably profiles DNA methylation, we validated our findings with two independent methods. These results supported the notion that MethylCap-seq is appropriate to generate genome-wide DNA methylation profiles. We have identified 57 highly significant DMRs common between all seven tumors profiled, including known and novel regions. We have also found regions which were present only in certain subtypes of NSCLC (squamous cell carcinomas or adenocarcinomas). This indicates that there may be a methylation signature which is characteristic for NSCLC subtypes. In this work we provide a unique resource containing genome-wide DNA methylation maps of NSCLC and control lung tissues, and comprehensive lists of known and novel DMRs and associated genes in NSCLC.

In **Chapter 3** we used the genome-wide methylation profiles for selection of the 25 most significant DMRs, based on their methylation scores. These DMRs might be involved in tumor biology. We tested these DMRs in 61 patients to see if they would show the same performance when a larger cohort of tumors were analyzed. We found that 18 DMRS showed a highly significant difference between tumor and control lung tissues. From these, 12 DMRs had not yet been reported as differentially methylated in NSCLC. In addition, these DMRs also presented higher sensitivity and specificity than previously proposed NSCLC methylation markers. We developed a panel of seven DMRs, composed of novel and known markers, which is capable of distinguishing tumors from control lung tissues with 90% accuracy.

The work in this thesis provides the first truly genome-wide DNA methylation profiles of NSCLC.

In **Chapter 4** I discuss the implications of our findings for this type of cancer, opening opportunities for further investigation that can lead to the development of sensitive biological markers for NSCLC. This may enable non-invasive diagnosis and histological classification. I also discuss how our work contributes to DNA methylation studies in general, and how this may lead to a better understanding of the biological roles of this type of modification.

SAMENVATTING

DNA methylatie is een chemische modificatie van het DNA dat regulatie van genen beïnvloedt zonder verandering in de gensequentie. Dit wordt **epi**genetica genoemd. Deze epigenetische modificatie is betrokken bij diverse cellulaire processen, zoals imprinting en X-chromosoom inactivatie. Veranderingen in de patronen van deze markeringen worden geassocieerd met de ontwikkeling van verschillende soorten ziektes. Hoewel de precieze rol van DNA-methylering in gen regulatie nog niet volledig wordt begrepen, is genoom-brede methylatie profilering in de afgelopen jaren uitgegroeid tot een belangrijk instrument voor de identificatie van genen die afwijkend zijn in complexe ziekten zoals niet-kleincellig longcarcinoom (NSCLC). NSCLC bestaat uit divers subtypen tumoren waardoor het een zeer heterogene ziekte is. Deze subtypen hebben biologische verschillen en ze reageren ook verschillend op behandelingen. Daarom is een juiste diagnose belangrijk voor een goede beoordeling van de ziekte, voor prognose voorspelling en voor het soort behandeling. Omdat DNA-methylering een stabiele epigenetische modificatie is die overgedragen kan worden tijdens de celdeling, kan dit gebruikt worden voor de ontwikkeling van diagnostische en prognostische markers. Daarbij is DNA-methylering in principe een omkeerbare mark, en dat biedt de mogelijkheid voor de ontwikkeling van behandelingen die gericht zijn om deze markering te verwijderen.

In het eerste deel van **hoofdstuk 1**, heb ik een overzicht gegeven van niet-kleincellig longcarcinoom, gericht op de belangrijkste kenmerken van dit soort longkanker. Hier krijgt u een idee van de complexiteit van NSCLC, de genetische en epigenetische factoren die betrokken zijn bij het ontstaan van de ziekte en de recente vooruitgang in de geneeskunde met betrekking tot diagnostiek en behandeling. In het tweede deel van dit inleidende hoofdstuk, leg ik uit wat DNA-methylatie is, wat er bekend is over het mechanisme en bespreek ik ook de meest voorkomende methoden om DNA-methylatie te bestuderen.

In dit proefschrift hebben we ons gericht op de profilering van DNA-methylatie in NSCLC.

