Dissertation

submitted to the
Combined Faculties for the Natural Sciences and for Mathematics
Of the Ruperto-Carola University of Heidelberg, Germany
For the degree of
Doctor of Natural Sciences

presented by

Hanna Jacobsson, M.Sc. in Molecular Bioscience

born in Landeryd, Sweden

Oral-examination: October 11th, 2013

Functional characterization of candidate risk CNVs in lung cancer

Referees: PD Dr. Odilia Popanda
PD Dr. Angela Risch

DECLARATION

According to § 8 (3) b) and c) of the doctoral degree regulations:

a) I hereby declare that I have written the submitted dissertation myself and in this process have used no other sources or materials than those expressly indicated.

b) I hereby declare that I have not applied to be examined at any other institution,
nor have I used the dissertation in this or any other form at any other institution as an examination paper, nor submitted it to any other faculty as a dissertation.

Heidelberg,____________________ ____________________
(Hanna Jacobsson

Contributions:

Parts of this thesis (material sections 2.6.1 and 2.6.2, and methods 3.1,3.5, 3.8.4, 3.9.4, 3.11, 3.13, and parts from the result sections 4.1, 4.3.2, Figure 15, 4.3.3.1, Figure 16, 4.3.3.2, 4.3.3.3, 4.3.3.4) were used for the manuscript “Post-GWA functional characterization of copy number gain identifies methylation dependent upregulation of miR-661 in NSCLC” (Jacobsson *et al,* submitted July 2013). These sections may contain suggestions and corrections from co-authors. Additional contributors of this study are indicated in the specific sections.

To my mother

Table of Contents

Summary 15

Zusammenfassung 17

Abbreviations 19

1. Introduction 21

1.1 Cancer 21

1.2 Lung cancer 22

1.2.1 Early-onset lung cancer...22

1.2.2 Genomics and lung cancer..23

1.3 Epigenetics 24

1.3.1 The Epigenome of lung cancer..25

1.3.2 miRNA biogenesis..26

1.3.3 miRNAs and lung cancer...27

1.3.3.1 miRNAs and clinical significance in lung cancer 28

1.3.3.2 Epigenetic deregulation of miRNAs in lung cancer 28

1.4 Epithelial to mesenchymal transition (EMT) in lung cancer 29

1.5 Genetic predisposition to lung cancer...29

1.6 Germline CNVs and cancer..30

1.7 Aim of the study 33

2. Materials 35

2.1 Computer software...35

2.2 Equipment..35

2.3 Reagents 36

2.4 Commercial kits..37

2.6 Study populations 38

2.6.1 GWA study..38

2.6.2 Study population primary NSCLC and adjacent normal tissue....39

3. Methods....41

3.1 CNV detection....41

3.2 CNV analysis with TyperAssay....41

3.2.1 Competitor concentration determination....42

3.2.2 iPLEX PCR....43

3.2.3 *SAP* treatment....43

3.2.4 Primer extension reaction....44

3.3 Fluorescence *in situ* Hybridization....44

3.3.1 Lymphocyte fixation....44

3.3.2 Bioprime direct labeling....44

3.3.3 Precipitation of Bioprime labeled BAC probes....44

3.3.4 Pre-treatment of BAC probes....45

3.3.5 Pre-treatment of slides....45

3.3.6 Detection direct labeling Bioprime system (sp orange)....45

3.4 Isolation of DNA and RNA from tissue samples....45

3.5 miRNA and mRNA expression with quantitative real time PCR (Cell lines)....46

3.6 Quantitative methylation analysis with the MassCLEAVE^{TM} Assay (Sequenom)......46

3.6.1 Bisulphite conversion....46

3.6.2 BT PCR....47

3.6.3 *SAP* treatment....47

3.6.4 *In vitro* transcription and RNAse A cleavage....47

3.7 Ectopic overexpression of miRNAs in lung cancer cell lines....48

3.8 3`UTR Luciferase reporter assay....48

3.8.1 Cloning 3`UTR....48

3.8.2 Restriction enzyme digest....49

3.8.3 Ligation pMIR reporter plasmid....49

3.8.4 Co-transfection of mimic miR 661 and mimic miR 210 with pMIR reporter constructs 49

3.9 PCPGL luciferase reporter assay 50

3.9.1 Cloning PCR of the miRNA 661 promoter 50

3.9.2 Restriction enzyme digest 51

3.9.3 Ligation PCPGL plasmid 51

3.9.4 Transfection PCPGL constructs 51

3.10 Western Blot 52

3.11 TGF β1 stimulation 52

3.12 Statistical analysis....52

4. Results....55

4.1 Germline CNV detection in early onset lung cancer 55

4.2 CNV validation 58

4.2.1 Sequenom Typer assay application....58

4.2.1.1 Determination of competitor DNA concentration 58

4.2.1.2 Copy number analysis of GSTM1 60

4.2.1.3 Copy number analysis of 8q24.3 and 11p15.5....61

4.2.2 CNV analysis using FISH....63

4.2.2.1 BAC test 64

4.2.2.2 Copy number analysis in patient lymphocytes using FISH....64

4.3 Chromosome 8q24.3 67

4.3.1 Expression and methylation on chromosome 8q24.3 67

4.3.2 miRNA 661 is upregulated and hypomethylated in lung cancer....69

4.3.2.1 Methylation and clinical outcome in NSCLC 71

4.3.2.2 Top hit genes targeted by miR-661....72

4.3.2.3 Top hit target expression after ectopic overexpression of miR-661 in lung cancer cell lines 74

4.3.2.4 EMT related gene targeted by miR-661 75

4.3.2.5 MEKK3 and E-cadherin are direct targets of miR-661....76

4.3.3 miR-661 is induced during TGFβ1 induced EMT....77

4.4 Chromosome 11p15.5 78

4.4.1 Expression analysis of genes and miRNAs on chromosome 11p15.5 78

4.4.1.1 PKP3 methylation and survival analysis....80

4.4.1.2 miRNA 210 targets RUNX3 in lung cancer 82

5. Discussion 85

5.1 CNV detection on GWA data for early-onset lung cancer....85

5.2 Method optimization for copy number analysis with the TyperAssay....86

5.2.1 CNV validation with TyperAssay 87

5.3 The role of 8q24.3 in lung cancer 88

5.3.1 Aberrant expression and regulation of miR-661 88

5.3.2 Clinical relevance of miR-661 hypomethylation 89

5.3.3 miR-661 targets 89

5.4 The role of 11p15.5 in lung cancer 92

5.4.1 *Plakophilin 3 (PKP3)* 92

5.4.2 miR-210 92

6. Strengths and weaknesses of the study....95

7. Conclusion and outlook....97

References 99

Supplemental data....111

Bibliography....121

Acknowledgement....123

List of figures

Figure 1. Hallmarks of cancer. 21
Figure 2. Illustration of the epigenetic interplay in mammalian cells. 25
Figure 3. Illustration of the miRNA biogenesis..27
Figure 4. Workflow for allele specific copy number analysis using the TyperAssay 42
Figure 5. CNV detection workflow in early onset lung cancer GWA study.. 56
Figure 6. EC 50 determination for 2N control and *GSTM1*.. 59
Figure 7. Copy number analysis for *GSTM1* 60
Figure 8. 8q24.3 copy number analysis with TyperAssay. 62
Figure 9. 11p15.5 copy number analysis with TyperAssay. 63
Figure 10. Genomic location for BAC probes used in FISH. 64
Figure 11. Copy number analysis for 11p15.5 and 8q24.3 with FISH.. 65
Figure 12. Quantification of FISH results with Image J. 66
Figure 13. Comparison of volume, surface and integrated density determined by Image J 66
Figure 14. RNA expression analysis of 8q24.3 region in lung cancer.. 68
Figure 15. Demethylation at a putative promoter for miR-661 70
Figure 16. Low methylation is associated with worse overall survival. 71
Figure 17. miR-661 top hit targets are downregulated in lung cancer. 73
Figure 18. Top hit target expression after miR-661 ectopic overexpression or inhibition 74
Figure 19. *CDH1* is a target of miR-661. 75
Figure 20. miR-661 directly targets *CDH1* and *MAP3K3*. 76
Figure 21. Upregulation of miR-661 in TGF β1 stimulated A549 cells. 77
Figure 22. mRNA and miRNA expression of genes at 11p15.5 in normal (N) and tumor (T).. ...79
Figure 23. *PKP3* is deregulated by hypomethylation in lung cancer. 81
Figure 24. *RUNX3* is a target of miR-210.. 83

Figure S1. *GSTM1* copy numbers in 3 plex assay for 8q24.3. 119
Figure S2. Validation of methylation at the putative promoter of miR-661. 119
Figure S3. E-cadherin immunohistochemical staining of primary AdC, SCC and bronchial and aveolar cells..120
Figure S4. Validation of methylation at the TSS of PKP3 in sample set 3...............................120

List of tables

Table 1. Computer Softwares..35

Table 2. Equipment..35
Table 3. Reagents...36
Table 4. Commercial kits...37
Table 5. Characteristics of GWA study population 38
Table 6. Characteristics of the study populations used for lung tumor and adjacent normal tissue analysis..39
Table 7. Descriptive overview of CNVs associated with lung cancer risk detected by QuantiSNP and PennCNV. 57
Table 8. miR-661 top hit genes for five different target prediction tools 72

Table S 1 Primers for quantitative allele specific CNV analysis-Typer assay...........................111
Table S2. List of expression primers for qPCR with Roche UPL system and SYBRgreen.......111
Table S3. List of primers for BT PCR for massCLEAVE^{TM}assay 112
Table S4. List of primers for cloning..112
Table S5. CNVs detected with PennCNV...113
Table S 6 CNVs detected with QuantiSNP...115
Table S7. Overlapped CNVs between QuantiSNP and PennCNV..117
Table S8. Significant CNVs associated with lung cancer risk (QuantiSNP). 118

Summary

Environmental exposure such as tobacco smoke is the principal cause of most lung cancer cases worldwide. However, only a small proportion of heavy smokers develop lung cancer which suggests that other factors such as genetic and/or epigenetic interindividual variations may be responsible for individual disease susceptibility. The overall aim of this study was to determine germline copy number variations (CNVs) associated with early-onset lung cancer risk and to further investigate the genetic and epigenetic interplay of microRNAs (miRNAs) and genes located in two candidate CNVs on 8q24.3 and 11p15.5 in lung cancer.

A genome wide association study (GWA) had been performed using the Illumina Infinium platform Human Hap550 BeadChip on 492 early-onset lung cancer cases and 487 population based controls. Two computational CNV detection algorithms, QuantiSNP and PennCNV, were applied to this existing data set to identify CNVs and the overlapping CNVs between the two algorithms were further analyzed for association with the disease. Ten CNVs were significantly associated with early-onset lung cancer. Two CNVs were selected for strength of association and for containing miRNAs and genes likely to be relevant for lung cancer. To assess their functional relevance in non-small cell lung carcinoma (NSCLC), qPCR based expression analysis and quantitative methylation analysis using the MassCLEAVE^{TM} assay of genes and miRNAs in these regions were performed on NCSLC and matched normal lung tissue. The expression analysis showed that miR-661 on 8q24.3 was significantly upregulated in lung tumor compared to normal. The putative miR-661 promoter was hypomethylated in tumor tissue and revealed a significant negative correlation with expression in tumor. Additionally, the loss of methylation at these sites was significantly associated with worse outcome independent from stage, histology and gender. The most significant changes in the gain CNV region on 11p15.5 were seen for miR-210 and *Plakophilin 3 (PKP3)* which both were significantly upregulated in NSCLC. Promoter hypomethylation at the transcription start site of *PKP3* was inversely correlated with expression in NSCLC, suggesting that methylation regulates the *PKP3* expression. For further functional analysis of the two miRNAs, predicted targets were identified *in silico* and 3´UTR luciferase reporter assays for the predicted targets and expression analysis after ectopic overexpression in A549, H1299 and H1703 lung cancer cell lines were carried out to determine whether a direct link between the miRNAs and the targets could be shown. The results showed that mitogen associated protein 3 kinase 3 (*MAP3K3)* and Cadherin1 (*CDH1)* are direct targets of miR-661, suggesting that miR-661 has oncogenic properties in lung cancer.

Furthermore, miR-210 was shown to target the tumor suppressor gene Runt related transcription factor 3 (*RUNX3),* a transcription factor known to be involved in lung development and to be a crucial regulator of cell proliferation.

The results from this study suggest that CNV analysis of GWAs data for lung cancer risk can point to functionally important regions in the genome that are deregulated in NSCLC and may contribute to lung tumorigenesis. Further investigation of the relevance of these CNVs to early-onset lung cancer risk is needed to confirm our suggested finding of two risk markers. Furthermore, additional analyses on the functional role of miR-661 in lung cancer are desirable to elucidate to what extent this miRNA contributes to tumorigenesis. Taken together, this study provides evidence that interplay between genetic variations and epigenetic deregulation plays a pivotal role in NSCLC pathogenesis.

Zusammenfassung

Umwelteinflüsse wie Tabakrauch sind die Hauptursache für die meisten Lungenkrebsfälle weltweit. Allerdings entwickelt nur ein kleiner Anteil der starken Raucher Lungenkrebs. Andere Faktoren, wie zum Beispiel genetische und / oder epigenetische interindividuelle Unterschiede, können dabei für die individuelle Krebsanfälligkeit verantwortlich sein. Das übergeordnete Ziel dieser Studie war es, die Variation der Genkopien (CNV) in der Keimbahn zu identifizieren die mit dem Risiko, an einer frühen Form von Lungenkrebs zu erkranken, assoziiert sind. Des Weiteren sollte das genetische und epigenetische Zusammenspiel von micro RNA (miRNA) und Genen aus den CNV Kandidaten Regionen 8q24.3 und 11p15.5 untersucht werden.

Eine genomweite Assoziationsstudie (GWA) wurde unter Verwendung der Illumina Infinium Plattform Hap550 BeadChip in 492 Patienten mit Lungenkrebs in jungen Jahren und 487 Populationskontrollen durchgeführt. Zur Detektion von CNVs wurden zwei rechnergestützte CNV Algorithmen, QuantiSNP und PennCNV, auf die existierenden Daten angewendet. Die überlappendenden CNVs der beiden Algorithmen wurden hinsichtlich der Assoziation mit der Krankheit weiter untersucht. Zehn CNVs waren signifikant mit frühmanifestiertem Lungenkrebs assoziiert. Zwei dieser CNVs wurden aufgrund der Stärke der Assoziation, sowie der Tatsache dass Sie miRNAs und Gene mit wahrscheinlicher Relevanz für Lungenkrebs enthielten, ausgewählt. Um ihre funktionelle Bedeutung in nicht-kleinzelligem Lungenkarzinom (NSCLC) zu auzuklären, wurden Gene und miRNAs in diesen Regionen mittels qPCR basierter Expressionsanalyse und quantitativer Methylierungsanalyse unter Verwendung des MassCLEAVE Assays in NCSLC und normalem Lungengewebe untersucht. Die Expressionsanalyse zeigte, dass miR-661 auf Chromosom 8q24.3 in Lungentumoren verglichen mit Normalgewebe signifikant hochreguliert war. Der mutmaßliche miR-661-Promotor war im Tumorgewebe hypomethyliert und zeigte eine signifikant negative Korrelation mit der Expression. Unabhängig vom Stadium, von der Histologie oder vom Geschlecht war der Verlust der Methylierung an diesen Stellen mit deutlich schlechterem Behandlungsergebnis assoziiert. MiR-210 und *PKP3 (*engl. *Plakophilin 3*) waren beide in NSCLC-Gewebe deutlich hochreguliert und zeigten damit innerhalb der CNV Region 11p15.5 die signifikantesten Änderungen . Promotor Hypomethylierung an der Transkriptionsstartstelle von *PKP3* war invers mit der Expression in NSCLC korreliert, was darauf hinweist dass Methylierung die *PKP3* Expression regulierthinweist. Zur weiteren funktionellen Analyse der beiden miRNAs wurden deren

prognostizierte Zielstrukturen *in silico* identifiziert. In den Lungenkrebs-Zelllinien A549, H1299 und H1703 wurden 3'UTR Luciferase-Reporter-Assays und Expressions-Analysen nach ektopischer Überexpression der prognostizierten Zielstrukturen durchgeführt, um eine direkte Verbindung zwischen den miRNAs und den Zielgenen zu zeigen. Die Ergebnisse zeigten, dass MAP3K3 (engl. Mitogen associated protein 3 kinase 3) und CDH1 (engl. Cadherin1) direkte Zielstrukturen von miR-661 sind, was auf die onkogenen Eigenschaften von miR-661 bei Lungenkrebs hinweist. Des Weiteren wurde gezeigt, dass das Tumorsuppressorgen *RUNX3* (engl. *Runt related transcription factor 3*), ein Transkriptionsfaktor der Lungenentwicklung und ein wichtiger Regulator für Zellproliferation, eine Zielstruktur von miR-210 darstellt.

Die Ergebnisse dieser Studie deuten darauf hin, dass die CNV Analyse der GWAS-Daten funktionell wichtige genomische Regionen aufzeigen kann, die in NSCLC dereguliert sind, und damit zur Lungentumorigenese beitragen können,. Weitere Untersuchungen der Bedeutung dieser CNVs für das Lungenkrebsrisiko sind erforderlich, um unsere zwei Kandidaten Risiko Marker zu bestätigen. Darüber hinaus sind zusätzliche Analysen über die funktionelle Rolle von miR-661 bei Lungenkrebs wünschenswert, um herauszufinden, inwieweit diese miRNA zur Tumorentstehung beiträgt. Zusammenfassend liefert diese Studie Hinweise darauf, dass das Zusammenspiel zwischen genetischen Variationen und epigenetischer Deregulierung eine entscheidende Rolle bei der NSCLC Pathogenese spielt.

Abbreviations

µg	microgram
3´UTR	three prime untranslated region
95% C.I.	95% confidence interval
AdC	Adeno carcinoma
AGO2	ARGONAUTE 2
BAF	B allele frequency
Bp	Basepair
BT DNA	Bisulphite converted DNA
Ca	Cases
CDH1	Cadherin 1
CDK2Na	Cyclin-dependent kinase inhibitor 2A
cDNA	complementary DNA
CNV	Copy number variation
CpG	Cytosine phosphate Guanine
Cy3	Cyanine 3
Da	Dalton
DEAF1	Deformed epidermal autoregulatory factor 1 homolog
DGV	Database for genomic variants
DIRAS3	Distinct subgroup of the ras family member 3
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferase
DNMT	DNA methyltransferase
dNTP	deoxyribonucleotide
EMT	Epithelial to mesenchymal transition
EXOSC4	Exosome Component 4
FISH	Fluorescence in situ hybridization
FITC	Fluorescein isothiocyanate
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase
GAS7	Growth arrest specific 7
gDNA	genomic DNA
GPAA1	Glycosylphosphatidylinositol anchor attachment 1
GRINA	Glutamate Receptor, Ionotropic, N-Methyl D-Aspartate-Associated Protein 1
GSTM1	Glutathione S transferase mu1
GSTT1	Glutathione S-transferase theta 1
GWA	Genome wide association
H3K4me3	Histone 3 lysine 4 trimethylation
HAT	Histone acetylase transferase
HDAC	Histone deacetylase
HPRT	Hypoxanthine-guanine phosphoribosyltransferase
HR	Hazard ratio
HRAS	v-Ha-ras Harvey rat sarcoma viral oncogene homolog
IGV	Integrative genome viewer
IHC	Immunohistochemistry
IRF7	Interferon regulatory factor 7
LCLC	Large cell lung carcinoma
LRR	Log R ratio
LOH	Loss of heterozygosity
MAF1	Repressor of RNA polymerase III transcription MAF1 homolog
MAP3K3	Mitogen associated protein 3 kinase 3
MBP	Methyl binding protein
miRNA	microRNA
mut	Mutant
ng	nanogram
NSCLC	Non-small cell lung cancer
ORF	Open reading frame
PCR	Polymerase chain reaction
pg	picogram
PKP3	Plakophilin 3
PLEC1	Plectin
pre-miRNA	Precursor microRNA
pri-miRNA	Primary microRNA
PTDSS2	Phosphatidylserine Synthase 2
RIPK2	Receptor-Interacting Serine-Threonine Kinase 2
RISC	RNA-induced silencing complex
RNA	Riboxynucleic acid
RNH1	Ribonuclease/Angiogenin Inhibitor 1
RUNX3	Runt-related transcription factor 3
SCC	Squamous cell carcinoma
SCLC	Small cell lung cancer
SHARPIN	SHANK-Associated RH Domain Interactor
SIGIRR	Single immunoglobulin and toll-interleukin 1 receptor (TIR) domain
SNAIL1	Snail homolog 1
SNP	Single nucleotide polymorphism
TERT	Telomerase reverse transcriptase
TGFβ1	Transforming Growth factor β 1
TWIST1	Twist basic helix-loop-helix transcription factor 1

1. Introduction

1.1 Cancer

A cancer cell can be characterized by the classic six hallmarks of cancer that enable the cell to proliferate, resist cell death, induce angiogenesis, evade growth supressors, disseminate, invade and metastasize [1]. The understanding of the underlying genomic, epigenomic and proteomic diversity of a tumor cell behind these changes, has over the last few years increased considerably, through advanced sequencing technology. The genome of a cancer cell harbors several mutations and only a subset is thought to be causal and crucial for the cancer progression by contributing to clonal growth advantage. They are called drivers. The rest are called passengers and are defined as those that do not affect the fitness of the cell, but are acquired during the progression through e.g. genomic instability which leads to increased mutation rates and chromosomal rearrangements [2, 3]. Furthermore, it has become evident that not only the cancer cell alone but also the surrounding tumor-microenvironment, is contributing to the hallmark properties. Another characteristics of cancer involves the inflammatory state that is driven by cells of the immune system, enabling tumor progression in various ways [4]. To understand the complexity of a tumor, the hallmarks of cancer proposed by Hanahan and Weinberg in 2000 and 2011, are valuable guidelines in the search for cancer risk factors, therapeutic targets or prognostic and diagnostic markers (Figure 1).

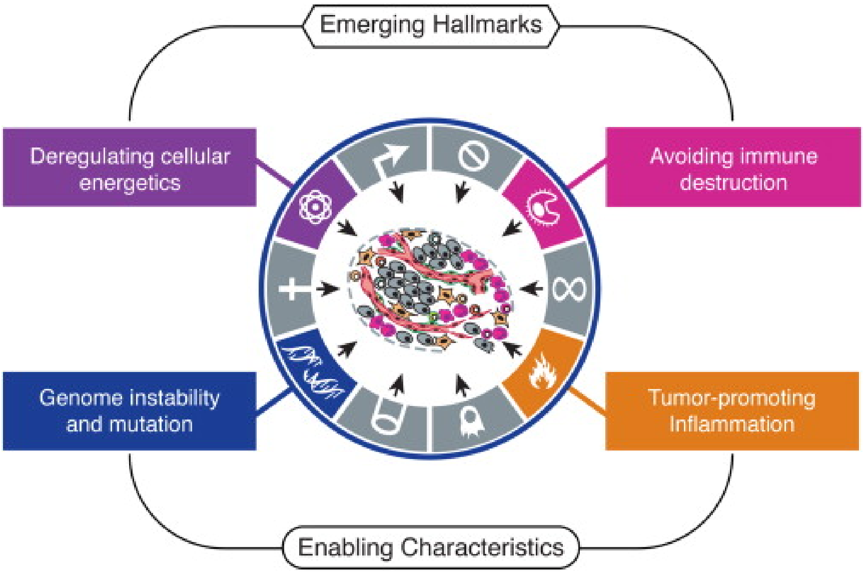

Emerging Hallmarks
Deregulating cellular energetics
Avoiding immune destruction
Genome instability and mutation
Tumor-promoting Inflammation
Enabling Characteristics

Figure 1. Hallmarks of cancer. (From the review by Hanahan and Weinberg: *Hallmarks of cancer: next generation, 2011* [4]).

1.2 Lung cancer

Lung cancer is the most prevalent cancer related death with 1.37 million deaths per year [5]. The majority of primary lung cancers are lung carcinomas and can be divided into two groups; Small cell lung carcinoma (SCLC) and Non-Small Cell Lung Carcinoma (NSCLC). SCLC is an aggressive neuroendocrine tumor consisting of small tumor cells deriving from epithelial and neuroendocrine cells. This type of lung cancer is strongly associated with smoking with a poor prognosis. Due to fast spread of these tumors, patients with SCLC are rarely operated [6]. NSCLC accounts for approximately 80% of all lung cancers and includes three histological subtypes; adenocarcinoma (AdC), squamous cell carcinoma (SCC), and large cell carcinoma (LCC)[7]. In recent years, AdC of the lung has replaced SCC as the most frequent histologic subtype for both men and women [8]. AdC arises from cells with glandular or secretary properties in the periphery of the lung [9]. The shifts in histologic types are related to increased rates of smoking in women and to modern cigarettes that contain higher concentrations of certain carcinogens [10]. Most AdC cases are linked to cigarette smoke and account for 20% of all lung cancers. Yet, among non-smokers and women, AdC accounts for most cases. SCC accounts for 30% of all lung cancers [11]. SCC originate from multilayered squamous cells, which are normally not present in the respiratory epithelium, but arise from glandular or secretory cells by metaplastic change as a result of tobacco smoke [9, 12]. NSCLC is staged from IA to IV, IA having the best prognosis and IV being the worst, based on the degree of spreading from the primary tumor [13].

1.2.1 Early-onset lung cancer

The mean onset age for lung cancer has been estimated between 60-70 years. Less than 10% of all cases develop lung cancer at an early age (younger than 51) [14]. Smoking is as well the major risk factor for this group, however, the histological type, gender distribution and genetic susceptibility have been shown to be different in early-onset lung cancer patients [15-17]. Risk studies have identified SNPs in matrix-metalloproteinase 1 (*MMP1*), Glutathione S transferase mu1 *(GSTM1), and* Cytochrome Cytochrome P 450 *(CYP450)* genes to be associated with early onset of the disease [18-20]. Additionally, risk studies in young lung cancer patients have shown an increased risk if the first degree relatives had cancer [17] or an even higher risk if the parent or a sibling was affected with lung cancer [21]. This suggests that genetic predisposition may have a stronger effect in early-onset lung cancer cases than among elderly cases.

1.2.2 Genomics and lung cancer

The traditional decision for therapy in lung cancer has been based on histology classification. Lung cancer is a molecularly heterogeneous disease, with an ever increasing understanding of genetic alterations, they have become increasingly important for treatment decisions [22]. The most common driver mutations with oncogenic features and therefore suitable as targets for therapy in AdC, appear in Epidermal growth factor receptor (*EGFR)* and Kirsten rat sarcoma viral oncogene homolog *(KRAS)* comprising between 5-15% of the cases. Other well defined genetic aberrations appearing in 5% of AdC are the Echinoderm microtubule-associated protein-like 4 (*EML4) and* anaplastic lymphoma kinase *(ALK)* fusion gene, estrogen-related receptor beta type 2 (*ERRB2), NRAS,* v-raf murine sarcoma viral oncogene homolog B1 *(BRAF),* phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha *(PIK3CA),* met proto-oncogene *(MET*) and cadherin-associated protein beta 1 (*CTNNB1)* mutations [23]'[24]. There appears to be a significant difference in the genomic landscape of SCC [25]. For example, *KRAS*, *EGFR* mutations and *ALK* rearrangements are rare in comparison to AdC and rather the *ERBB* genes, Fibroblast growth factor receptor 1 (*FGFR1),* the tyrosine kinase *DDR2* and the JAK/STAT pathway are frequently altered by mutations or amplifications in SCC. This suggests that subtype specific alterations play a crucial role in treatment decisions in lung cancer. For instance, *FGFR1* amplification and *DDR2* mutations are treatment targets specifically for SCC [26, 27]. Specific tyrosine kinase inhibitors are used in targeted treatment for patients with *EGFR* mutations, but can cause fatal complications for patients without [28]. A common feature for both AdC and SCC, is a strong correlation between smoking status and number of mutations. Smokers have a 10 fold higher mutation rate compared to non-smokers [29]. Mutations in *BRAF, JAK2, JAK3, TP53* and mismatch repair genes are strongly associated with smoking, whereas *EGFR, ROS1,* and *ALK* rearrangements appear as well in never smokers [29].

1.3 Epigenetics

Epigenetic modifications are stable marks, resulting from the covalent modification of proteins and DNA, which controls gene expression without involving a change in the DNA sequence itself (reviewed in [30]). Cells store their epigenetic information as histone modifications, DNA methylation, nucleosome positioning and non-coding RNAs. The histone modifications take place at well conserved amino acid residues located in N-termini of the histone tails and include lysine acetylation, arginine and lysine methylation and serine phosphorylation. These modifications make up a code, the histone code, which affects the chromosomal architecture and is involved in a range of nuclear processes such as gene transcription, DNA repair or DNA replication [31]. Active gene transcription involves interplay between DNA methylation, nucleosome positioning and histone modifications (Figure 2). Active promoters lack DNA methylation and have nucleosome depleted regions (NDR) upstream of the transcription start sites (TSS) [32]. The nucleosomes that flank these NDRs are marked with active histone modifications (H3K4me3, lysine acetylation and H2A.Z) which may destabilize nucleosomes to facilitate transcriptional initiation [33]. Enhancer regions also harbor histone modifications e.g. H3K4me1 and H3K27ac and deoxyribonuclease 1 (DNase 1) sensitivity and nucleosome depletions [34].

DNA methylation in mammals occurs mainly at the 5´-carbon position of cytosine at CpG dinuclotides. This epigenetic modification occurs at long stretches of CG rich sequences present in satellite repeat sequences, middle repetitive ribosomal DNA sequences, centromeric repeat sequences and CG rich sequences (CpG islands). CpG island sequences range from 200bp to 4kb in length, are found in promoter regions of almost half of the genes in the mammalian genome and are generally unmethylated in normal cells [30]. There are three DNA methyltransferases (DNMTs) present in the mammalian cells which are responsible for the *de novo* methylation during embryonic development and maintenance of the methylation pattern after replication [35]. The major role of DNA methylation is associated with transcription repression in several processes during development which require this function, such as differentiation and embryonic development, tissue specific gene expression regulation or gene silencing on the inactive X chromosome and imprinted genes. Additionally, methylation has been proposed as a genome defense against transposable elements [36]. Methyl-DNA binding proteins (MBPs) bind to methylated DNA sequences and are associated with histone deacetylases, building a bridge between the two epigenetic modifications. Moreover, hypermethylated DNA is often associated with an inactive chromatin mark, including deacetylated histone H3 and H4, H3K9 methylation and H3K27 methylation [37].

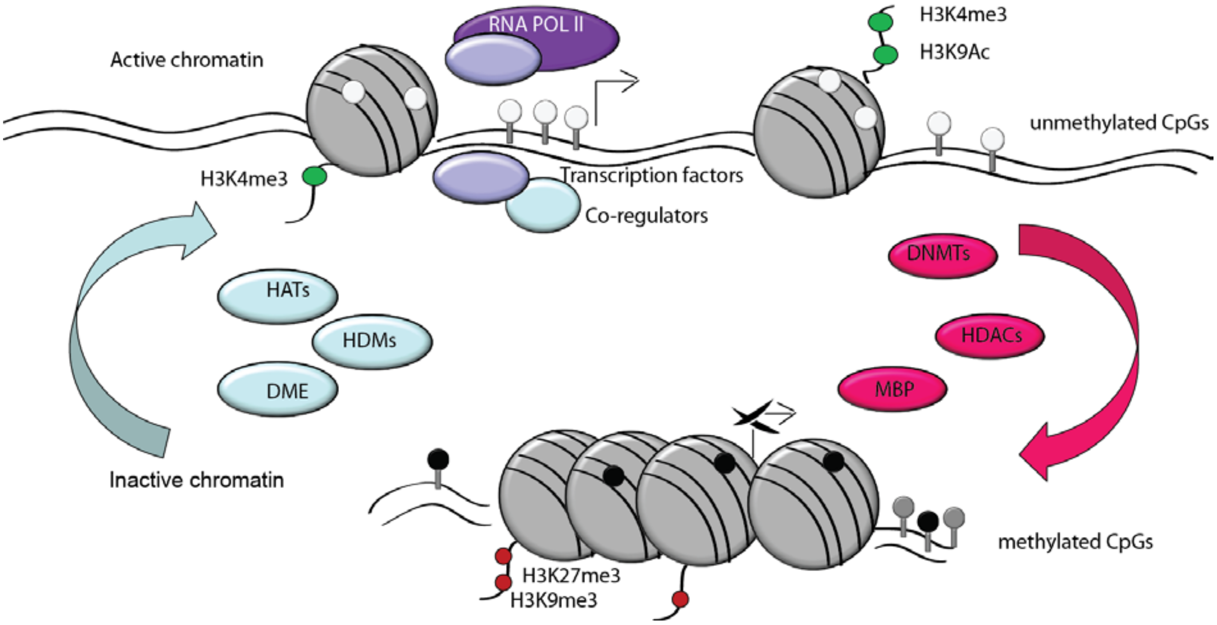


Figure 2. Illustration of the epigenetic interplay in mammalian cells (Adapted from [38]).

1.3.1 The Epigenome of lung cancer

The cancer epigenome is generally characterized by a loss of global methylation (hypomethylation) and increased methylation (hypermethylation) enriched at TSS in CpG islands or at CpG island shores [30, 39, 40]. Hypomethylation was the first known epigenetic mechanism associated with cancer development [40]. It can contribute to cancer development in several distinct ways e.g. chromosomal instability, transcriptional activation of retrotransposons, loss of imprinting and up regulation of oncogenic genes [30, 40-42]. The epithelial cell marker 14-3-3 sigma gene upregulation in NSCLC is one example where hypomethylation significantly correlates with increased expression in NSCLC and correlates with increased resistance to chemotherapy [43, 44]. Another example is the TP73 hypomethylation and its correlation with a global loss of LINE1 methylation [43]. Hypermethylation in lung cancer is associated with gene silencing of genes involved in e.g. cell cycle, DNA repair, carcinogen metabolism, cell to cell interaction, apoptosis and angiogenesis [45]. During tumorigenesis, both alleles of a tumor suppressor gene need to be inactivated by for example chromosomal deletions or loss of heterozygosity (LOH) in the coding region of a gene or hypermethylation of CpG islands located in promoter regions of the gene. Epigenetic alterations are thought to be a key pathway for long term silencing of tumor suppressor genes, and thus, can constitute the second lesion in Knudson´s two hit model of how cancer develops [46]. A good example in lung cancer for this model is cell cycle regulator *CDKN2A* (p16). The CDKN2A tumor suppressor is frequently inactivated in NSCLC by methylation (21%), mutations (18%), exon 1β skipping (4%) or homozygous deletion (29%) [25, 47-49]. Interestingly, *CDKN2A* promoter methylation has been shown to be correlated with smoking [50]. Other evidences that smoking contributes to methylation changes comes from studies in cell line systems and mouse models showing an effect on *DAP kinase* methylation [51, 52].

1.3.2 miRNA biogenesis

microRNAs were discovered for the first time in 1993 in *Caenorhabditis elegans* (*C. Elegans)* and since then the field has evolved rapidly and they are now one of the most studied molecules [53, 54]. microRNAs (miRNAs) are small, non-protein coding RNA molecules (18-24nt long) and function as endogenous inhibitors of gene functions by pairing to the 3´ untranslated region (3´UTR) of the target gene triggering either degradation of the messenger RNA (mRNA) or translational inhibition [55] (Figure 3). They are transcribed mainly by RNA polymerase II into a primary transcript that ranges from a few hundred up to 20kb or more [56],[57]. These long transcripts are characterized by a hairpin like structure and are further processed in the nucleus by RNAse III DROSHA complex which trims the primary transcript down to a precursor miRNA (pre-miRNA). An alternative pathway in the miRNA biogenesis without the DROSHA mediated cleavage, takes place during the splicing machinery. This is mainly true for intronic miRNAs (miRtrons) [58, 59]. The pre-miRNA is transported by Exportin 5 out of the nucleus to the cytoplasm where another RNAse III enzyme, DICER, cuts the hairpin loop and generates a miRNA duplex [60]. One strand acts as a guide strand incorporated in the miRNA-associated RNA induced silencing complex (RISC). The complementary passenger strand is thought to be degraded or further selected as a functional strand [61]. The mature, single stranded miRNA together with ARGONAUTE (AGO) act as guides to bring the RISC complex to its target. Preferentially, miRNAs regulate gene expression by binding to the complementary strand in the 3´UTR of the mRNA leading to mRNA degradation or translational inhibition. However, it is also now known that miRNAs can bind to 5´UTRs or open reading frames (ORF) and additionally also upregulate their targets [62].

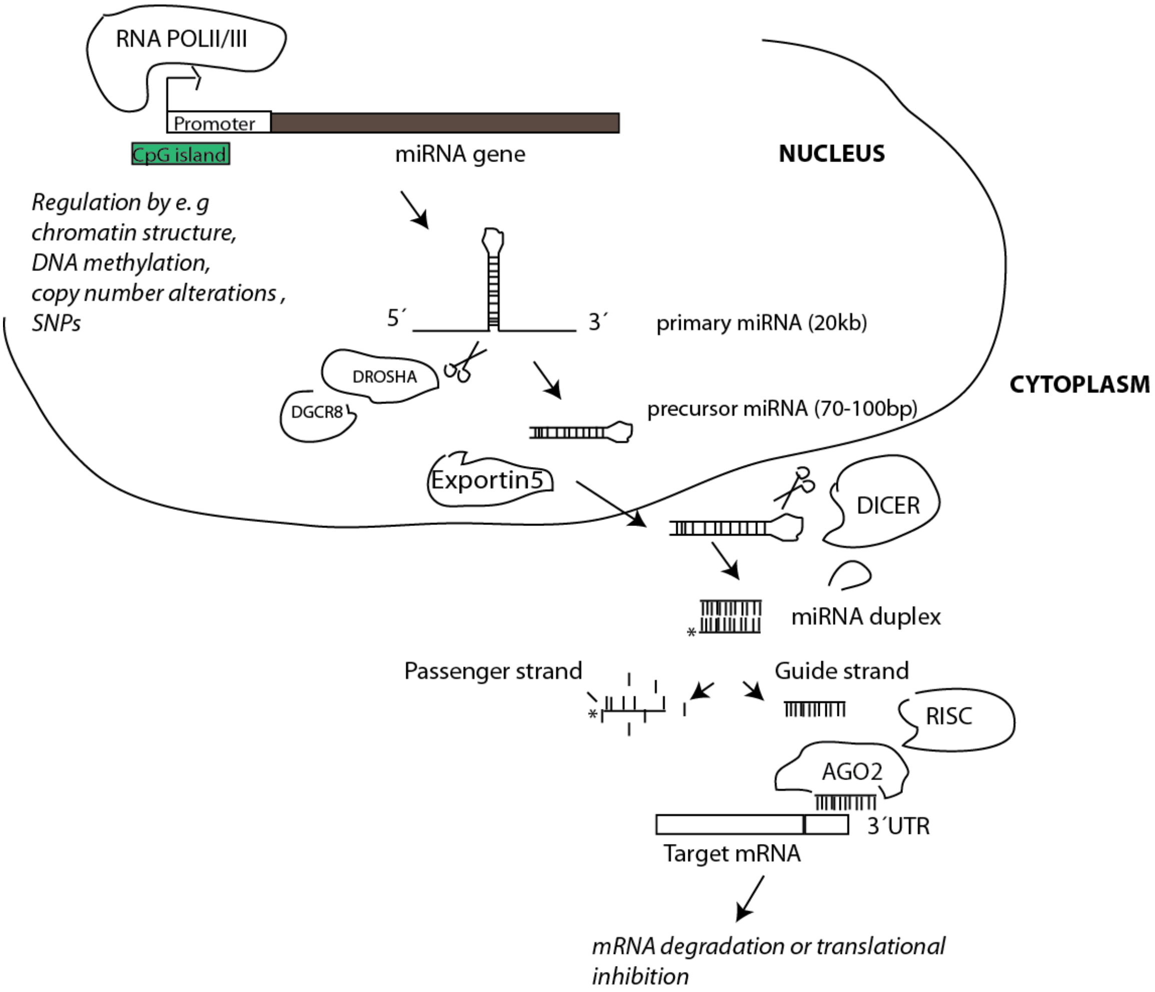


Figure 3. Illustration of the miRNA biogenesis (Adapted from [63]).