In **hoofdstuk 2** hebben we zeven NSCLC en aangrenzend long weefsel geprofileerd. Met behulp van MethylCap-seq, waar gemethyleerd DNA wordt neergeslagen door proteïne affiniteit en vervolgens gesequenced met behulp van next-generation sequencing, hebben we een uitgebreide kaart verkregen van differentieel gemethyleerde gebieden (DMRs) in de tumoren versus aangrenzend weefsel. Om te bewijzen dat deze methode betrouwbare DNA-methylatie profielen kan geven, hebben we onze bevindingen bewezen met twee onafhankelijke methoden. Deze resultaten bewijzen dat MethylCap-seq geschikt is om genoom-brede DNA-methylatie profielen te genereren. We hebben 57 zeer significante DMRs gevonden bij alle zeven onderzocht tumoren, met inbegrip van bekende en nieuwe gebieden. Wij hebben ook regio's gevonden die slechts aanwezig waren in bepaalde subtypen van NSCLC (squamous cell carcinomen

of adenocarcinomen). Dit geeft aan dat methylatie mogelijk een kenmerk is karakteristiek voor NSCLC subtypen. Dit werk levert unieke informatie op namelijk genoom-brede DNA-methylatie kaarten van niet-kleincellig longcarcinoom en controle longen, en uitgebreide lijsten van bekende en nieuwe DMRs en de bijbehorende genen in NSCLC.

In **hoofdstuk 3** hebben we de genoom-brede methylatie profielen gebruikt om de 25ste belangrijkste DMRs te selecteren, op basis van hun methylatie scores. Deze DMRs zouden betrokken kunnen zijn in de tumor biologie. We hebben deze DMRs bij 61 patiënten geanalyseerd om te zien of ze hetzelfde zouden laten zien in een groter cohort van tumoren. Wij vonden dat 18 DMRs een zeer significant verschil tonen tussen de longtumoren en het controle weefsel. Hiervan zijn 12 nieuwe DMRs differentieel gemethyleerd in NSCLC. Bovendien hebben deze DMRs ook een hogere gevoeligheid en specificiteit dan de eerder voorgestelde NSCLC methylatie markers. We hebben ook een panel van zeven DMRs ontwikkeld, bestaande uit nieuwe en bekende markeringen die in staat is longtumoren van controle longweefsel met 90% nauwkeurigheid te onderscheiden.

Het werk beschreven in dit proefschrift levert het eerste echte genoom-brede DNA-methylatie-profielen van niet-kleincellig longcarcinoom op.

In **hoofdstuk 4** beschrijf ik de consequenties van onze bevindingen voor dit type van kanker. We bieden mogelijkheden voor verder onderzoek dat kan leiden tot de ontwikkeling van gevoelige biologische markers voor NSCLC, wat weer kan leiden tot ontwikkeling van niet-invasieve diagnostiek en histologische classificatie. Hier beschrijf ik ook hoe ons werk bijdraagt aan de DNA-methylatie studies in het algemeen en hoe dit kan leiden tot een beter begrip van de biologische rol van DNA-methylering.

Curriculum Vitae

Rejane Hughes Carvalho
Salvador, Brazil

Education

PhD in Biomedical Genetics — March 2008 – Feb. 2012
Erasmus MC - Rotterdam

Master in Molecular Cell Biology and Bioinformatics — Sep. 2005 – Sep. 2007
University of Amsterdam

Specialization in Forensic medicine — May 2002 – Apr. 2003
University of the State of Bahia and Technical Police

BSc in Biological Science — May 1996 – Sep. 2001
Federal University of Bahia

Research experience

Trainee — Feb. 2007 – Aug. 2007
LRRK-2 and GBA genes in Parkinson Disease
Laboratory of Clinical Genetics – ErasmusMC, Rotterdam

Trainee — Feb. 2006 – Oct. 2006
NRL and CRX genes in Medulloblastoma development
Laboratory of Human Genetics – AMC-UvA, Amsterdam

Research assistant — Sep. 2002 – Oct. 2003
Vaccine development against Leishmaniasis
Gonçalo Moniz Research Center-FIOCRUZ-Salvador, Brazil

Technician — Nov. 2001 – Jul. 2002
Diagnostics of chromosomal disorders
Health Science Institute – Salvador, Brazil

Trainee — Mar. 1999 – Oct. 2001
Makers for Diabetes Type II and Metabolic Syndrome
Hospital Edgard Santos – Salvador, Brazil

Trainee — Dec. 1996 – Dec. 1998
Chromosomal abnormalities (awarded with two research grants)
Laboratory of Human Genetics – Salvador, Brazil