1.3.3 miRNAs and lung cancer

The main function miRNAs have in the cell is to regulate cellular processes such as development, differentiation, proliferation and apoptosis [63]. It is well recognized how important miRNAs are for the normal cell function, and it is therefore evident that if miRNA deregulation takes place, this can have a severe impact on the cellular function. In 2004, the first evidence for the role of miRNAs in lung cancer was published [64]. The study identified the miRNA let-7 expression to be correlated with post-surgery survival in NSCLC. Additionally, overexpression of let-7 in A549 cells led to inhibition of growth, suggesting a tumor suppressor function of this miRNA. Later studies on let-7 in lung cancer have strengthened this hypothesis by showing that it targets the oncogene family *RAS*, a gene family that is mutated and upregulated in the majority of lung adenocarcinomas [24, 65]. Other genes such as *CDC25a*, *CDK16* and *Cyclin D* involved in the G1/S transitions and BCL-2 involved in apoptosis, are further examples of oncogenes being regulated by let-7 [66, 67]. The first oncogenic miRNA (oncomiR) reported in lung cancer was the miR cluster mir-17-92 [68]. This cluster consists of several miRNAs located on 13q31.3, a frequently amplified region in small cell lung cancer [69]. miR-21 is another well studied oncomiR in lung cancer and other cancers. The miR 21 is upregulated by EGFR signaling in lung cancers. This miRNA has been shown to target tumorsupressor PTEN in adenocarcinoma [70]. Also, genome wide copy number variation (CNV) discovery studies have revealed that very few miRNAs are located in common CNV regions in healthy individuals [71, 72]. Marcinkowska and colleagues studied specifically the co-localization of all miRNA loci with known CNV regions and found few overlaps throughout the genome, suggesting that miRNAs are underrepresented in polymorphic regions due to their biological importance and therefore might be affected by dosage [73, 74]. The functionality of miRNAs in CNVs associated with disease risk is still poorly understood.

1.3.3.1 miRNAs and clinical significance in lung cancer

miRNAs are stable molecules and thus might be suitable as biomarkers with clinical significance. The clinical outcome of lung cancer patients could be significantly improved by using non-invasive tools e.g. biomarkers for early detection, prognosis or for treatment decisions. Several studies have shown that miRNA expression signatures measured in body fluids e.g saliva, plasma, bronchoalveolar lavage fluid or sputum can be used to distinguish subtypes of lung cancer or be used as prognostic markers. One example of this was shown in study carried out in sputum from patients and healthy controls which showed it was possible to diagnose AdC using the expression pattern of miR-486, miR-21, miR-200b and miR-375 with 80.6% sensitivity and 91.7% specificity [75]. Another study using three overexpressed miRs in lung cancer; miR-205, mir-210 and miR-708 could distinguish SCC from healthy individuals (96% specificity and 73% sensitivity) [76]. Additionally, miRNA signatures for prognostics have been detected in plasma up to 2 years before CT diagnosis [77]. Moreover, lentiviral based delivery systems have successfully been used in animal models for miRNA let 7, which makes this miRNA and others potential targets for therapeutics [78].

1.3.3.2 Epigenetic deregulation of miRNAs in lung cancer

Deregulation of miRNA expression as a consequence of altered promoter DNA methylation has recently been shown in CLL and other cancers [79-81]. In lung cancer, this has been shown for miR-886-3p for which expression suppression by promoter methylation correlates with poor outcome in SCLC [82]. Other examples are miR-193 and miR-9-3 promoter hypermethylation associated with poor survival in SCC [83]. Additionally, upregulation through hypomethylation has been reported for miRNA let 7a-3 which targets *IGF2* leading to an enhanced tumorigenic phenotype [84]. miRNAs can also themselves be involved in the epigenetic regulation by targeting components of the epigenetic machinery exemplified by the miR-29 family which targets DNMT3a and DNMT3b in lung cancer [85]. The role of epigenetic regulation of miRNAs located in CNV regions as dosage compensation is not well understood in lung cancer and should be further investigated to understand the potential genetic and epigenetic interplay.

1.4 Epithelial to mesenchymal transition (EMT) in lung cancer

Lung cancer is often diagnosed at late stage, when local invasion and metastasis already have taken place. Therefore, a better understanding of the metastatic potential and molecular mechanisms behind in lung cancer is of high importance. One of the most crucial hallmarks in cancer is the ability of a cell to evade the extracellular matrix (ECM), migrate, and invade a new site and form a metastatic lesion [1]. For this, the cell needs to undergo epithelial to mesenchymal transition (EMT), a process defined as a phenotypic change where the cells lose their intercellular junctions through loss of e.g. E-cadherin, and the apical-basal polarity and becomes migratory and invasive [86]. E-cadherin mediates cell to cell adhesive interactions and contributes to intracellular signaling by its interaction with sigma Catenin which, through GTPase activation regulate EGFR activity [86]. Downregulation of E-cadherin can be explained by DNA hypermethylation, mutations or miRNA mediated repression [87-91]. The best studied transduction pathway of EMT is the Transforming growth factor β1 (TGFβ1) signaling pathway. TGFβ1 binds to type I and type II receptors which triggers EMT through activation of SMAD2 and 3. The activated SMAD proteins relocate to the nucleus where they together with transcription factors such as SNAIL1 and 2, TWIST and ZEB, regulate the expression of target genes leading to downregulation of epithelial markers e.g. E-cadherin and upregulation of mesenchymal markers e.g. n-Cadherin, Fibronectin, Vimentin and of metalloproteinases MMP2, 3, and 9 [92].

EMT in lung cancer is less studied in comparison to other epithelial cancers e.g. breast cancer and colorectal cancer (reviewed in [93]). There is supporting evidence that invasive cells in lung cancer undergo EMT. For example, TGFβ expression has been shown to be positively correlated with lymph node metastasis in late stage and loss of E-cadherin has been shown to correlate with poor patient prognosis [90, 94] and higher risk of developing brain metastasis [95]. Additionally, TGF β-induced EMT in NSCLC leads to upregulation of SNAIL1 and TWIST. Upregulation of these proteins is associated with shorter overall survival [96-98]. The complexity of this multifaceted process, however, requires further investigation of EMT key players and their role in the progression of lung cancer.

1.5 Genetic predisposition to lung cancer

Environmental exposure such as tobacco smoke is the principal cause of most lung cancer cases worldwide. However, only a small proportion (10-15%) of heavy smokers develop lung cancer which suggests that additional underlying factors such as genetic and/or epigenetic inter-individual variations may influence the individual disease susceptibility [99]. This is supported by several genome-wide association (GWA) studies for lung cancer risk which have pointed to risk SNPs associated with lung cancer on 15q25.1, 5p15.33 and 6p21.33 [100-103]. For example the 15q25.1 locus harbors subunits of the neuronal nicotine acetylcholine receptors (nAchR) CHRNA3, CHRNA5 and CHRNB4. SNPs in the *CHRNA3* and *CHRNA5* region have also been associated with smoking addiction [104, 105]. [100-103]. Some genes within GWA-identified risk loci (e.g. *CHRNA3* and *CHRNB4* on 15q25.1 and *TERT* on 5p15.33) have been found to play a role in tumorigenesis and to be deregulated by DNA methylation [106, 107]. Other polymorphisms associated with increased risk for lung cancer includes carcinogen-metabolizing genes (*CYP1A, GSTM1, NAT2* and *MPO*), DNA repair gene *XRCC1*, inflammation related genes e.g. Matrix metalloproteases (MMPs) and cell cycle genes such as *MDM2* and *TP53* [108-110]. Moreover, miRNA polymorphisms have shown to be functionally related to disease and to be associated with poor survival [111]. Taken together, there is evidence that the genetic background plays a role in susceptibility to lung cancer by acting on crucial pathways that together with environmental factors participate in tumorigenesis.

1.6 Germline CNVs and cancer

The human population shows extensive genomic variations, consisting of both gains and losses of chromosomal regions known as copy number variations (CNVs) [71, 112-117]. CNVs are characterized as DNA segments that are 1kb or larger and present at variable copy numbers compared to a reference genome. A CNV can be simple in structure, such as tandem duplications, or may involve gains or losses of homologous sequences at multiple sites in the genome. Until recently, CNVs were thought to be a rare event in the human genome. However, population based genome wide studies have identified thousands of CNVs throughout the genome and it is now thought that CNVs encompass more total nucleotides and arise more frequently than SNPs [71, 118-120]. Several studies have shown that CNVs may contribute to the phenotypic differences between two individuals and additionally play a role in disease susceptibility by altering gene dosage, disrupt other genes or interfering with regulatory elements such as enhancer sequences or promoter regions[121-123]. However, the phenotypic variation associated with CNVs has not been evaluated thoroughly.

The role of constitutional CNVs in cancer predisposition and development is so far not well explored. One of the first studies showing that common CNVs may contribute to cancer susceptibility was investigated in patients with Li Fraumeni syndrome, a familial cancer associated with TP53 mutations [124-128]. In this study, a screen of common cancer related genes showed that 49 of these overlapped with CNV regions in more than one person in a large reference population. Based on these findings, they hypothesized that constitutional CNVs present in the germ line can predispose to cancer development and also initiate acquired CNVs in the tumor [129]. Furthermore, the Database of Genomic variants (DGVs) revealed that close to 40% of cancer related genes are interrupted by a CNV. Additionally, strong evidence that CNVs can be associated with disease risk have been shown for prostate cancer [130] and neuroblastoma [131].

CNV polymorphisms in the metabolic enzymes *GSTM1* and GSTT1 belonging to the Phase II enzymes have been intensively studied in COPD and lung cancer [132]. For example, *GSTM1* homozygous deletion has been associated with a small increased risk for lung cancer [133]. Furthermore, correlations have been shown between GST CNVs and mRNA expression in lung, suggesting that these CNVs are gene dosage dependent and may have a functional impact on the disease [134]. Also, CNVs in the Cytochrome P450 genes have been studied intensively in lung cancer [135]. Both losses and gains have been found to be associated with risk of different *CYP* genes. However, mRNA expression of the *CYP2D6* gene, has not shown concordance with numbers of copies, suggesting epigenetic mechanisms to play a crucial role in regulation of this gene family [136]. Moreover, recent studies have identified new novel functional susceptibility genes for risk and prognosis of lung cancer in the Chinese population, e.g. deletion of the mitogen associated kinase *MAPKAPK2* and gain of the *WWOX* gene [137, 138]. The impact of CNVs associated with lung cancer risk and disease progression is not well characterized and more investigations are necessary to explore how miRNAs and genes located in these regions are regulated and what functional impact they have on lung tumorigenesis.

1.6.1 CNV detection

GWAs using SNP based arrays are frequently used to identify risk loci associated with disease susceptibility. The Illumina Infinium platform Human Hap550 BeadChip has more than 500 000 SNP probes and is a hybridization based assay which uses allele specific primer extension and signal amplification for genotype calls [139, 140]. Several different CNV detection algorithms based on Hidden markov models (HMM) have been developed over the last years, making it feasible to use array based SNP data for CNV identification genome wide (reviewed in [141]). For example, PennCNV was developed to identify copy number changes by using an integrated HMM algorithm [142]. This method is based on combining the log R Ratio (LRR), the measure of normalized total signal intensity, and the B allele frequency (BAF), a measure of normalized allelic intensity ratio together with the distance between neighboring SNPs. This information is incorporated and used in a sliding window over the chromosome. Another CNV detection algorithm, QuantiSNP, is using a similar approach with the addition of the Objective Bayes HMM [143]. The Bayes factor provides a probability measure for the presence of a copy number variant in a region. The higher the Bayes factor is the stronger the evidence that the CNV exists. A combination of different algorithms is often used to increase the probability of detecting non-false positives. However, the robustness of these algorithms is not well reported. Therefore, replication studies and technical validations are important to determine whether CNVs detected from genome wide platforms are trustable. For validation, low throughput methods are used e.g. PCR based methods [144], Fluorescence *in situ* hybridization (FISH) [145] or multiplex ligation probe amplification (MLPA) [146]. The absolute quantification of copy number changes, is, however, challenging, and robust high-throughput quantitative methods are missing.

1.7 Aim of the study

GWA studies in lung cancer have suggested several risk loci to play a role in the predisposition to the disease with some evidence for functionality in lung cancer. However, larger variations such as CNVs have not been well characterized and also not the function these variants may have in the tumorigenesis of the lung. Therefore, based on a GWA study on early-onset lung cancer, the overall aim of this study was to explore the functional impact of CNVs associated with lung cancer risk on lung tumorigenesis.

The working hypotheses were:

1. There are germline CNVs existing in the genome that contribute to disease predisposition in early-onset lung cancer.
2. CNVs associated with lung cancer risk harbor genes and miRNAs that play a functional role in lung cancer progression.
3. Epigenetic analysis in lung cancer and normal tissue can contribute to a better understanding of how miRNAs and genes located in risk CNVs are regulated and could be useful to identify possible tumor suppressors or oncogenes.

The main objectives of the study were to

1. Determine novel germline CNVs associated with lung cancer risk based on a GWA study carried out on early-onset lung cancer
2. Establish and optimize a protocol for a high-throughput quantitative CNV analysis in blood.
3. Perform a technical validation of the candidate CNVs associated with lung cancer risk
4. Investigate the impact of candidate CNVs on tumorigenesis by
    a. Determination of mRNA and miRNA expression in NSCLC and adjacent normal tissue.
    b. Determination of DNA methylation as a potential expression regulator of genes and miRNAs in candidate CNV regions.
    c. Functional determination of miRNAs and genes in candidate CNVs in lung cancer cell lines to further understand how CNVs associated with risk may be functionally relevant for the disease progression.

2. Materials

2.1 Computer software

Table 1. Computer Softwares

Name	Company
Multiple Experiment Viewer software suite	L. Craig Venter Institute, San Diego, USA,
Graphpad Prism 5	GraphPad Software Inc, La Jolla, CA, USA
Typer 4.0	Sequenom, Hamburg, Germany
Epityper 1.2	Sequenom, Hamburg, Germany
ImageJ	Image/ImageJ, National Institute of Health, USA

2.2 Equipment

Table 2. Equipment

Name	Company
Transilluminator	Herolab E.A.S.Y 442
Thermocycler	Applied Biosystems
Centrifuge 5430	Eppendorf, Hamburg, Germany
Masspectrometer	Sequenom, Hamburg
LightCycler 480	Roche, Mannheim, Germany
Electrophoresis Power Supply-300	Pharmacia Biotech, Wienna, Austria
Nano drop Spectrophotometer ND-1000	peqLab Biotechnology, Erlangen, Germany
E. coli pulser	Bio RAD Labs, Munich, Germany
Thermo cycler	Eppendorf, Hamburg, Germany
Biofuge fresco	Heraeus,Hamburg, Germany
MassARRAY nanodispenser	Sequenom, Hamburg, Germany

2.3 Reagents

Table 3. Reagents

Name	Company
Shrimp Alkaline Phosphatase	Sequenom, Hamburg, Germany
10x Buffer	Sigma, Hamburg, Germany
HOT STAR taq polymerase	Qiagen, Hilden, Germany
Poly Acrylamide	Carl Roth GmbH, Karlsruhe, Germany
Tris-Borat-EDTA	Carl Roth GmbH, Karlsruhe, Germany
Ammonium persulphate	Carl Roth GmbH, Karlsruhe, Germany
TEMED (N´tetramethylethylenediamine)	Carl Roth GmbH, Karlsruhe, Germany
Transcleave mix:	Sequenom, Hamburg, Germany
RNAse A	Sequenom Hamburg, Germany
T cleavage mix	Sequenom Hamburg, Germany
DTT	Sequenom Hamburg, Germany
T7 RNA polymerase	Sequenom Hamburg, Germany
Resin	Sequenom Hamburg, Germany
RNAse/DNAse free water	Sequenom Hamburg, Germany
384-well plate	Thermo Fischer Scientific, Waldorf, Germany
Pipettes	Matrix, Gilson, Lab Systems
Pipette tips	TipOne, Ahrensburg, Germany
Ampicillin	GE Health care, Neu-Isenburg, Germany
TOP10 chemocompetent cells	Invitrogen, Karlsruhe, Germany
pCpGL vector	Gift from Michael Rehli
pMIR luciferase vector	Qiagen, Hilden, Germany
T4 ligase	New England Biolabs, Ipswich, USA
T4 ligation buffer	New England Biolabs, Ipswich, USA
Zeocin	Invitrogen, Karlsruhe, Germany
Bacto TM AGAR	Becton, Dickinson and Co., New Jersey, USA
Bacto TM Peptone	Becton, Dickinson and Co., New Jersey, USA
Bacto TM Yeast Extract	Becton, Dickinson and Co., New Jersey, USA
Millipore water	GIBCO, Invitrogen, Karlsruhe, Germany
peqGold Universal Agarose	peqLab Biotechnology, Erlangen, Germany
6xLoading dye	Fermentas; Leon-Rot, Germany

HF and GC buffer	Thermo Fischer Scientific, Waldorf, Germany
Pfu tag polymerase	Thermo Fischer Scientific, Waldorf, Germany
RPMI 1640 medium	GIBCO, Invitrogen, Karlsruhe, Germany
F12 Kaighn´s L-glutamine	GIBCO, Invitrogen, Karlsruhe, Germany
Dulbecco´s Phosphate Buffered Saline	GIBCO, Invitrogen, Karlsruhe, Germany
Trypsin EDTA	GIBCO, Invitrogen, Karlsruhe, Germany
TRANS-IT	MirusBio. Madison, WI, USA

2.4 Commercial kits

Table 4. Commercial kits

Name	Company
QIAMP DNA mini kit	Qiagen, Hilden, Germany
EZ DNA Methylation kit	Zymo Research, Orange CA, USA
QIAquick PCR purification kit	Qiagen, Hilden, Germany
MassCLEAVE^{TM} Kit-T7	Sequenom Hamburg, Germany
Spectro CHIP Arrays and Clean Resin Kit	Sequenom Hamburg, Germany
REPLI-g Mini Kit	Qiagen, Hilden, Germany
HI SPEED Midi kit	Qiagen, Hilden, Germany
miRscript Assay	Qiagen, Hilden, Germany
RNeasy kit II	Qiagen, Hilden, Germany

2.6 Study populations

2.6.1 GWA study

A GWA study in early-onset lung cancer had previously been carried out in a collaborative project between DKFZ Heidelberg, Helmholtz center Munich (H.-E. Wichmann, J. Heinrich) and Göttingen University (H. Bickeböller, A. Rosenberger). Genome-wide SNP analysis was performed on 492 early-onset lung cancer cases and 488 population-based controls (Illumina, 550K) of the Helmholtz-Gemeinschaft Deutscher Forschungszentrum (HGF) lung cancer GWA study [19] including lung cancer cases diagnosed at ≤ 50 years from the Lung Cancer in the Young (LUCY) study [14], a multicenter study within 31 German hospitals, and the Heidelberg lung cancer study, a hospital-based case-control study conducted by the German Cancer Research Center (DKFZ). Controls were selected from the Cooperative Health Research in the Region of Augsburg [KORA][147]).The characteristics of the study population are presented in Table 5.

Table 5. Characteristics of GWA study population

Subjects	**Controls (n=488)**	**Cases (n=492)**
Age at recruitment/diagnosis	45.6 ± 3.7	45.6 ± 3.7
Gender		
Female	240 (49%)	187 (38%)
Male	248 (51%)	306 (62%)
Pack years (py)	11.1 ± 18.4	30.2 ± 19.6
Smoking status		
Never	218 (45%)	38 (8%)
Ex-smokers	141 (29%)	84 (17%)
Current	129 (26%)	370 (75%)
Histology		
SCC		98 (20%)
AdC		178 (36%)
SCLC		113 (23%)
LCLC		19 (4%)
Mixed type		1 (0%)
NSCLC mixed type		33 (7%)
Other		17 (3%)
Unknown		33 (7%)

2.6.2 Study population primary NSCLC and adjacent normal tissue

Tumor and distant matched unaffected lung tissue were received from NSCLC patients at the Thoraxklinik at Heidelberg University hospital, Germany. Patients gave their consent for the use of their resected tissue for the study. Tissues were snap-frozen within 30 minutes after resection and stored at -80°C until the time of analysis. Tumor histology was classified according to the 3rd edition of the World Health Organization classification system [148]. The study protocol was approved by the local Ethics Committee of the Heidelberg University, Germany (270/2001; 199/2001, 186/1996, 201/1998). Characteristics of study populations used for lung tumor and adjacent normal tissue are presented in Table 6.

Table 6. Characteristics of the study populations used for lung tumor and adjacent normal tissue analysis

Sample set	***set 1***		***set 2***		***set 3***	
	SCC	AdC	SCC	AdC	SCC	AdC
Subjects	(n=19)	(n=18)	(n=24)	(n=24)	(n=23)	(n=20)
Age at diagnosis	61±13	64±10	66±10	62±9	61±10	60±10
Gender						
Female	6 (32%)	5 (45%)	7(39%)	11 (61%)	4(44%)	5 (55%)
Male	13 68%)	13 (50%)	17(59%)	12(41%)	19(56%)	15 (44%)
Pack years (py)						
	42±24	40±22	40±20	28±24	41±21	40±21
Smoking status						
Never	0 (0%)	3 (17%)	1(4%)	5(16%)		
Ex-smokers	14 (74%)	9 (50%)	9(38%)	11(44%)	9 (20%)	8 (19%)
Current	5 (26%)	5 (28%)	14(58%)	10(40%)	14 (33%)	12 (28%)
Unkown		1 (5%)				
Stage						
I	7 (36%)	9 (50%)	10 (41%)	9 (38%)	3 (7%)	3 (7%)
II	6 (32%)	2 (11%)	9 (38%)	10 (41%)	13 (30%)	3 (7%)
III	6 (32%)	7 (39%)	5 (21%)	5 (21%)	5 (12%)	11 (26%)
IV					2 (5%)	3 (7%)

3. Methods

3.1 CNV detection

The CNV detection was carried out by Agnes Hotz-Wagenblatt (Division of Bioinformatics, Genomics and Proteomics, Core Facility, DKFZ, Heidelberg). CNV detection was carried out with PennCNV and QuantiSNP according to recommendations in reference [142] and [143], respectively. In brief, for PennCNV, the input signal intensity files were prepared with BeadStudio from the original Illumina data as described by PennCNV (http://www.openbioinformatics.org/PennCNV/PennCNV_input.html#_Toc214852004). After splitting, the CNV detection program of PennCNV was run with the data files hh550.hg18.pfb, hh550.hmm, and hh550.hg18.gcmodel. Stringent cut offs were applied with a minimum of 7 SNPs having a log R ratio (LRR) > 0.25 for duplications and < 0.25 for deletions. To combine the CNVs of the cases and the controls, perl scripts calculated the minimal overlaps for overlapping CNVs, number of cases and controls, and filtered for CNVs detected in more than 10 samples (either cases or controls). QuantiSNP was run according to the manual with gc corrections. A perl script then filtered the QuantiSNP output files according to SNP number >= 7 and LRR std < 0.25. Further analysis for the combination of the CNVs was accomplished as described for PennCNV.

3.2 CNV analysis with TyperAssay

For quantitative, allele specific copy number analysis, a protocol for the TyperAssay application using the MassARRAY platform from Sequenom was optimized [149]. Absolute quantification for CNV analysis combines a multi-plex competitive PCR with the iPLEX primer extension reaction provided by Sequenom, followed by detection by MALDI-TOF and allele ratio analysis (Figure 4). In order to analyze allele specific and absolute copy number variation the assay was designed for the wild-type (wt) allele at an informative SNP representing the genomic DNA (gDNA), and a mutant allele which represents the competitor DNA that serves as an internal standard. A known concentration (copy number) of competitor DNA was added to the reaction as a template, and by comparing the ratio of competitor allele to wt allele it is possible to determine the absolute copy number of the wt allele. A 2N control was included in the same plex for intra-assay normalization standard to compare sample to sample loading variation and regions of interest against a known diploid control.

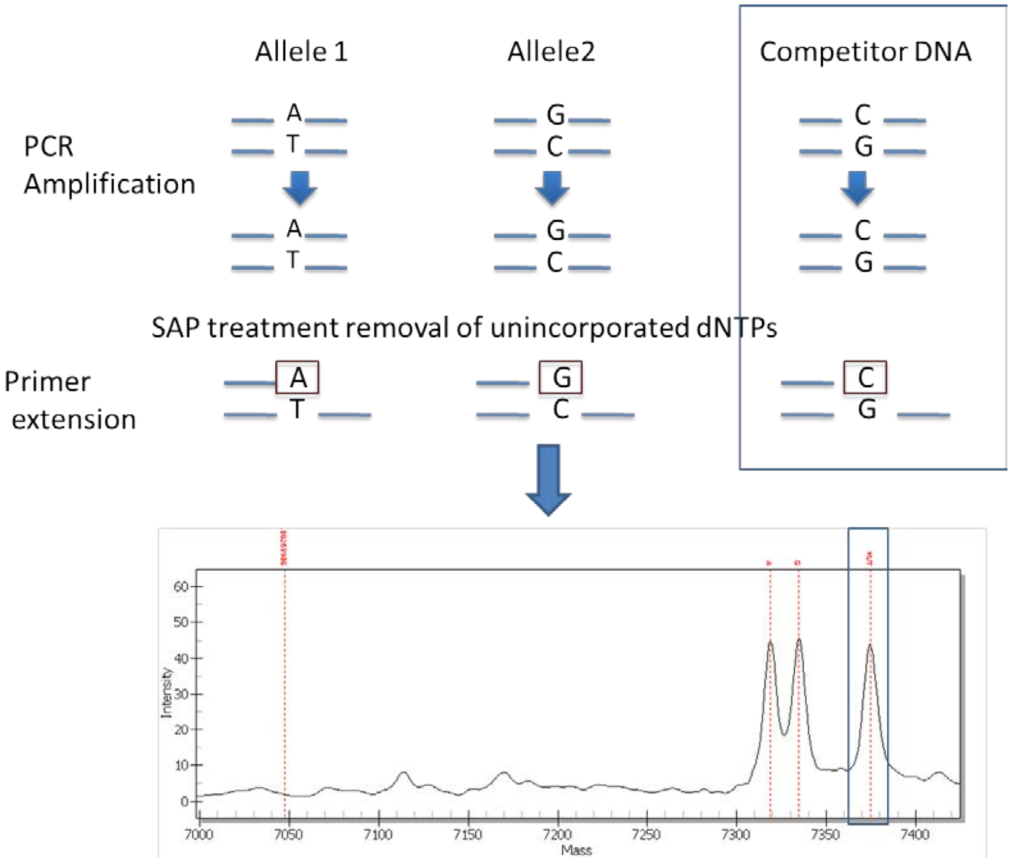


Figure 4. Workflow for allele specific copy number analysis using the TyperAssay application (Sequenom). PCR amplification is carried out to amplify a region of interest e.g. a SNP to be able to distinguish two alleles. In the same reaction, a co-amplification of a synthetic DNA identical to the target region except for the SNP base is carried out to be able to quantify the absolute copies in the reaction. After *SAP* treatment (see section 3.2.3), a primer extension step is performed to extend the PCR product with one base which distinguishes the alleles from each other in the masspectrometry.

3.2.1 Competitor concentration determination

A known concentration (copy number) of competitor DNA was included in the assay to serve as an internal control. To obtain the concentration of the competitor DNA representing the same amount of DNA molecules in the reaction as the gDNA, a serial dilution with competitor together with a fixed amount of DNA template (copies of molecules) of gDNA was analyzed. For each genomic region of interest, we used a competitor in the iPLEX reaction. Competitor PCR is a method that was established for quantification of mRNA by spiking a known molar quantity of a synthetic DNA template identical to the gene of interest that competes with the mRNA PCR amplification[150]. The amount of haploid copies in 20ng gDNA is 6210 (Equation 1). The theoretical concentration of competitor DNA needed to equal 6210 copies is 2.06E-15 M (Equation 2).

Equation 1. Calculation for absolute copies of 20ng gDNA

Size (bp) of 1 haploid genome:

3.15*10^9 bp [151]

Weight (pg) of 1 nucleotide pair:

1.02310^$^{-}$9 pg

Size (bp) of 1picogram (pg) of DNA:

0.978x10^9 bp

Absolute copies of 20ng DNA:

(0.978*10^9)*20000pg/3.15*10^9 bp= **6210** copies

Equation 2. Concentration for competitor DNA that equals 20ng DNA

Avogadro's constant:

6.0123*10^23 mol $^{-1}$

Amount of substance (n):

6210/ 6.0123*10^23 = **1.03122^-20 mol**

Concentration needed for iPLEX PCR in a final volume 5µl:

1.03122^-20 mol / 5*10^-6 l = **2.06244E-15 M**

3.2.2 iPLEX PCR

Target regions were amplified using primers listed in Table S1. The PCR reagents were purchased from Qiagen, Hilden, Germany. Multiplex PCRs were carried out in a final volume of 5µl using 20ng of template gDNA and competitor DNA amount equal to input gDNA (see section 4.2.1) containing 1X PCR buffer with 2nM MgCl$_2$, 500µM dNTP mix, 100nM primermix and 0,5U Hot Star Taq. Thermal cycling was performed with following conditions: Initial pre-incubation at 95°C for 15 min followed by 35 cycles of denaturation at 94°C for 20 s, annealing at 56°C for 30 s, elongation at 72°C for 1min and a final elongation step at 72°C for 3 min. 1µl of PCR product was run on a 3% agarose gel containing 5µg/ml Ethidium Bromide. A 100bp ladder (Fermentas, St-Leon Rot, Germany) was used to determine the size of the PCR products.

3.2.3 *SAP* treatment

For dephosphorylation of un-incorporated dNTPs, Shrimp Alkaline Phosphatase (*SAP*) treatment was carried out in a final volume of 2µl with 0.255U *SAP* and 1X *SAP* buffer and added to 5µl PCR product and incubated at 37°C for 40 min followed by deactivation of the enzyme at 85°C for 5 min.

3.2.4 Primer extension reaction

To determine the two alleles of a SNP with MALDI-TOF, an extension primer reaction was carried out with an extension primer using the PCR product. The extension primers were designed to bind adjacent to a SNP to extend with one base dependent on the genotype. The mass shift dependent on the nucleotide was analyzed with MALDI-TOF using the Typer software. The extension reaction was carried out in a final volume of 9µl with 0.222X iPLEX plus buffer, 0.5 x iPLEX termination mix and 0.625µM or 1.25µM extension primer mix and 0.5x iPLEX enzyme. The extension primer mix concentration was determined by high or low mass. Thermal cycling was performed under following conditions: Initial denaturation at 94°C followed by 40 cycles including 94°C for 5 s and 5 cycle steps at 52°C for 5s and 80°C for 5s followed by elongation at 72°C for 3 min.

3.3 Fluorescence *in situ* Hybridization

3.3.1 Lymphocyte fixation

The lymphocytes were washed 3x with 1xPBS at 2400rpm for 10min. To preserve the cell morphology while permeabilizing the cells for labeled oligonucleotides, 2 ml of Fixative (3 parts Methanol and 1 part Acetic Acid) was added dropwise while vortexing. Hypotonic treatment was done with KCL 0.075M and added dropwise while vortexing to a final volume of 12ml. The samples were incubated for 20-30min at 37°C. After incubation, samples were centrifuged for 10min at 2400rpm and the supernatant was removed. Fixative was added dropwise by vortexing and centrifuged for 10min at 2400rpm. The fixation step was repeated one time followed by centrifugation for 10min at 2400rpm.

3.3.2 Bioprime direct labeling

20µl 2.5x Random primer solution (Invitrogen, Karlsruhe, Germany) was added to 300-500ng BAC DNA in 10µl H_2O and incubated for 5min at 100°C for denaturation and directly followed by incubation on ice. The denatured BAC DNA was incubated overnight at 37°C with 5mM dACG, 5mM dTTP, 1mM spectrum Orange-UTP, 40U/µl KLenow (Klenow Fragment in 50mM Potassium Phosphate pH 7.0, 100 M KCl, 1mM DTT, 50% Glycerol) in a total volume of 50µl. The size of the DNA was determined on 1.2% agarose gel and treated with DNAse I for 30-60min at 15°C with 1:50 DNAse I (1.5µl) and 10x NT buffer (500 mM Tris, pH 7.5, 100 mM MgCl2, 10 mM DTT, 0.5 mg/ml BSA) in a final volume of 60µl (48µl labeled BAC, 6µl 10x NT buffer and 1.5 µl DNASe and H_2O) to obtain DNA fragments with the size 500-700bp. 5µl Stopmix was added after treatment to stop the amplification reaction.

3.3.3 Precipitation of Bioprime labeled BAC probes

Bioprime labeled BACs were precipitated together with 5µl salmon sperm, 30µl Cot DNA in a 2.5X final volume of EtOH (100%) and incubated overnight at -20 °C.

3.3.4 Pre-treatment of BAC probes

Precipitated BAC probes were spun down at 13000rpm at 4°C for 30 min. The supernatant was removed and the pellet was air-dried at room temperature. The pellet was dissolved in 2.5µl of deionized formamide at 37°C on a shaker. 20% dextransulfate in 2x SSC was added to a final volume of 5µl. The probes were denatured at 73°C for 5 min followed by 30 min incubation at 37°C for pre-annealing to allow cot DNA to bind repetitive elements.

3.3.5 Pre-treatment of slides

Slides were equilibrated in 2x SSC followed by dropwise adding 150µl RNAse-solution (RNAse 20mg/ml in 10mM Tris-Hcl pH 7.5, 15mM NaCl diluted 1:200 in 2x SCC) to the slides and incubated upside down in a humid chamber for 1h at 37°C. The RNAse treated slides were washed 3x 5min in 2x SSC on a shaker at room temperature followed by pepsin treatment (60µl pepsin in 10mM 80ml HCl) for 8-10 min depending on the quality of the slides. To remove the HCl solution, slides were washed 2x 5min in 1x PBS on a shaker followed by 1x5min wash in 1xPBS plus 50mM $MgCl_2$. To preserve the cell structure, slides were incubated in PBS-$MgCl_2$-Formaldehyd (1%) for 10min and sub-sequentially washed in 1x PBS for 5 min. To dry the slides 70, 90 and 100% subsequent wash was carried out for 3 min in each ethanol concentration and afterwards air-dried. 100 µl denaturation mix (4.9mM HCl, 1.6XSSC, formamide) was added drop wise on the slides and covered by a cover glass. Denaturation was carried out at 73°C for 1.45 min and directly transferred to ice cold 70% EtOH and incubated for 3 min followed by 3min wash in 90 and 100% EtOH subsequently. 5µl of pre-annealed precipitated BAC probes (described in chapter 4.3.4) was added to the slides and covered with 15x15mm^2 cover glass and sealed with fixogum. The hybridization was carried out over night at 37°C in a dry chamber.

3.3.6 Detection direct labeling Bioprime system (sp orange)

Slides were washed in 2x SSC for 10min at room temp or 3x SSC a 5min at room temperature followed by 2x 7 min wash in 0.2 x SSC at 52 °C. The slides were then shortly washed in 4x SSC/Tween 20% before DAPI staining for 2-5min. The DAPI stained slides were washed in water and thereafter air dried. 2 drops of anti-fade was added on the coverslips and slides were stored at 4°C in dark.

3.4 Isolation of DNA and RNA from tissue samples

DNA and RNA isolation from tissue samples were carried out by Michael Meister, Thoraxklinik Heidelberg with the following protocol: for nucleic acid isolation 10 – 15 tumor cryo sections (10 – 15 µm each) were prepared for each patient. Only samples with a viable tumor content of ≥50% were used for subsequent analyses. Matched tumor free lung tissues were removed distant to the tumor and macroscopically reviewed to be devoid of tumor. Frozen cryo sections were homogenized with the TissueLyser mixer-mill disruptor (Qiagen, Hilden, Germany) and normal lung tissues were homogenized using a Miccra D-8 rotor–stator homogenizer with DS-5/K1 (Art-moderne Labortechnik, Muelheim, Germany). Genomic DNA and total RNA were isolated using an AllPrep DNA/RNA Kit (Qiagen, Hilden, Germany) according to the manufacturer's instruction with following modifications: the flow-through from DNA spin columns was applied onto gDNA eliminator column (Qiagen, Hilden, Germany) and 1.5 volume of 100% ethanol was added to the flow-through before RNA isolation. Buffer RW1 was replaced with RWT buffer (Qiagen, Hilden, Germany). DNA and RNA quantification was carried out with a NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies, Wilmington, MA, USA). The quality of total RNA was assessed with an Agilent 2100 Bioanalyzer and Agilent RNA 6000 Nano Kit (Agilent Technologies, Boeblingen, Germany). RNA was considered sufficient for further analyses if it had an RNA integrity number (RIN) of at least 8.0.

3.5 miRNA and mRNA expression with quantitative real time PCR (Cell lines)

mRNA and miRNA isolation was carried out with RNeasy PLUS KIT II (Qiagen, Hilden, Germany) with following modifications; 350µl Buffer RLT Plus was added to the samples and mixed by vortexing for 1 min. The homogenate was transferred to the gDNA eliminator spin column and centrifuged for 30s at 8000xg. 1.5 volumes 100% ethanol was mixed with the flow-through and transferred to RNeasy Mini spin column and centrifuged for 15s at 8000xg. cDNA synthesis was carried out on 0.5 or 1µg RNA using miScript II RT kit from Qiagen according to the manufacturer´s recommendation with following modifications; The miScript HiFlex buffer was applied to synthesize both miRNA and mRNA. The cDNA synthesis was diluted in RNAse free water to 1:20 for subsequent expression analysis. miRNA expression was carried out in duplicates (for tissue samples) and in triplicates (cell lines) with miScript Sybr Green PCR kit (Qiagen, Hilden, Germany) according to manufacturer´s recommendations. SCARNA17, and RNU6B, and SNORD25 expression were used for normalization. mRNA expression was carried out in duplicates with Roche Universal probe library system according to manufacturer´s recommendations. *GAPDH* and *HPRT* housekeeping genes were used for normalization.

3.6 Quantitative methylation analysis with the MassCLEAVE^{TM} Assay (Sequenom)

3.6.1 Bisulphite conversion

DNA polymerase cannot distinguish between methylated and unmethylated cytosines in the genomic sequence and therefore the methylation patterns are lost after the PCR amplification. To preserve methylation patterns, genomic DNA was treated with sodium bisulfite, leading to deamination of unmethylated cytosine to uracil. The procedure includes sulfonation at the carbon 6-position of cytosine followed by irreversible hydrolytic deamination at the carbon 4-positions to generate uracil sulfonate. The final step includes a subsequent desulfonation step under alkaline conditions to generate uracil. The rate of deamination of 5-methylcytosine is much slower and it is assumed that any remaining cytosines after this treatment were originally methylated [152]. Sodium Bisulfite treatment (BT) was carried out with the EZ DNA Methylation KIT (ZYMO research, Freiburg, Germany). In brief, 0.5µg or 1µg of genomic DNA was diluted with ddH_2O to a total volume of 45µl and 5µl of M-Dilution buffer in eight strip tubes (Thermo Fischer Scientific, Karlsruhe, Germany). The DNA was denatured to single strands by the addition of sodium hydroxide and incubated for 15 min at 37°C. For deamination by sodium bisulfite, 100 µl of conversion reagent was added to the samples and incubated in the dark for 16 h at 50°C followed by incubation at 4°C for 10 min. Samples were transferred onto a column with the addition of 400 µl M-Binding buffer. The columns were centrifuged at 3000 g for 5 min and washed with 500 µl M-wash buffer. For desulphonation, 200 µl of M desulphonation buffer was added and columns were incubated at room temperature for 15-20 min followed by two additional washing steps with 500 µl M- wash buffer for removal of bisulfite salts and other interfering chemicals. The DNA was eluted two times with 30 µl Elution Buffer.

3.6.2 BT PCR

The PCR reaction was conducted in a 5µl total volume reaction containing 1µl BT DNA, 200mM forward and reverse primers, 200nM dNTPs, 1x PCR Buffer and 0.2 U HOTSTAR polymerase. The PCR reaction was carried out in a Thermocycler (Eppendorf, Hamburg, Germany) with following conditions: Initial pre-incubation at 95°C for 15min, 45 cycles of denaturation at 95°C for 30 s followed by annealing at 54-60°C for 30 s and elongation at 72°C for 1min followed by an additional extension step at 72°C for 7 min. 1µl PCR product for all samples was subjected to agarose gel electrophoresis with 2% Borat agarose gel. Gels were run at 150V for 40min and stained for 5 min in an ethidium bromide solution (5µg/ml). A 100bp ladder (Fermentas, St Leon-Rot, Germany) was used to determine the size of the PCR products.