PhD Portfolio

Summary of PhD training and teaching

Name PhD Student:	Rejane Hughes Carvalho	**PhD period:**	March 2008 – Feb. 2012
ErasmusMC Department:	Cell Biology	**Promotor:**	Prof. Dr. Sjaak Philipsen
Research Schoool:	Medisch-genetisch center	**Co-promotor:**	Dr. Joachim.G.J.V. Aerts

1. PhD training

General academic skills	**Year**	**Workload (hours)**
- BioBusiness Summer School	2008	80
- Biomedical English writing and communication	2011	84
- Partek training course	2011	24
General courses		
- Safely working in laboratory	2008	8
- Analysis of microarray gene expression	2008	32
- Epigenetic mechanisms in health and disease	2008	16
- Genome maintenance and cancer	2009	16
- Next generation sequencing data analysis	2009	24
- Technological facilities	2009	8
- Experimental approach to molecular and cell biology	2010	168
Meetings and Conferences		
- CGC/CBG meeting "Molecular mechanisms and mouse models in cancer	2008	16
- CGC annual meeting	2009	8
- CGC/CBG meeting "Molecular mechanisms in cancer"	2010	16
- 4th Benelux NGS meeting	2011	16
- CGC annual meeting (oral presentation)	2011	16
- Welcome trust Conference – Epigenetics of common diseases (poster)	2011	40
- Bridge Meeting (oral presentation)	2011	2
- Monday Morning Meetings (Cell Biology weekly meeting)	2008-2012	200
- Work discussions	2008-2012	80

Seminars and workshops		
- 18th MGC Symposium	2008	12
- 15th MGC Workshop	2008	8
- 19th MGC Symposium	2009	12
- Bioinformatic Analysis workshop	2009	24
- Writing successful Grant Proposals	2009	8
- 20th MGC Symposium	2010	12
- EuTRACC Bioinformatics Workshop	2011	32

ACKNOWLEDGEMENTS

The past four years felt like being in a roller-coaster. Now that I'm reaching the final meters of this adventure, I look back and realize that the "ups" were much more frequent than the "downs" and for this I have to thanks the people that helped me keeping on track.

On the top, of course, I want to thank my promoter: Sjaak Philipsen. Dear Sjaak, first let me thank you for this opportunity. I think you had no idea what you were getting yourself into when you accepted me to do this project. Therefore, thank you for putting up with my "latin temperament". My appreciation for you has also so many more reasons: for being present even when absent, and for being missed when not present. For your optimism, which had the weird power to change my mood (in a good way) after every talk we had. For your "super-vision", your brightness; for your silent, though markedly strong guidance; for always coming up with solutions I hadn't thought about, no matter how much I had searched for (frustrating but very "leerzaam"). Thank you! You made the past years easier than they otherwise would have been.

Joachim Aerts, thanks for accepting to be my co-promoter. Thanks also for your prompt replies to my questions and requests. It has been a pleasure working with you.

To my small committee, Prof. Ruud Delwel, Prof. Rudi Hendriks and Prof. Dies Meijer. I very much appreciate your suggestions and I thank you for making time to talk personally with me and advising me on my thesis.

Professor Frank Grosveld. Thank you for accepting me as a member of the Cell Biology Department. My appreciation also for your insights and mostly for recommending the assay that ended up as the basis of my project.

I extend my thanks also to our collaborators in Norway. Prof. Boris Lenhard, Vanja Haberle and Supat Thongjuea for all the working on the sequencing data and for your good insights on the normalization and statistical analysis.

Marike, Jasperina and Bep, we would all be lost without you (I specially). Thanks for your patience, for answering all my weird questions and request, which some times had nothing to do with your duties. And specially to make me feel I was not alone when I got stuck in the middle of nowhere in London.

Melle, Leo, Koos and Enno, what would we do without you? Thanks for dealing with my request, processing all the financial issues for me and tracking back my packs, when they would get lost, and doing all of it with care and an incredible efficiency.

Pim, Leo, Ton and Sjozef, thanks for taking care of my computer and all questions about updates, passwords, programs and also for the advices about software's and hardware's. Pim, thanks for keeping me mentally sane when you saved my files, and for being always in such a good mood, even when I would come to bug you with all kind of requests.

My sincere gratitude to Nynke. First, thanks for the translation of my summary. But above all thank you for listening, for always making time for me, for thinking with me, for your care and advices, for your support at all times. I will never be able to show you my gratitude for everything you have done for me (and Ton too, to whom I also extend my thanks). It felt/feels good to know that I could/can count on you whenever I need.