3.6.3 *SAP* treatment

For dephosphorylation of un-incorporated dNTPs in the PCR reactions, *SAP* treatment was carried in a final volume of 2µl using 0.3 U *SAP* and added to 5µl PCR product and incubated at 37°C for 20 min. The enzyme was deactivated in a subsequent heat inactivation for 5 min at 85°C.

3.6.4 *In vitro* transcription and RNAse A cleavage

During *in vitro* transcription, T7 polymerase starts to transcribe the double stranded BT DNA from the T7 promoter tag on the reverse primer sequence resulting in single stranded RNA. RNase A specifically cleaves single-stranded RNA at 3’ of every rCTP and rUTP residues.

The assay uses a specific rNTP/dNTP nucleotide mixture together with T7 polymerase, which allows incorporation of both types of nucleotides. Because the nucleotide mixture contains dCTP (instead of rCTP), rUTP, rGTP and rATP, this results in unique ‘T-cleavage’ of only rUTP by RNaseA. Methylation events are identified as a G-A sequence change (representing C-T differences at the DNA level introduced during BT), which leads to a 16Da mass shift of the cleaved products. 2µl *SAP* treated PCR was applied on 384 well plate (Thermo Fischer Scientific, Karlsruhe, Germany), and used as a template in a 7 µl transcription reaction, containing 3.14mM DTT, 2.5mM dCTP, 1mM rUTP, 1mM rGTP, 1mM rATP, 20U T7 R and DNA polymerase and 0.9 mg/ml RNAse A in 0.64 x T7 polymerase buffer. Transcription and digestion were performed simultaneously at 37°C for 3 h. Samples were diluted in 20µl ddH_20 and 6mg of an ion-exchange CLEAN Resin (Sequenom) was added to the samples to condition the phosphate backbone of nucleic acid fragments for the MALDI-TOF analysis. 22µl of cleavage reaction was dispensed automatically onto silicon matrix preloaded chips (SpectroCHIPs; Sequenom). The mass spectra were analysed using MassARRAY mass spectrometer (Sequenom). The spectra’s methylation ratios were generated with MassARRAY v1.2 software (Sequenom). The software generated quantitative results for each cleaved CpG site or an aggregate of multiple CpG sites.

3.7 Ectopic overexpression of miRNAs in lung cancer cell lines

A549 lung cancer cells were seeded in F12K medium (Gibco, Invitrogen, Freiburg, Germany) with 10% fetal bovine serum (FBS) and transfected using Dharmafect transfection agent (Dharmacon) with either 5nM Syn-hsa-miR-661 miScript miRNA mimic, or 5nM miR mimic control Allstar (Qiagen, Hilden, Germany) for 24, 48 and 72 hours on a 6 well plate or 10cm culture plates. H1299 and H1703 lung cancer cell lines were seeded in RPMI 1640 medium (Gibco, Invitrogen, Freiburg, Germany) with 10% FBS and transfection was carried out as in A549 cells.

3.8 3`UTR Luciferase reporter assay

3.8.1 Cloning 3`UTR

PCR amplification was conducted using *Pfu* taq DNA polymerase (Thermo Fischer Scientific, Karlsruhe, Germany). The *Pfu* enzyme allows for high fidelity PCR with fewer errors compared to other thermostable polymerases and processes blunt-ended products. The PCR was carried out with 100 ng genomic DNA in a 20 µl total volume reaction containing 1x GC or High-Fidelity buffer (Thermo Fischer Scientific, Karlsruhe, Germany), 400nM dNTP mixture (Invitrogen, Karlsruhe, Germany), 200mM of forward and reverse primers (Sigma Aldrich, Munich, Germany), 3% DMSO, and 0.2 U *Pfu* HF (Thermo Fischer Scientific, Karlsruhe, Germany). Thermal cycling was carried out in Master Cycler Gradient (Eppendorf, Hamburg, Germany) with the following conditions: initial pre-incubation at 98°C for 30 s followed by a second denaturation step at 98°C for 10 s, 40 cycles of annealing temperature 56°C or 58°C for 30s followed by elongation at 72°C for 1 min and an additional elongation step at 72°C for 8 min. The PCR was performed with forward and reverse primers listed in Table S2**.** The PCR products were run with 6x Loadingbuffer (Fermentas, St Leon-Rot, Germany) and separated by electrophoresis on 1.0 % 1X borate agarose gel. A 1kb and 100bp size ladders were run on the same gel to determine the size of the PCR product. The correct fragments were cut out from the gel and purified with Qiaquick PCR purification kit (Qiagen, Hilden, Germany) according to manufacturer´s protocol. To verify the correct size of the DNA fragment, 200ng was separated on a 1.2% 1X borate agarose gel and photographed.

3.8.2 Restriction enzyme digest

The PCR product (500ng) and the pMIR report vector (1µg) was digested with *MluI* and *HindIII* in a final volume of 30µl containing 1x NEB buffer, 1% BSA, 0.5U *SAP* and 0.2U *MluI* and *HindIII* ,respectively, for 3h at 37°C. The digested product was run on a 1% 1X borate agarose gel and the correct product was cut out and purified with Qiaquick PCR purification kit (Qiagen, Hilden, Germany) according to the manufacturer´s protocol.

3.8.3 Ligation pMIR reporter plasmid

MluI/HindIII digested DNA was ligated with pMIRreport vector in a 10 µl total reaction volume containing 50ng pMIR report vector, 5:1 insert-vector, 1x T4 buffer and 0.5 U *T4* ligase (New England Biolab) at room temperature for 1h followed by 10min heat-inactivation at 65 °C. Chemical competent Ecoli cells (Top10) were transformed according to manufacturer´s recommendations (Invitrogen, Karlsruhe, Germany). Clones were picked and incubated in 2.5ml Ampincillin LB medium at 37°C over night in 15ml tubes. For mini preparation, a Miniprep kit (Qiagen, Hilden, Germany) was used according to manufacturer´s instruction. The concentrations were measured with Nanodrop (peqLab Biotechnologies GMBH, Erlangen, Germany). The DNA quality was checked by electrophoresis separation on 1% 1X Borate agarose gel.

3.8.4 Co-transfection of mimic miR 661 and mimic miR 210 with pMIR reporter constructs

A co-transfection assay was carried out to determine whether an ectopic overexpression of the miR- 661 and miR- 210 can affect the luciferase activity of the PMIR reporter construct with a cloned 3´UTR of target genes by binding to the predicted target sites. A transfection mix consisiting of 0.05µl Dharmafect reagent (Thermo Fischer Scientific, Karlsruhe, Germany) and 5µl RPMI 1640 (without FCS) per well were incubated for 5 min at room temperature to allow for liposome complex formation. 5nM final concentration of mimic RNAs were incubated together with transfection mix for 30 min at room temperature and applied to the wells on a 384 well plate. As a negative control for the mimic overexpression, cells were transfected with a mimic all-star, which has no target in the human genome. 3000 human embryonic kidney (HEK T 293) cells in 50µl RPMI 1640 (10% FCS) were added to the 384 well plate incubated for 24h at 37°C. A *Trans*IT LT1 transfection assay (MirusBio. Madison, WI, USA) was carried out after 24h to introduce the pMIR report constructs to the mimic transfected HEK 293T cells. The cells were transfected with 10ng TK Renilla control vector (promega, Madison, WI, USA), 500pg pMIR reporter construct plasmids (Qiagen, Hilden, Germany) and 30ng empty vector in 15µl RPMI 1640. TK Renilla control vector was used as a control for transfection efficiency and viability of the cells. Basic pMIR report was used to normalize the background noise. A positive control for each miRNA mimic was created with 4x perfect matched sequence cloned in PMIR reporter vector. For each assay 6 biological replications of each assay were conducted. Firefly luciferase activity of individual transfections was normalized against *Renilla* luciferase activity and pMiR report-basic activity.

3.9 PCPGL luciferase reporter assay

A luciferase reporter assay was carried out for determination of promoter activity and impact of CpG methylation on expression on a CpG island located 5kb upstream of the pre-miRNA 661. A pCpGL vector with a firefly luciferase reporter gene was used. The luciferase reporter gene is located downstream of the multiple cleavage site of the vector and the activity of the gene is driven by the inserted promoter region. The enzyme luciferase catalyzes a reaction with a luciferin substrate to produce light and the photon emission can be detected by a luminometer (SpectraMax® M5e) to determine the activity of the promoter [153]. The pCpGL vector is CpG free allowing for optimal *in vitro* methylation assays for functional analysis of promoter CpG methylation [154].

3.9.1 Cloning PCR of the miRNA 661 promoter

For molecular cloning of GC rich fragments, PCR amplification was conducted using *Pfu* DNA polymerase. The PCR was carried out with 100 ng genomic DNA in a 20 µl total volume reaction containing 1x GC or High-Fidelity buffer (Thermo Fischer Scientific, Karlsruhe, Germany), 400nM dNTP mixture (Invitrogen), 200mM of forward and reverse primers, 3% DMSO, and 0.2 U *Pfu* HF Thermal cycling was carried out in Master Cycler Gradient with the following conditions for a touch-down program: initial pre-incubation at 98°C for 30 s followed by a second denaturation step at 98°C for 10 s, 10 cycles with starting temperature at 65°C decreasing 0.5 °C per cycle, 25 cycles at 60°C followed by elongation at 72 °C for 1 min and an additional elongation step at 72°C for 7 min. The PCR was performed with primers listed in Table S2**.** The PCR products were run with 6x loadingbuffer and separated by electrophoresis on 1.0 % 1X borate agarose gel. A 1kb and 100bp size indicator was run on the same gel to determine the size of the PCR product. The correct fragments were cut out from the gel and purified with Qiaquick PCR purification kit according to manufacturer´s protocol. To verify the correct size of the DNA fragment, 200 ng was separated on a 1.2% 1X borate agarose gel and photographed.

3.9.2 Restriction enzyme digest

The PCR product (500ng) and the PCPGL vector (1µg) was digested with *BAMH1* and *HindIII* in a final volume of 30µl containing 1x NEB buffer, 1% BSA, 0.5U *SAP*, and 0.2U *BAMHI* and *HindIII* ,respectively for 3h at 37°C. The digested product was run on a 1% 1X Borat agarose gel and the correct product was cut out and purified with Qiaquick PCR purification kit according to the manufacturer´s protocol.

3.9.3 Ligation PCPGL plasmid

BAMHI/HindIII digested DNA was ligated with a pCpGL vector in a 10 µl total reaction volume containing 1µl pCpGL vector, 3µl insert DNA, 1x T4 buffer and 0.5 U T4 ligase over night at 16°C. To remove salt and bi products, the plasmid was precipitated with 1µl Glycogen, 6µl NH4Ac 7.5 M, 2.5 vol EtOH 100% and 100 µl EtOH 70%. The pellet was air dried and diluted in 5µl ddH2O. 35 µl electro-competent E. coli cells (PIR strain) were transformed with 1µl purified pCpGL plasmids using E pulser (Bio RAD Labs, Munich, Germany). Transformed E. coli were incubated with LB medium at 37°C for 45 min to activate the Zeocin resistance. 50 µl was applied on Zeocin positive agar plates and incubated over night at 37 °C. Positive clones were picked and incubated in 50 ml Zeocin LB medium at 37°C over night in Erlenmayer flasks. For midi preparation, a HISPEED MIDI kit (Qiagen, Hilden, Germany) was used according to manufacturer´s instruction with the following modifications: After isopropanol precipitation, samples were transferred to a 50ml Falcon tubes and centrifuged at 10000 rpm for 30 min. The pellet was washed with 70% EtOH followed by centrifugation for 30 min at 10 000 rpm at 4°C. The pellet was air dried at room temperature and diluted in 100 µl ddH2O (Qiagen, Hilden, Germany). The concentrations were measured with Nanodrop and adjusted to 100ng/microliter with EB buffer. The DNA quality was determined by electrophoresis separation on 1% 1X borate agarose gel.

3.9.4 Transfection PCPGL constructs

3000 HEK 293T cells were seeded in 50 µl RPMI 1640 10% FCS medium in a 384 well plate. The cells were grown for 24 h at 37°C prior transfection. The cells were transfected with a mixture of 30ng luciferase reporter vector and 10ng TK Renilla control vector, using TransIT-LTI reagent. In brief, 0.15µl TransIT reagent and 4.85µl RPMI 1640 medium without FCS per well was mixed and incubated at room temperature for 5 min. 20ngTK renilla control vector diluted in RPMI was added to the transfection mix and a master plate for each PCPGL construct was prepared. For each pCpGL construct, a master mix was conducted for 7 wells. 30ng/ well pCpGL plasmid DNA was added to the TransIT and TK Renilla mix and incubated for 20-30 min at room-temperature. 20 µl Transfection mix (TransIT mix together with pRL TK and pCpGL plasmid DNA) was applied to each well and incubated for 48 hours at 37°C. pRL TK vector was used as a control for transfection efficiency and viability of the cells. For each assay, 6 biological replicates were applied and a CMV-PCPGL vector was used as a positive control for the assay. Basic pCpGL was used to normalize the background noise. Firefly luciferase activity of individual transfections was normalized against Renilla luciferase activity and pCpGL-basic activity.

3.10 Western Blot

Western blot was carried out on whole cell lysates from H1299, H1703, A549 and HEK293T cells. Cells were harvested with PBS and spun down at 700rpm for 5 min. 50-100µl SDS-lysis buffer was added to the cell pellet and samples were incubated on ice for 30 min followed by 10min centrifugation at 13000xg to separate cell debris from supernatant. The supernatant was transferred to a new tube and boiled at 95°C for 10 min. Protein quantification was carried out with the BCA method described in [155]. 15-20µg protein was separated on a 10% SDS PAGE gel at 50-75 volt for 2h in 1x running buffer. The gel was transferred by wet blot in 1x transfer buffer at 200mA for 2h. Proteins were visualized using Horse Radish peroxidase (HRP) conjugated secondary antibodies (Santa Cruz, Biotech, Heidelberg, Germany) using the enhanced chemiluminescence detection system (Amersham Pharmacia Biotech, Little Chalfont, UK). Band intensities were quantified with ImageJ software (Image/ImageJ, NIH, USA).

3.11 TGF β1 stimulation

A549, H1299 and H1703 cells were treated with 2, 5 or 10ng/ml TGF β1 ligand (Invitrogen, Hilden, Germany) between 8h-144h. Medium exchange with TGF β1 was carried out after three days. Cells were harvested with cold PBS and spun down at 600rpm for 10min and cell pellet were put at -80°C until RNA isolation. RNA isolation and expression was carried out according to protocol in chapter 3.6.

3.12 Statistical analysis

For case control comparison, the PennCNV tool was used to identify those stretches of SNPs that had significant copy number changes in cases versus controls using Fisher's Exact Test with the raw CNV files. P-values for each SNP in one defined CNV region was combined using the Stouffer´s test described in [156]. The statistical analysis of the CNVs was performed by Agnes Hotz-Wagenblatt, DKFZ, Heidelberg. Significant differences in miRNA, mRNA expression and DNA methylation between tumor tissue and matched adjacent normal tissue were calculated by Wilcoxon matched pairs signed rank test where p < 0.05 was set to be statistically significant. Two sided T-tests were used to assess statistical significance in mRNA expression between treated and untreated cell lines. Correlation between miRNA expression and target gene expression and between methylation and expression was carried out using Spearman´s correlation test where p< 0.05 was set to be statistically significant. Heat maps were created with the Multiple Experiment Viewer software suite (L. Craig Venter Institute, San Diego, USA, (http://www.tm4.org/mev/)) to visualize % methylation for each CpG site or CpG unit. Kaplan Meier plots were created with GraphPad Prism 5 and the significance was calculated with Log Rank (Mantel-Cox Test). The cutoff value for methylation state high and low was calculated by using the crit-level procedure described in [157] and ADAM statistical software SPSS 21.0 (DKFZ, Heidelberg, Germany). The survival time was calculated from the date of operation to the date of death or the date of last observation. Multivariate analysis was done using the Cox-proportional hazard regression analysis applying a stepwise backward procedure [158]. Thomas Muley, Thoraxklinik Heidelberg, was performing the multivariate analysis.

4. Results

4.1 Germline CNV detection in early onset lung cancer

To identify CNVs associated with lung cancer risk, the Illumina Infinium platform Human Hap550 BeadChip from a GWA study previously performed on 492 early onset lung cancer cases and 487 population-based controls was used (clinical characteristics are listed in Table 5). In collaboration with Agnes Hotz-Wagenblatt (DKFZ Core facility for Genomics and Proteomics, DKFZ, Heidelberg), two CNV detection algorithms, QUANTISNP and PennCNV, were applied to the SNP data (see methods, 3.1). 41 CNVs were detected by QuantiSNP and 38 by PennCNV (a detailed description can be found in Supplemental Table S5 and Table S6). Among these, 25 CNVs were overlapping between the two algorithms. To identify CNVs associated with early onset lung cancer risk, Fischer´s exact test with a cut-off *p* value < 0.05 was performed on the 25 regions that overlapped between these algorithms. Ten of these CNV regions were significantly associated with lung cancer risk for both CNV detection algorithms (Figure 5 and Table 7). Associations with losses were detected on 1q21.1, 8q24.23, 6q12, 11q11 and 19p12 and associations with gains were detected on 8q24.3, 11p15.5, 12p12.3 and 22q11.21. The CNV loss on 1q21.1 is overlapping with a CNV region associated with neuroblastoma [131]. Two CNVs were found on 11q11 that harbor three members of the olfactory receptor family, *OR4P4, OR4S2* and *OR4C11*. The olfactory receptor family is known to be variable between individuals and belongs to the most common CNVs in the genome [71, 112, 118, 119, 159]. CNVs without annotated genes were found on 8q24.23, 6q12, and 19p12. A CNV identified on 12p12.3 harbors pleckstrin homology domain containing, family A (PLEKHA5) was more commonly observed in controls. A CNV on 22q11.21 overlaps with a microdeletion that has been associated with DiGeorge syndrome [160]. This gain region was detected in controls and not in cases. In order to focus the study on the CNVs that may have a putative functional impact on lung cancer, the CNVs on 8q24.3 and 11p15.5 were selected based on their restricted detection in cases and that the annotated genes and miRNAs in these regions have a potential role in lung tumorigenesis.

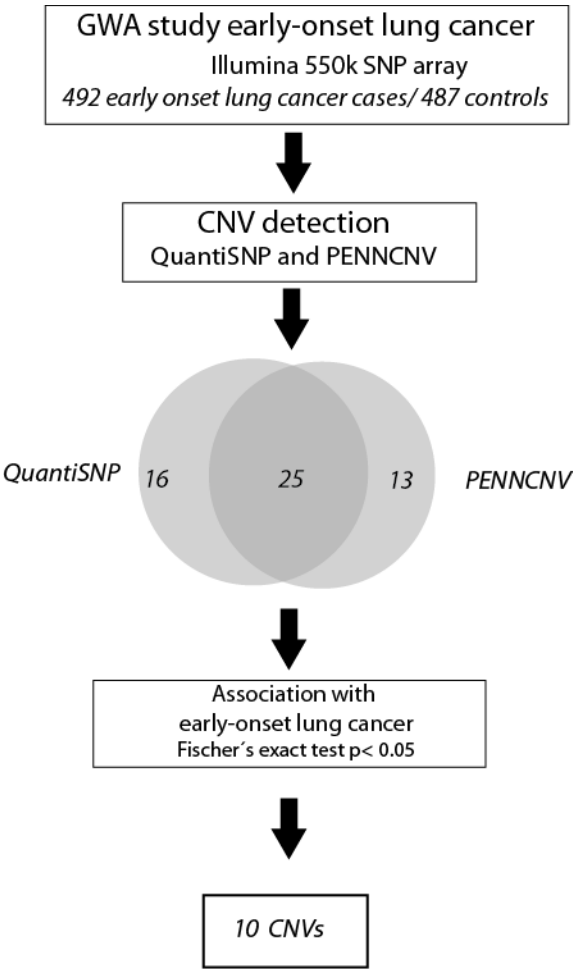


Figure 5. CNV detection workflow in early onset lung cancer GWA study. A GWA study was carried out on 492 early onset lung cancer cases and 487 population based controls using the Illumina Infinium platform Human Hap550 BeadChip. 41 and 38 CNVs were detected by QuantiSNP and PennCNV, respectively. Ten CNVs show a significant association with lung cancer for both algorithms (Fischer´s exact test p<0.05).

Table 7. Descriptive overview of CNVs associated with lung cancer risk detected by QuantiSNP and PennCNV.

Cytoband	CNV coordinates (hg18)[a]	CNV type	PennCNV n cases	PennCNV n controls	P value[b]	QuantiSNP n cases	QuantiSNP n controls	P value[b]
1q21.1	chr1:147305744-147478120	loss	Gain: 5 Loss: 17	Gain:5 Loss: 37	$9.9x10^{-13}$	Gain: 4 Loss: 22	Gain: 9 Loss: 28	$<10^{-18}$
8q24.23	chr8:137898044-137913669	loss	Gain: 0 Loss: 0	Gain: 0 Loss:22	$<10^{-18}$	Gain:0 Loss: 15	Gain: 1 Loss:28	$2.2x10^{-16}$
8q24.3	chr8:145090342-145223898	gain	Gain: 18 Loss: 0	Gain: 0 Loss: 0	$<10^{-18}$	Gain: 27 Loss: 0	Gain: 0 Loss: 0	$<10^{-18}$
6q12	chr6:67093085-67105019	loss	Gain: 0 Loss: 39	Gain: 0 Loss: 67	8.9E-13	Gain: 0 Loss:55	Gain: 0 Loss: 84	3.7E-10
11p15.5	chr11:548884-609789	gain	Gain: 11 Loss: 0	Gain: 0 Loss: 0	$4.4x10^{-16}$	Gain: 11 Loss:1	Gain: 0 Loss: 0	$<10^{-18}$
11q11	chr11:55139733-55179162	loss	Gain: 0 Loss: 57	Gain: 0 Loss: 0	$<10^{-18}$	Gain: 0 Loss: 80	Gain: 0 Loss: 2	$<10^{-18}$
11q11	chr11:55127597-55139733	loss	Gain: 0 Loss: 0	Gain: 0 Loss: 104	$4.1x10^{-10}$	Gain:0 Loss: 0	Gain: 0 Loss: 117	$1.93x10^{-3}$
12p12.3	chr12:19360345-19431361	gain	Gain: 0 Loss: 0	Gain: 11 Loss: 0	$<10^{-18}$	Gain: 4 Loss: 0	Gain: 13 Loss: 0	$<10^{-18}$
19p12	chr19:20422200-20473895	loss	Gain: 0 Loss:22	Gain: 0 Loss:38	$1.7x10^{-8}$	Gain: 0 Loss:25	Gain: 0 Loss:33	$4.08x10^{-2}$
22q11.21	chr22:17257787-17355587	gain	Gain: 0 Loss: 0	Gain: 12 Loss: 0	$<10^{-18}$	Gain: 2 Loss: 3	Gain: 12 Loss: 0	$<10^{-18}$

[a] genomic positions are defined according to start and end SNP detected by PennCNV. QuantiSNP genomic positions are presented in Table S8.

[b] Fisher´s exact test was applied for each SNP and a combined *p* value for the CNV was calculated with Stouffers test [156].

4.2 CNV validation

4.2.1 Sequenom Typer assay application for quantitative, allele-specific copy number analysis

In order to establish a protocol for validation of the CNV candidates, an allele specific copy number analysis (ACN) application using the Sequenom massARRAY platform [149] was chosen .This method has been described to be able to distinguish copy numbers on specific alleles based on SNP information, and theoretically as being able to determine the absolute copy number in a sample. For the establishment of a protocol we used the *GSTM1* locus which is known to be variable and associated with early onset lung cancer [20]. A former PhD student in the group, Maria Timofeeva, had previously identified the *GSTM1* copy number in blood samples from lung cancer patients with a multiplex real time PCR method [144]. A subset of these samples was used for the method establishment. For normalization to a 2N control, recommendations from Sequenom were given for genomic regions without variations (Caren Vollmert, Sequenom, Hamburg, personal communication). Nine assays were designed for these regions and one control was identified as being 2N for *GSTM1* (SNP rs6715929 on chr2:172640000-172700000, hg 18). This 2N control was included in all assays.

4.2.1.1 Determination of competitor DNA concentration

Competitor PCR is a method that was established for quantification of mRNA by spiking a known molar quantity of a synthetic DNA template identical to the gene of interest that competes with the mRNA PCR amplification[150]. To obtain the concentration of competitor DNA which represents the same amount of DNA molecules in the reaction as the gDNA, a serial dilution with competitor together with a fixed amount of DNA template (copies of molecules) of gDNA was analyzed. The amount of haploid copies in 20ng gDNA is 6210 Equation 1, see section 3.2.1). The theoretical concentration of competitor DNA needed to equal 6210 copies was 2.06E-15 M (Equation 2, see section 3.2.1). The actual competitor concentration needed for a specific assay was determined by competitor titration against a fixed amount of gDNA to reach an equilibrium value (EC50) (Figure 6). To cover a wide range of copies, the following concentrations were used for each competitor:

$1.0*10^{-13}$ M, $6.0*10^{-14}$ M, $3.0*10^{-14}$ M, $1.0*10^{-14}$ M, $6.0*10^{-15}$ M, $1.0*10^{-15}$ M, and $1*10^{-17}$ M.

EC50 values were calculated to determine at which concentration the competitor equals the gDNA. An average EC50 concentration over the sample set for each assay was subsequently used for in the CNV analysis.

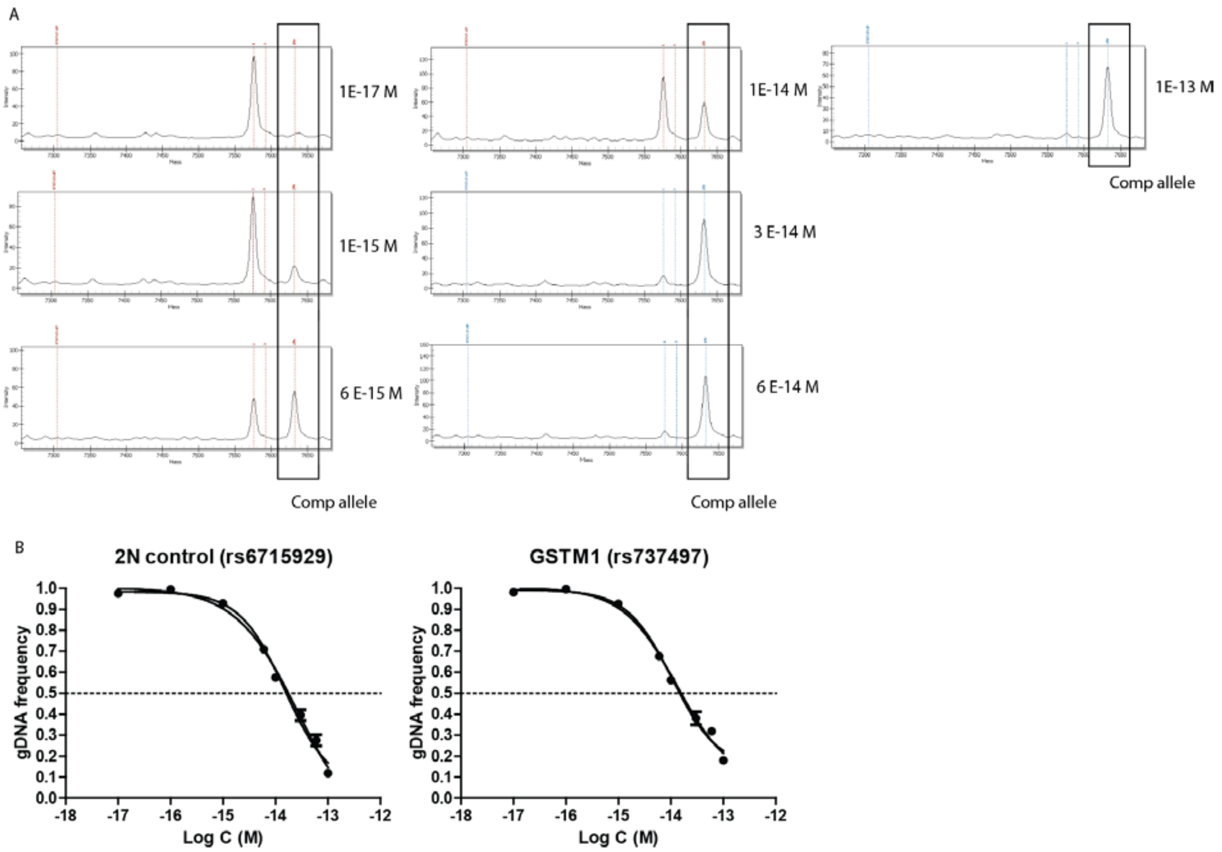


Figure 6. EC 50 determination for 2N control and *GSTM1*. A. Intensity signals conducted by MALDI TOF for the 2N control rs6715929 in a sample for this SNP. A third allele, the competitor allele (Comp allele) is highlighted in a square. Dose dependent increase in signal intensity of Comp allele and decrease of WT allele is illustrated in each graph. **B**. The gDNA frequency is plotted against the log2 concentration of the competitor (Log C M). EC50 values were calculated to determine at which concentration the competitor equals the gDNA. Each dot represents the average of technical triplicates and the errorbars show standard deviations.

4.2.1.2 Copy number analysis of GSTM1

CNV analysis was carried out in 13 patients with known *GSTM1* copy number to determine whether the competitor PCR based method is valid for further analysis of the CNV candidates. The competitor concentration was set to 5.26E-15M based on the average EC50 for rs6715929 for both assays. We could confirm the copy numbers for 12/13 patients for *GSTM1* in a two plex consisting of the 2N control plus *GSTM1* (Figure 7A). Additionally, the *GSTM1* copy number in the same samples was determined using the EC50 concentration for *GSTM1* and the EC50 concentration for the 2N control in order to compare the results (Figure 7B). The homozygous deletion could be confirmed. However, the other copy numbers were shifted upwards, suggesting that the EC50 concentration established for the 2N control is more suitable than using individual concentrations.

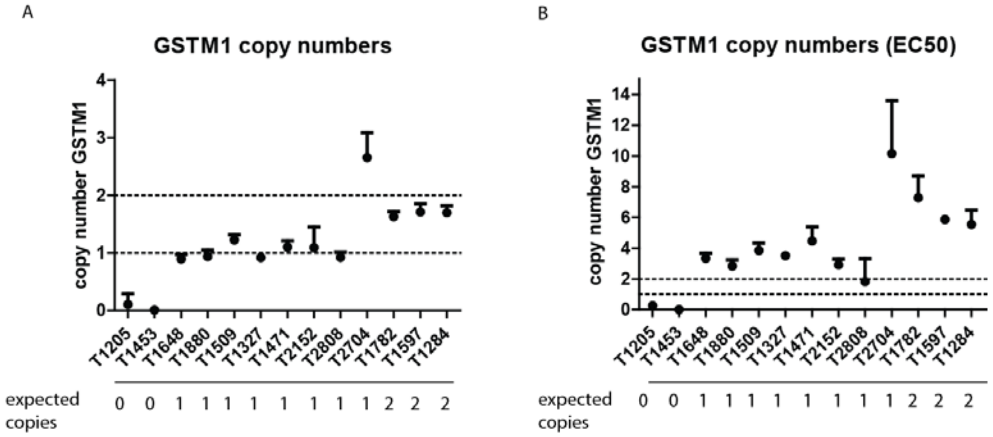


Figure 7. Copy number analysis for *GSTM1* using two different concentrations for the competitor. **A**. The same competitor concentration for 2N control and *GSTM1* was used based on the EC50 for the 2N control. **B**. Copy number analysis where the individual concentrations for the competitor were used based on EC50 determination.

4.2.1.3 Copy number analysis of 8q24.3 and 11p15.5

In order to validate two CNVs detected by PennCNV and QuantiSNP, we used the protocol established for the TyperAssay to determine copy numbers of *GSTM1*. We included *GSTM1* in a three plex for the 8q24.3 region (SNP of interest, 2N control and *GSTM1*) as a control for those samples where the copy number is known (Figure S1). The assays for copy number determination of the CNV on 8q24.3 were designed for the SNPs rs4977177 and rs13255347 located in the minimal overlapped region (Figure 8A). The analysis was carried out on 11 patients where no copy number variation was detected by either PennCNV or QuantiSNP and 12 patients with detected copy number gain. The selection of samples for validation was based on DNA availability. The start and end of the CNV detected by QuantiSNP and PennCNV in each patient are shown in (Figure 8A). The copy number analysis for rs4977177 was validated for 7/11 with 2N but did not confirm the expected copy gain in the samples analyzed (Figure 8B). The *GSTM1* copy number were confirmed for 19/23 samples for rs4977177 assay (Figure S1A). Three patients with expected copy number gain at 8q24.3, sample T530, T699 and T2187, were confirmed in the analysis of rs13255347 (Figure 8C). The *GSTM1* copy number was confirmed for T614 and T2187 but not for T530 in the same plex (Figure S1B). For copy number analysis at 11p15.5, two assays were designed in the overlapped region for rs1062099 and rs746707 (Figure 9A). The copy number variation for samples with expected copy number gain or loss was not confirmed for either assay (Figure 9B-C).

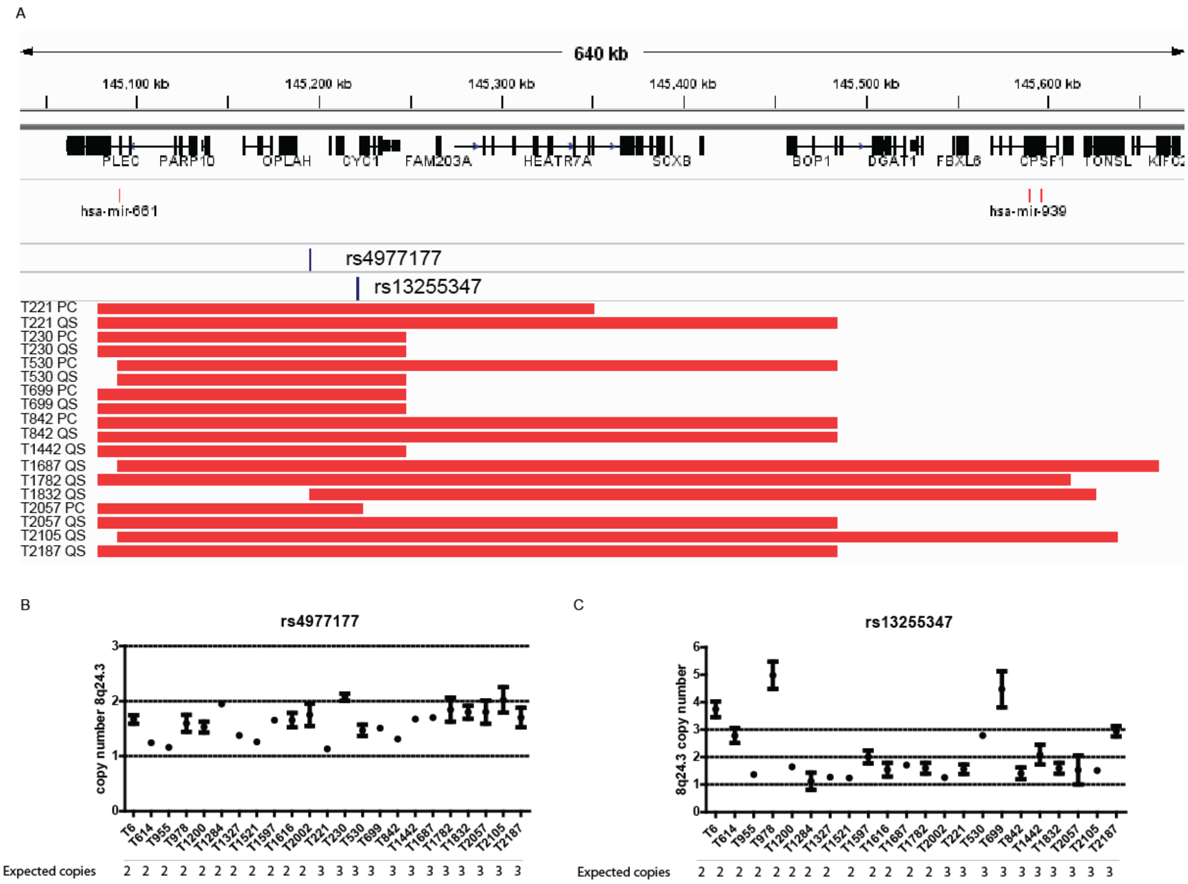


Figure 8. 8q24.3 copy number analysis with TyperAssay. A. Locus overview with positions for rs4977177 and rs13255347 and genomic positions for CNVs detected by PennCNV and QuantiSNP in patients (red). **B-C.** Normalized copy number in blood from patients from TyperAssay for rs4977177 (B) and rs13255347 (C). The graph shows the average copy number from three technical repeats and standard deviations in bars. The expected copies are based on PennCNV and QuantiSNP algorithms.

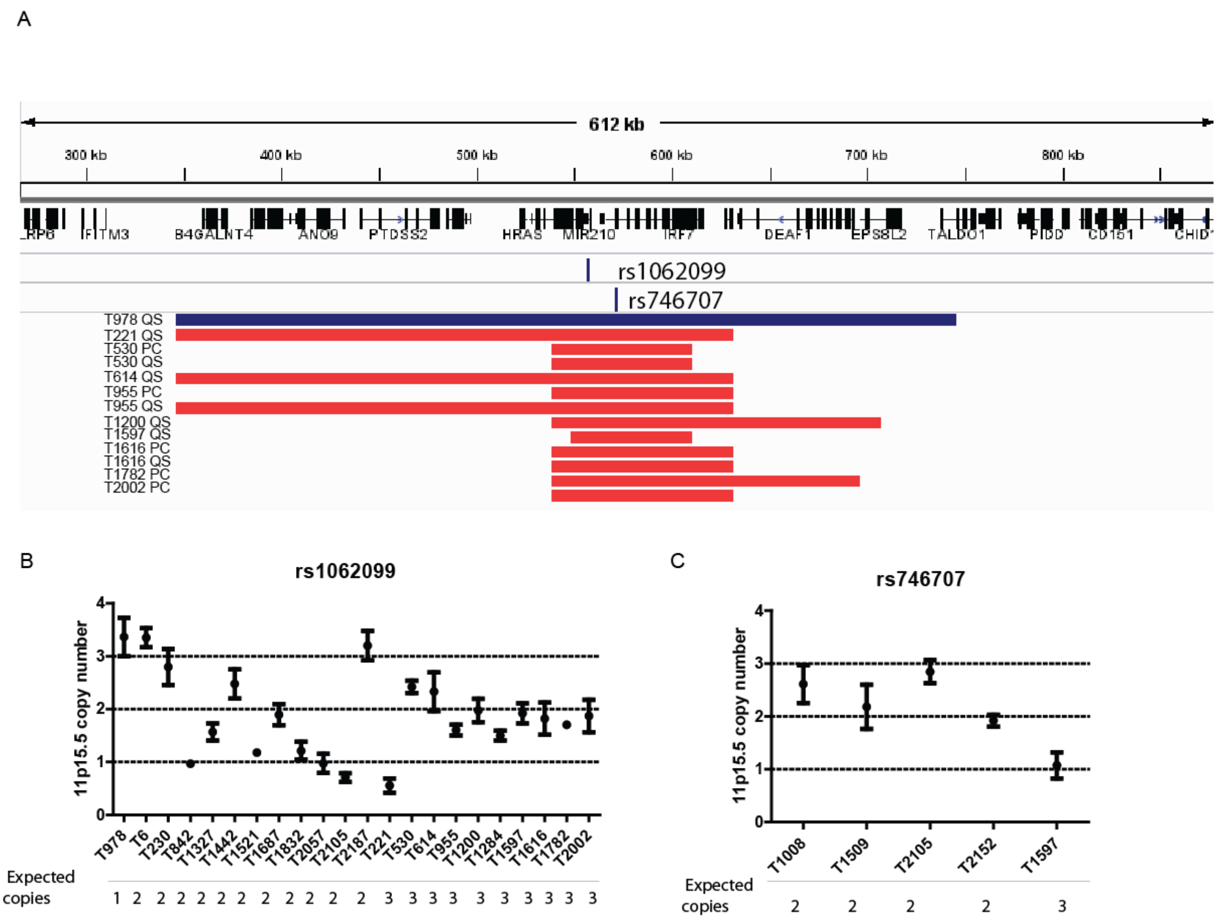


Figure 9. 11p15.5 copy number analysis with TyperAssay A. Locus overview with positions for rs1062099 and rs746707 and genomic positions for CNV gains detected by PennCNV and QuantiSNP in patients with gain (red) and loss (blue). **B-C**. Normalized copy number in blood from patients from TyperAssay for rs1062099 (B) and rs746707 (C). The graph shows the average copy number from three technical repeats and standard deviations in bars. The expected copies are based on PennCNV and QuantiSNP algorithms.