I also want to thank my colleagues. Thamar, I appreciate that at any time I came with a question you made some time for me. Thanks also for making blind eyes when I would "steal" one of your coffee pads ☺, and mostly for providing the way to have a good cup of coffee; thanks also, of course, for lending me your mobile charger every time I would ran out of battery (what was quite often). Teus, our statistics counsellor, thanks for your help with the statistics and graphics. Sahar, in the past 4 years we had our up and downs and I like to believe that the up's overruled the down's. Your cheerful words and your encouragement helped a lot. Thank you for always showing available to help and for the quick thee-breaks in your apartment ☺. I wish you success in the next steps of your life. Sylvia, Pavlos, Anna and Divine, it has been great working alongside you and I enjoyed your company. Jun, thanks for your analysis, for being prompt to look and process the uncountable lists Sjaak and I kept sending you.

To all colleagues from Cell Biology. Thank you all for making of Cell Biology such a nice place to work. Alex (ja, zeker!!! ☺), Dubbi and Eric Soler, thanks for your advices and for being available to help.

Thanks also to the colleagues of Genetics, with special thanks to Romana. Thanks you for your time, your tips and our talks, and for listening to my silly doubts about "dutch behaviour". Thanks also to Pim, for landing me the gel combs and trays and for putting a smile in my face greeting me “Hi Jane” every time I would show up in the department. Gosia thanks for listening and cheering me up. Keep on with the dancing!

My gratitude and appreciation to all colleagues of Clinical Genetics. A special thanks to Bianca, thanks for being such a nice person to me and all your help in the short time I was in Clinical Genetics. And Leontine, thanks for borrowing me the bisulfite kit when I would run out of it.

My thanks and appreciation to everybody from the different departments, without whom I would have had a hard time to go on with my project. Wilfred, thank you for your help and advices with the next-generation sequencing, and for your availability though I know you have a busy schedule. Mirjam, thanks for helping me since the time you were working in the Bioinformatics department. I still remember the first exchanged messages when you patiently explained me what “md5sum” was. Thanks for being always so helpful and patient. Rutger, thanks for the analysis of the motifs, for all explanations about the programming languages, and for the chit-chats; it has been great to exchange ideas with you. I know you are both overloaded with work, so many thanks for squeezing me in your agendas. Christel, Zeliha and Antoine, thanks for the help with the equipments, all explanations and for not giving me a hard time when I needed to use the PCR machines. Erik-Jan, thanks for helping with the sequenicng data in the very begining and the solex program.

My gratitude also to Margaretha and Joost Hegmans. Margaretha, I cannot thank you enough for your help with the staining, but above all for your willingness to help. I enjoyed a lot the few times we worked together, and even more the times we had coffee breaks. Lot of success with your new house and the whole “verbouwing”. Joost, it was a pleasure to work with you when we had the monthly discussions. Thanks for being such an encouraging (euphemism for provoking) person, it helped keeping my mind sharp☺.

I thank also Monique Oomen, Peter Hegman, Anieta Sieuwerts and John Foekens, for your promptness and help on providing me with the tissues. Illa Grigoriev, for kindly giving me the MRC5 cell line, without it I would have lost precious time in the beginning of my project. Walter de Valk, thanks not only for your help with all the bisulfite sequencing samples, but for the nice and friendly talks.

I need to extend my gratitude to Marcel Koos, Jens Bunt and Mohamed Hamdi. Although the three of you were not directly involved in the past 4 years that led to my degree, I will be always grateful for the opportunity you gave me in my return-to-science here in Holland, Marcel. There is where everything started, and I enjoyed very much working to/with you in the AMC along with Jens and Mohamed.

Ik wil ook Jopie, Natalina en hun collegas bedanken voor alle hulp en voor het zorgen dat alles schoon en in order was. Natalina, het was altijd een leuk begin van de dag om uw vrolijke “bom dia” te horen. Jopie, dank u ook wel voor uw geduld en hulp.

There are many people that I need to tell them how much they are appreciated. Among them: Herve, Piere, Sara, Bianca, Francesca, Luna, Carol and Weng Yee, thanks for, in course of those years, in one way or another, making my days birghter and the evenings more festive.