4.2.2 CNV analysis using FISH

Fluorescence *in situ* hybridization (FISH) is commonly used in cytogenetics and diagnostics to detect chromosomal aberrations associated with disease. In collaboration with Anna Jauch (Human Genetics, Heidelberg University), a FISH assay was carried out on lymphocytes from patients where the expected copies of 8q24.3 and 11p15.5 was determined by PennCNV and QuantiSNP. Two additional control samples were used known not to have any alterations on 8q24.3 and 11p15.5 (supplied by Brigitte Scholl, Human genetics, Heidelberg University, unpublished data).

4.2.2.1 BAC test

To determine the specificity of the bacterial artificial chromosomes (BACs) used for FISH, the hybridization was carried out on cells in metaphase to determine the chromosomal location together with a probe for 8q and 11p. The genomic location of the BACs used for copy number analysis is illustrated in Figure 10. The genomic localization was confirmed for two BACs for each region, RP11-714N16 and CTD-3202E22 for 8q24.3 and for 11p15.5, CTD-2647G13 and RP11-1007G14 (example in Figure 11).

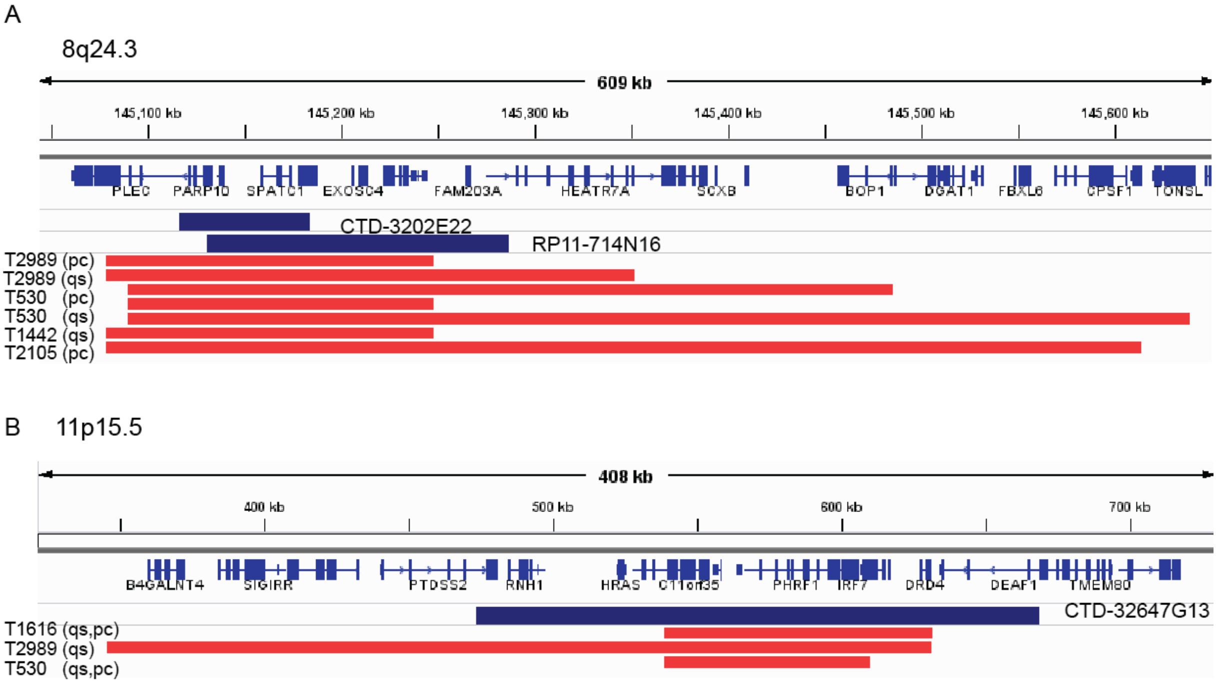


Figure 10. Genomic location for BAC probes used in FISH. A. Two BAC probes, CTD-3202E22 and RP11-714N16 were used for 8q24.3 CNV determination and was carried out in four patient samples. The red lines show the start and end genomic location for detected CNV by either PennCNV (pc) or QuantiSNP (qs). **B**. BAC probe CTD-32647G13 was used to determine the CNV on 11p15.5 in three samples.

4.2.2.2 Copy number analysis in patient lymphocytes using FISH

FISH analysis was carried out on lymphocytes in interphase from patients with detected copy number gains. In order to quantify whether the region showed a gain, 50- 100 cells were analysed per patient to determine the frequency of the event. Separate fluorescent signals for the BAC probes were not expected since the BAC size and size of CNV ratio were minor (Figure 10). Instead we hypothesized that if we observe size or fluorescent increase between the two alleles, this may indicate gain on one allele. To answer this question, quantification was carried out for the volume, particle size and integrated density of the signals using stacked pictures of each FISH result with Image J. Significant changes were observed in volume, size and integrated intensity in the patients (Example shown for patient T530 in Figure 12) indicating differences between the two alleles. The same parameters were analyzed in the control samples and we compared whether the differences (ratios) between allele 1 and allele 2 could be distinguished between cases and controls. The parameters for quantification analyzed showed no significant difference between case and controls, thus, the copy number gain in these patients could not be confirmed by FISH (Figure 13).

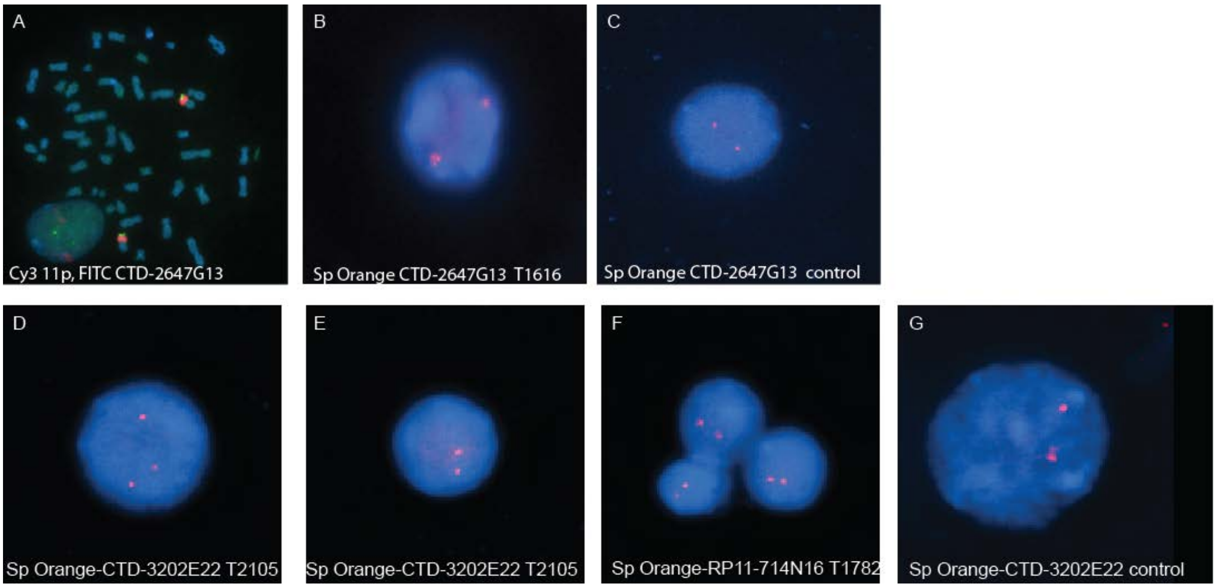

A
Cy3 11p, FITC CTD-2647G13
B
Sp Orange CTD-2647G13 T1616
C
Sp Orange CTD-2647G13 control
D
Sp Orange-CTD-3202E22 T2105
E
Sp Orange-CTD-3202E22 T2105
F
Sp Orange-RP11-714N16 T1782
G
Sp Orange-CTD-3202E22 control

Figure 11. Copy number analysis for 11p15.5 and 8q24.3 with FISH. A. Cy3 coupled 11p probe (Red signal) was co-hybridized with FITC labeled probe CTD-2647G13 (green signal) to determine the right chromosomal location for 11p15.5**. B-G**. Lymphocytes in interphase were hybridized with Sp Orange labeled probes (red signal) for 11p15.5 (B-C) and 8q24.3 (D-G).

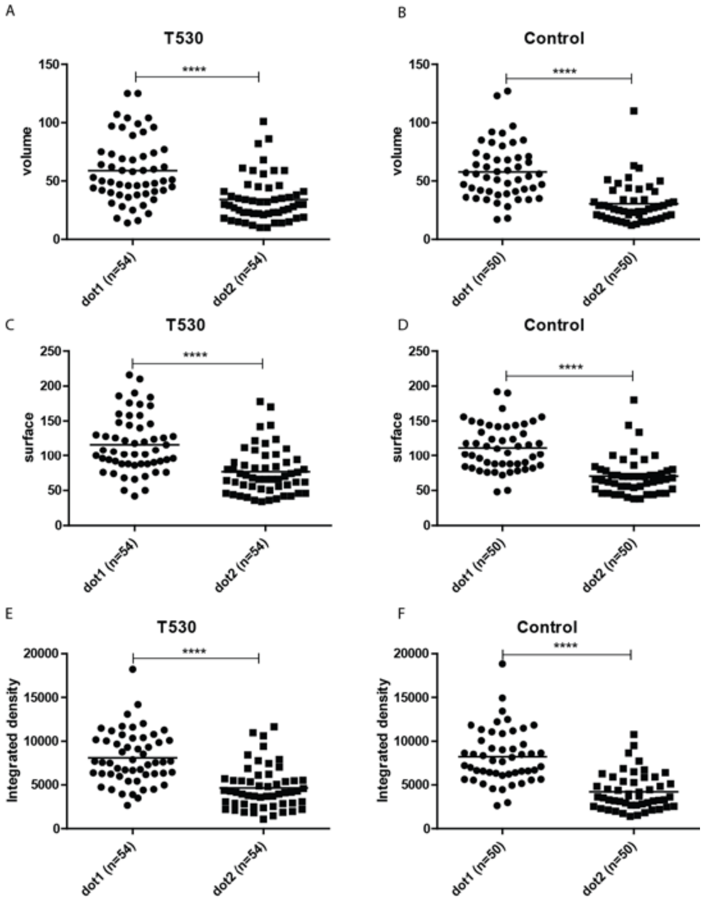


Figure 12. Quantification of FISH results with Image J software. Three parameters were analyzed between the two allele signals (dot1 and dot2) for patient T530 and control; volume **(A-B),** surface **(C-D),** and integrated density **(E-F).** The two sided unpaired T test was used to determine significance (**** *p*< 0.0001).

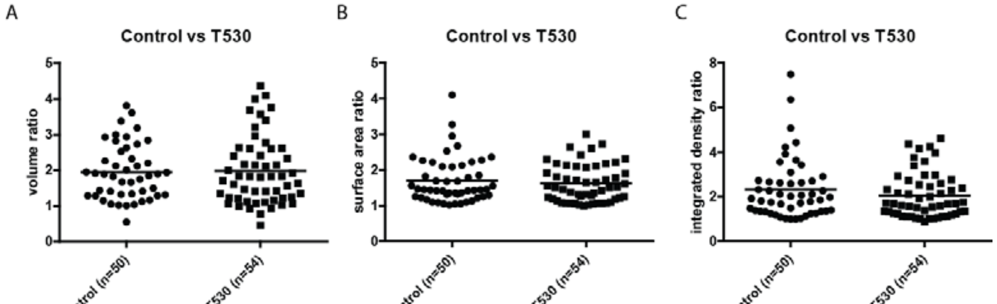


Figure 13. Comparison of volume, surface and integrated density determined by Image J between patient and control. A-C. The volume (A), surface (B) and integrated density ratio (C) between two alleles was compared between control sample and patient sample. No significant difference was observed between control sample and patient sample (two sided unpaired T test p> 0.05).

4.3 Chromosome 8q24.3

4.3.1 Expression and methylation on chromosome 8q24.3

In order to elucidate the mechanism by which the detected CNV gain on 8q24.3 affects lung tumorigenesis, an expression analysis was performed using RT qPCR in 36 tumor-normal pairs for all genes and the miRNA located in the minimal overlapped region (Figure 14). Among the genes and miRNAs expressed in lung tissue, the most significant difference was seen for miR- 661 which was shown to be consistently overexpressed in tumor tissue (p< 0.0001, Wilcoxon matched pairs signed rank test) (Figure 14B). Upregulation was observed in 32/36 patients. *PLEC1*, *GRINA*, and *MAF1* genes were significantly downregulated (downregulated vs upregulated: N=36, 26/1, *p*< 0.0001, 27/9 *p*<0.0001, and 9/3 *p*= 0.03) and *EXOSC4* was upregulated in 12/36 tumors and downregulated in 2/36 tumors (p< 0.0001). The log2 foldexpression between tumor and normal is shown in (Figure 14I). *PARP10, SPATC1* and *OPLAH* were detected at a high CT value in the normal and lung cancer (>33) and excluded from further analysis (data not shown). *GPAA1* and *SHARPIN* and showed no significant difference between tumor and normal (Figure 14F-G).

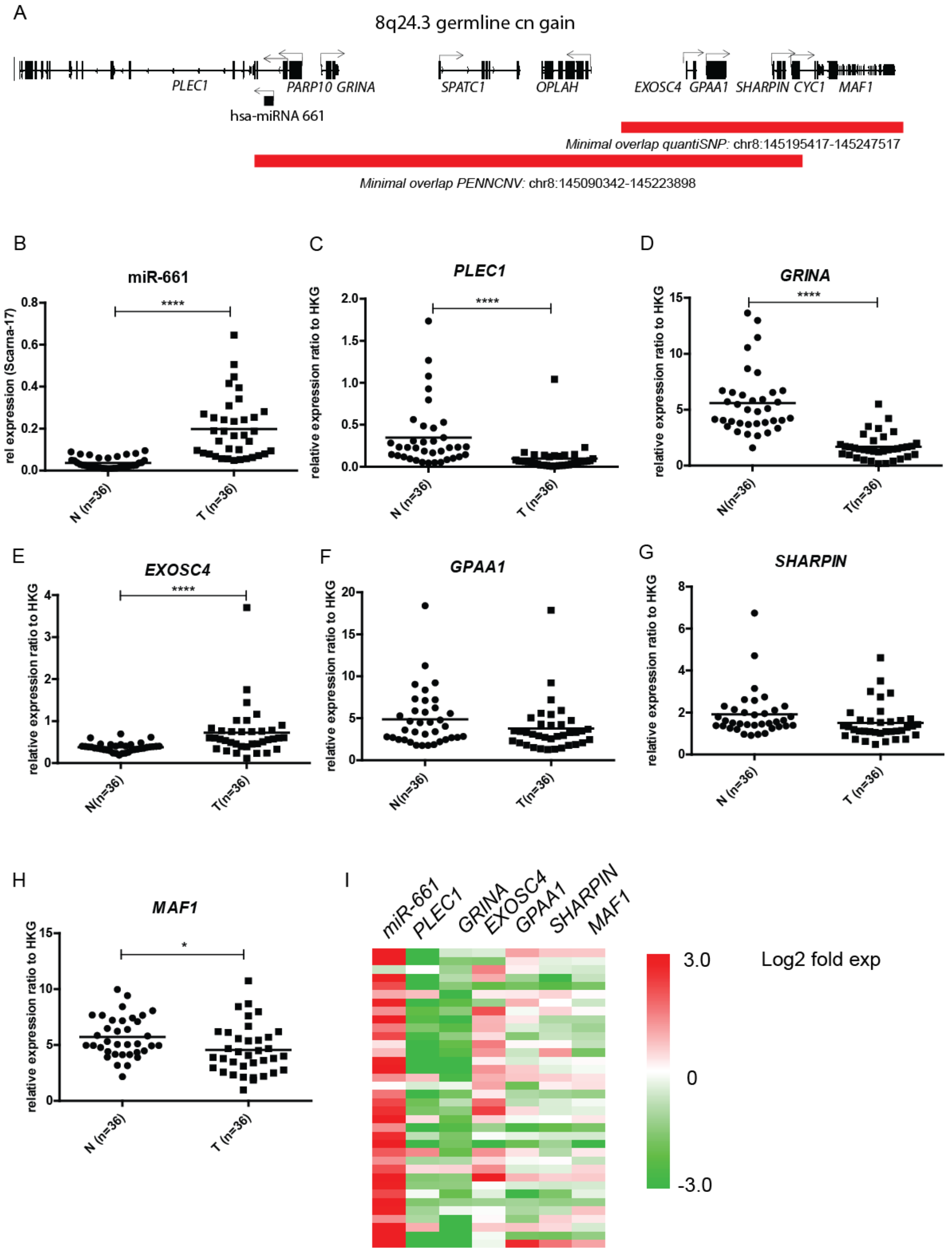


Figure 14. RNA expression analysis of 8q24.3 region in lung cancer. A. 8q24.3 genomic representation with the genomic location of the minimal overlapped region detected with PennCNV and QuantiSNP. **B-H.** Expression analysis of miR-661, *PLEC1*, *GRINA*, *EXOSC4*, *GPAA1*, *SHARPIN* and *MAF1* was carried out in 36 tumor and adjacent lung tissue. The expression is shown relative to *HPRT*. I. Heatmap over Log2 foldexpression in tumor tissue of 8q24.3 genes. The scale ranges from 3 log2 in red to -3.0 log2 in green. The foldexpression is the average of *HPRT* and *GAPDH* normalization. Wilcoxon matched pairs signed rank test was used to determine significant differences between matched normal and tumor tissue (* *p*= 0.05, ** *p*=0.01 *** *p*= 0.0001, **** *p*< 0.0001).

4.3.2 miRNA 661 is upregulated and hypomethylated in lung cancer

Deregulation of miRNA expression as a consequence of altered promoter DNA methylation has recently been shown in CLL and other cancers [79, 80]. Thus DNA methylation was analyzed as a possible mechanism of miR-661 overexpression in lung cancer. Amplicons were designed to cover the closest upstream CpG island (CpG island 61) located 5kb upstream of the precursor-miRNA and the amplicons were overlapped with publically available data for the active transcription mark H3K4me3 (ENCODE project [161]) (Figure 15A). A heterogeneous methylation pattern was observed over the region (Figure 15B). The highest significant methylation differences between tumor and matched normal tissue were seen at amplicon A3. Hypomethylation in the tumor tissue with at least 20% methylation differences was seen for 23/36 matched tumor and normal pairs (Figure 15C). DNA methylation at amplicon A3 and miR-661 expression in the tumor tissue showed a significant inverse correlation (Spearman r =- 0.35 *p*= 0.04) indicating that loss of methylation at these CpG sites is associated with increased levels of miR-661 expression. The methylation data was validated in an additional sample set with 88 lung tumor and normal pairs and hypomethylation in 48% of the tumors was observed (Figure S2). Next, we determined whether the hypomethylated region harbors an active promoter site. The region was cloned in a 1.2kb DNA fragment covering the sequence of amplicon A3 into a luciferase reporter vector. A 9 fold increase in luciferase expression over the basic vector was seen (Figure 15E). A truncated version of the fragment excluding the hypomethylated sequence significantly reduced the luciferase activity, indicating that this region is important for promoter activity.