Ik waardeer en ben erg dankbaar om onze vriendschap, Ronald. Dankjewel voor de gezelligheid, voor jouw tijd, de koffie pauzes en voor het gezellige diner uitjes. Ik zal altijd een leuke herinnering hebben aan de dag dat jij op een koude avond jouw warme huisje verliet om met mij een kopje koffie te drinken en mij te helpen wakker te blijven omdat ik overnacht moest werken. Mocht je ooit in Brazilië op vakantie gaan, dan zou ik met alle plezier jou gids willen zijn.

Marco, Elly en Inge. Ik zal jullie voor altijd dankbaar zijn voor alles wat jullie voor mij hebben gedaan, vanaf het begin in 2003 tot nu toe. Ik had dit promotietraject niet zo succesvol kunnen doorlopen, of dit stuk kunnen schrijven zonder jullie hulp en steun, in welk opzicht dan ook. Marco, bedankt voor je hulp bij de saaie klus van het tellen van gel bandjes. Maar bovenal kan ik je niet genoeg bedanken voor het redden van mijn bestanden van de kapotte harde schijf. Dank jullie/je wel voor alles.

Ook bedank ik Monique, Debby en Armand. Bedankt voor de vriendschap, de steun en het begrip. Ik hoop jullie binnenkort weer te kunnen zien.

Vojka, my friend, I know you don't like to listen to me saying "thanks" to you, but this time you will have to read it ☺. Thanks for your support, for understanding, for checking up on me; for being such a great friend. Our political debates and philosophical talks were/are always stimulating and I doubt I will ever find anyone like you with whom I can have this type of talk without running the risk to be taken to a mental institution. And of course also Vera, Milos and Frans, for the "gezellig" avonden.

A big thanks to Alex. Thanks a lot for the R script, it helpped me immensily, and thank you mainly for standing my heated temperament and discussion with Lena without sending me out of your home ☺.

Lena and Cíntia, my paranimfs and, above all, the best friends one could wish for. There are just not words enough to express how much you both mean to me. Lena, who could say that it is possible to find in Holland a Russian person with so much of a strong character as myself? How can two people be so alike in personality but also have such a different taste about almost everything? I will never forget our heated discussion about politics or our late "happy hours", either in your house or in the pool-bar, or just in a restaurant, talking about everything and nothing. Or that day where we stayed outside the gym for almost an hour discussing books and movies; first exercise the mind, then the body. Nor will I forget the best news-years eve of my life. So many nice moments. Lena, спасибо, my dear friend.

Cí, você é uma amiga incrível. Pela primeira vez sinto uma amizade bi-lateral, onde não só sabemos que podemos contra uma com a outra, mas que realmente nos abrimos e confiamos uma na outra. Obrigada por tudo, minha amiga, pelo carinho e atenção, principalmente nesses últimos meses. Acho que você não faz idéia de como, mesmo estando longe, você me ajudou. Tenho certeza que nossa amizade vai continuar mesmo quando não estivermos mais aqui. Adoro você. Ricardo, meu amigo, eu acredito que para amizades se formarem, assim como para muitas outras coisas, precisa-se das pessoas certas, no momento certo, você e Cí entraram na minha vida no momento certo, e não há dúvida de que foram/são as pessoas certas. Obrigada por tudo, Ric.

Gostraia de agradecer também Dr. Judith, meu querido amigo Adriano e minha amiga Adriana. Doutora, a senhora não é só uma mentora, mas também uma amiga. Obrigada pelos ensinamentos e principalmente pelo apoio durante todos esse anos. Adriano e Adriana, meus queridos e grande amigos. Vocês não se conhecem, mas são o símbolo de que quando duas pessoas são amigas, distância e tempo não importa. Será que tem algo a ver com o nome de vocês? ☺

À minha família: tias, tios, primas e primos. Mesmo eu estando longe, sei que faço parte dessa família grande, complicada, mas que se ama. Se tem algo que realmente apredi aqui é que minha família são vocês, são vocês que me aceitam como sou e me perdoam quando erro. Família não se toma emprestado.

Painho, Ginho, Séu, Manda, Mari, Dani, Beca, Xande e Teu. Vocês são meus pilares; são exatamento como os fundamentos de uma construção, a gente raramente vê, mas sem eles a construção desaba. São minha bússola (nesse caso ao contrário, já que o ponteiro da minha direciona pro sul ☺). Só de saber que tenho vocês, nunca me sinto perdida. Eu não conseguiria ter estado aqui todos esses anos, se não tivesse certeza que posso contar com todos e cada um de vocês. Amo vocês.

To everyone, once more, T-H-A-N-K-S!!!!

Jane