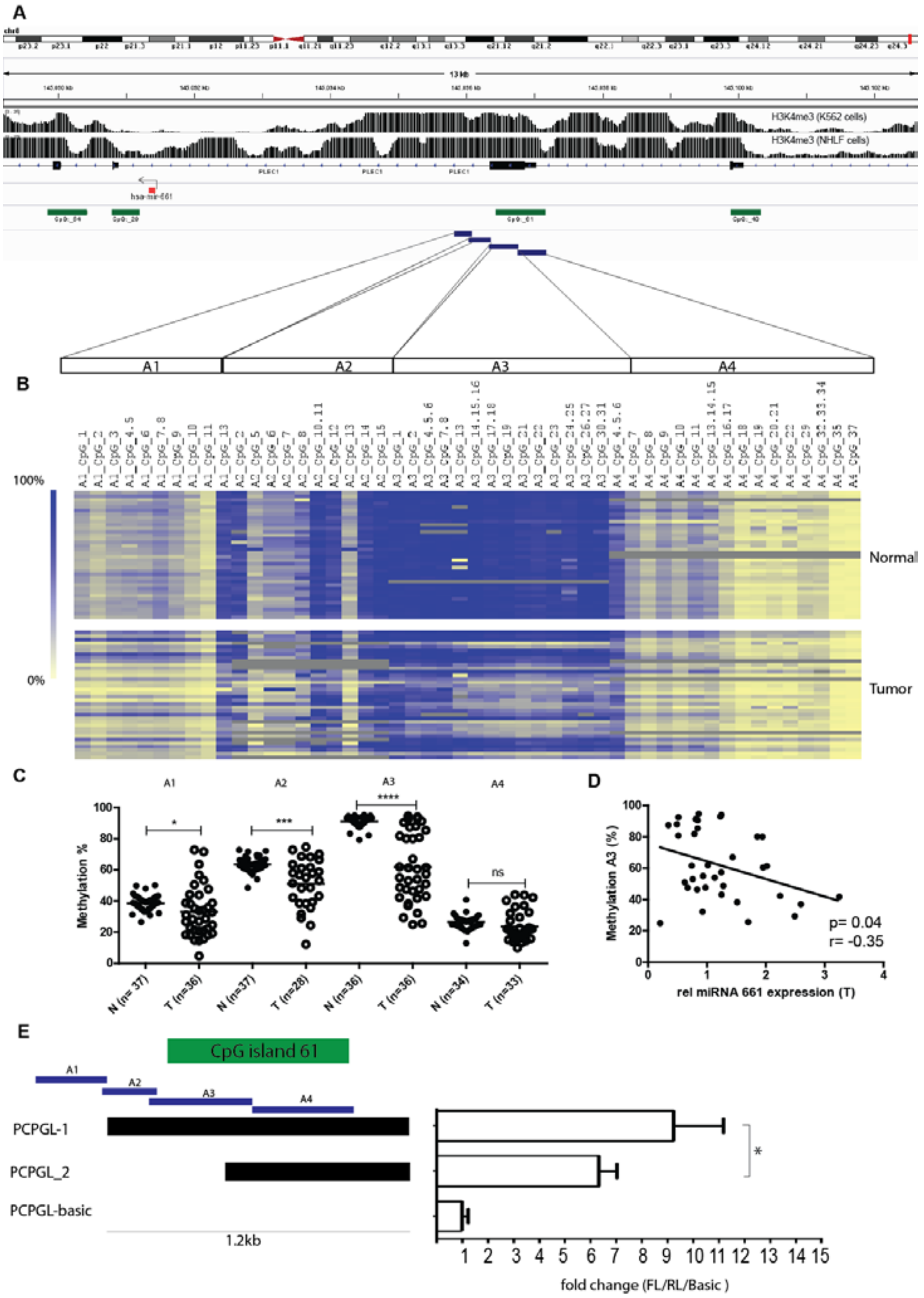
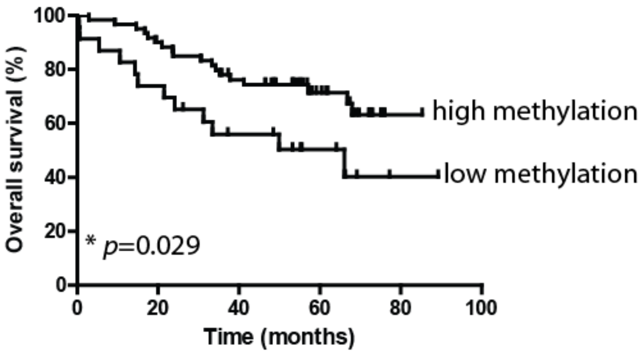


Figure 15. Demethylation at a putative promoter for miR-661 correlates with expression in tumor tissue. A. Overview of miR-661 locus with ENCODE data for H3K4me3 from K562 cells and NHLF cells (Extracted from Integrated Genome Viewer, Broad Institute (IGV [162])). Amplicons A1-A4 cover the CpG island 61 5kb upstream from pre-miR-661**. B**. Heatmap shows methylation in % on a scale from yellow (0% methylation) to blue (100% methylation). Methylation for amplicon A1-A4 is shown for 36 normal lung and tumor pairs. Lines in grey depict missing data points. **C**. Average methylation of the amplicons A1 to A4 in normal (N) and tumor (T) tissue. Wilcoxon matched pairs signed rank test determines the significance **D**. Average % methylation at amplicon A3 in tumor (y axis) plotted against relative miR-661 expression in tumor (x axis). Correlation between methylation and expression was determined using Spearman correlation test (**p* < 0.05,). **E**. Illustration of promoter luciferase construct PCPGL-1 (1.2kb) covering CpG island 61 and PCPGL-2 (750bp) without the hypomethylated region. Significance was determined by an unpaired two sided T- test (**p* <0.05).

4.3.2.1 Methylation and clinical outcome in NSCLC

Next, we investigated whether mir-661 promoter methylation correlates with clinical outcome in NSCLC. The analysis was carried out in set 1 and 2 (n=83). We defined two methylation states, high and low, based on the average methylation in amplicon A3 defining the high (>47%) and low (< 47%) methylation groups. The 5-year survival rate was 71.4% for high methylation state and 50.3% for low methylation state. In the multivariate analysis the methylation rate was a significant prognostic factor independent of stage, histology and gender with a hazard ratio (HR) of 2.27 (95% CI: 1.09-4.72) (p=0.029) (Figure 16).



Number at risk						
High meth	60	54	43	22	2	0
Low meth	23	18	12	7	2	0

Figure 16. Low methylation is associated with worse overall survival. Kaplan Meier plot with high and low methylation state. The *p* value was obtained with Log rank (Mantel Cox) test.

4.3.2.2 Top hit genes targeted by miR-661

miR-661 has over 2000 predicted targets. To determine which genes or pathways are regulated by miR-661, an overlap between five different prediction algorithms was made and targets present in at least two of the chosen prediction algorithms were further evaluated for expression in lung tissue. The selection of targets from each algorithm was carried out using the highest scores per prediction with cut-offs based on recommendations for each algorithm. For TargetScan 6.2, selection was based on the total context score with the cut-off at -0.50 [163]. For Diana microT-v5 cut-off recommendations were set to 0.8 [164]. For miRanda the “good” mirSVR score was set to <= -0.1 [165]. For miRDB a score > 80 was recommended [166]. miRWALK combines scores from selected data bases and calculates a p-value where significance was set to p<0.05 [167]. mRNA expression was analyzed in 36 lung tumor -normal tissue pairs of four targets, *MAP3K3*, *RIPK2*, *DIRAS3* and *GAS7* (Table 8). All four genes were significantly downregulated in lung tumor (Figure 17A,C,E,G). The *MAP3K3* mRNA expression and the *DIRAS3* mRNA expression in tumors were negatively correlated with miR-661 expression in the same tumors

(Spearman r = -0.38, *p*= 0.02 and r-0.45, *p*=0.005, respectively), indicating that the expression may be regulated by the miR-661. *RIPK2* and *GAS7* expression in tumor tissue showed an inverse correlation trend (Spearman r= -0.13, *p*= 0.42 and r= -0.19, *p*=0.27, respectively) (Figure 17B, D, F, H).

Table 8. miR-661 top hit genes for five different target prediction tools

Gene name	TargetScan 6.2 [163]	Diana -v5[164]	miRanda[165]	miRDB [166]	miRwalk [167]
MAP3K3	-0.95	0.97			*p* <0.05
RIPK2			-1.0445		*p* <0.05
DIRAS3			-1.04		*p* < 0.05
GAS7		0.82		81	*p* <0.05

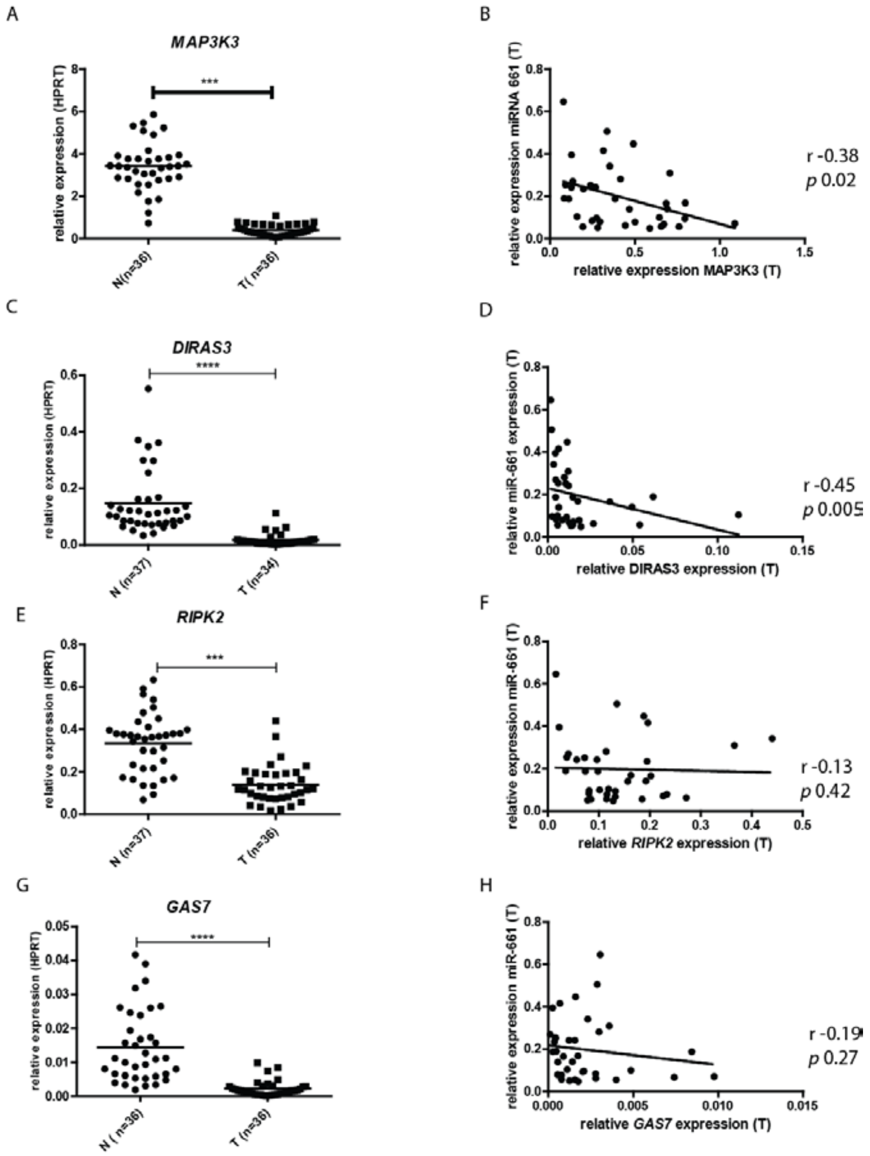


Figure 17. miR-661 top hit targets are downregulated in lung cancer. **A-B**. *MAP3K3* was significantly downregulated in tumor tissue and negatively correlated with miR-661 expression Spearman r =-0.38, *p*=0.02). **C-D**. *DIRAS3* was significantly downregulated in lung cancer and showed a reverse correlation with miR-661 expression in tumor (Spearman r =-0.45, *p*=0.005). **E-G**. *RIPK2* and *GAS7* were downregulated in tumors but showed no significant inverse correlation with miR-661 expression.

4.3.2.3 Top hit target expression after ectopic overexpression of miR-661 in lung cancer cell lines

To further explore the relation between the predicted top hit genes listed in Table 8 and miR 661 H1703, H1299 and A549 lung cancer cell lines were transfected with mimic miR-661 during 24, 48 or 72h and a mRNA expression analysis of the targets was carried out (Figure 18). *GAS7* is not expressed in H1703, H1299 and A549 and therefore not further analyzed (Data not shown). *MAP3K3, DIRAS3* and *RIPK2* showed a significant reduction after ectopic overexpression of miR-661 (Figure 18A, B, C). *MAP3K3* was additionally analyzed on protein level (MEKK3) and the same effect was observed (Figure 18D and E). When inhibiting the endogenous miR-661, MEKK3 protein was induced in H1703 and H1299 cells after 24h, and 72h transfection (Figure 18D and E). This effect was not seen on mRNA level (data not shown).

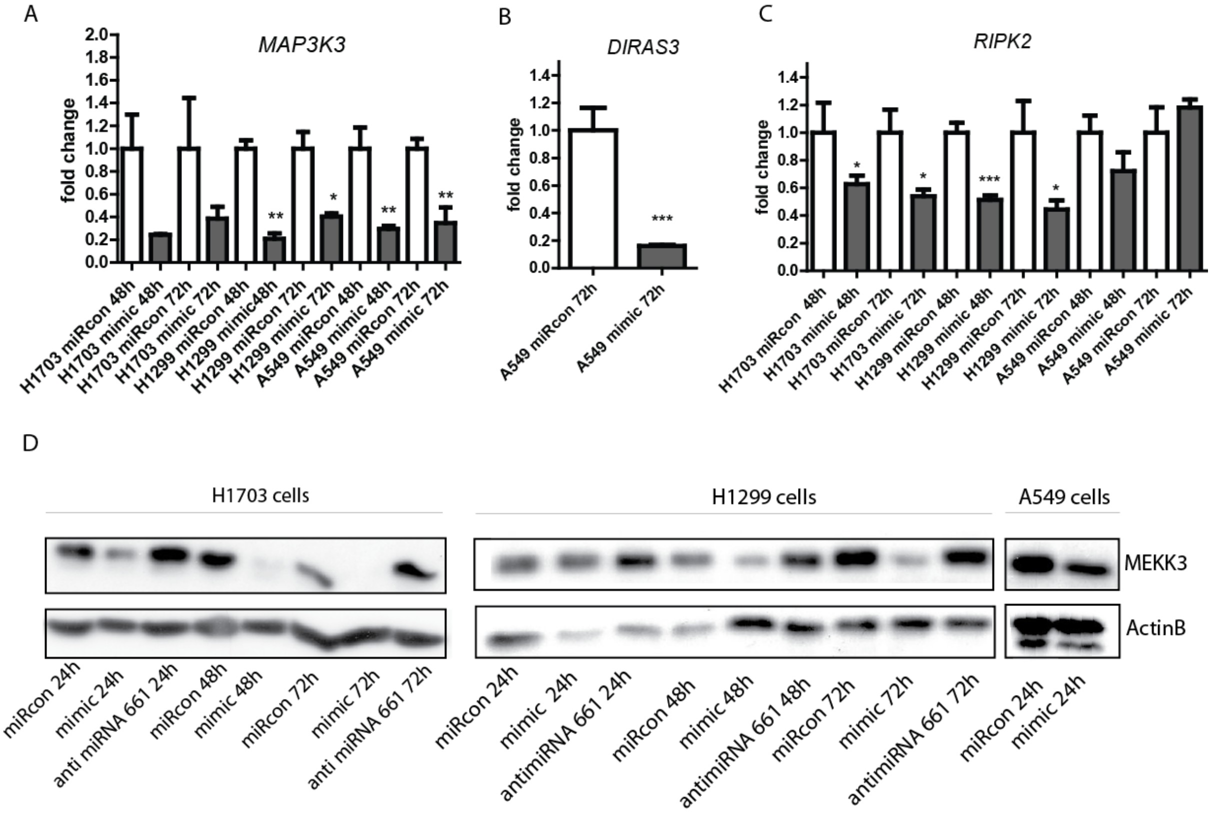


Figure 18. Top hit target expression after miR-661 ectopic overexpression or inhibition. A. *MAP3K3* mRNA expression in H1703, H1299 and A549 cells after mimic and miRcon treatment. The expression analysis was carried out after 48 and 72 h treatment. **B**. *DIRAS3* mRNA expression in A549 cells after 72h mimic or miRcon treatment. **C**. *RIPK2* mRNA expression in H1703, H1299 and A549 cells after 48 and 72h mimic and miRcon treatment. *GAPDH* was used for normalization for mRNA expression. The fold change expression shows the average of three technical replications with standard deviations. The experiment was repeated three times. **D**. Protein expression of MEKK3 in H1703, H1299 and A549 cells after miRcon, mimic and anti-miR treatment. ACTINB was used as a loading control for protein expression.

4.3.2.4 EMT related gene targeted by miR-661

miR-661 has previously been shown to play a role in EMT in breast cancer cells [168]. A second approach was therefore to further refine searches for target genes functioning in cell adhesion, cell to cell contact or known markers in EMT that may contribute to lung cancer. The MirWAlk data base was used to determine predicted targets of miR-661 with cut-off *p* value=0.05 for four target prediction algorithms (Diana-T, miRanda, PITA, and Targetscan) and with this approach transmembrane glycoprotein E-cadherin (*CDH1*) was identified. *CDH1* mRNA expression in lung tumor showed a minor downregulation and upregulation with no significant inverse correlation with miR-661 expression (Figure 19A and B). Ectopic overexpression with miR-661 mimic in a lowly invasive cell line A549 reduced the mRNA expression significantly after mimic treatment as well as the protein (Figure 19C and D).

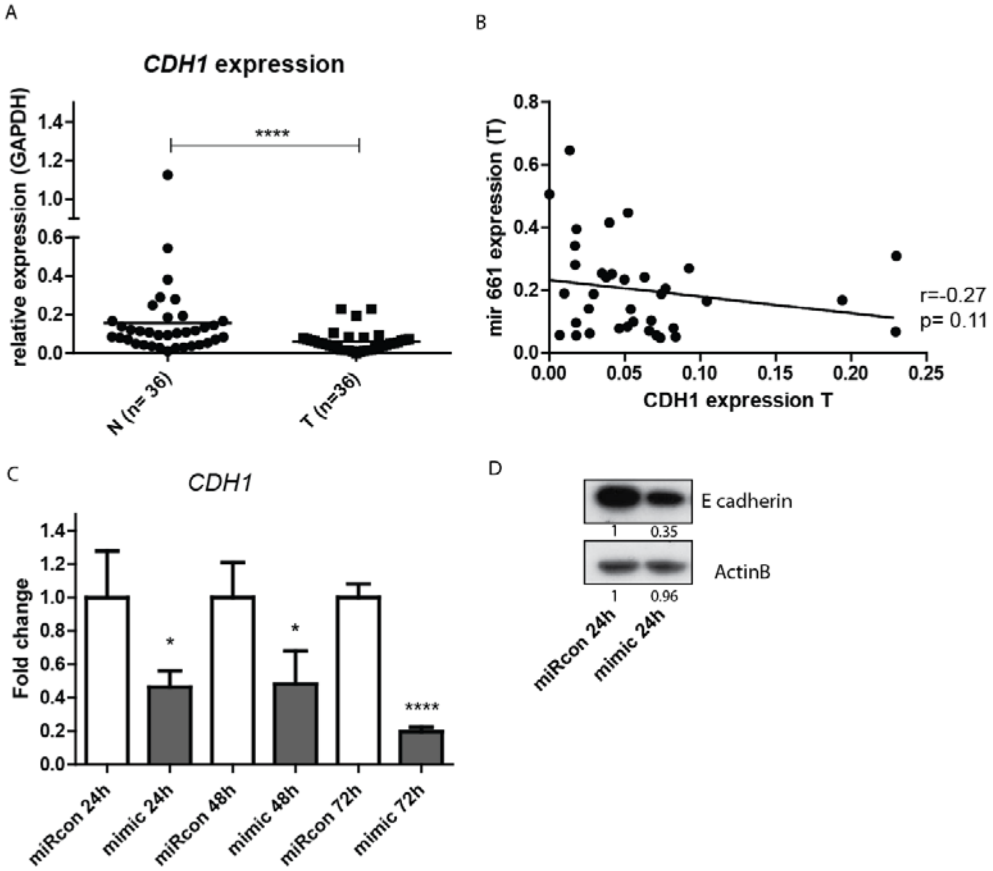


Figure 19. *CDH1* is a target of miR-661. A. *CDH1* mRNA expression in normal (N) and tumor (T) tissue. **B**. No significant correlation between miR 661 expression and *CDH1* in tumor tissue. (Spearman r= -0.27 *p* = 0.11). **C-D**. Ectopic overexpression of miR-661 in A549 cells represses *CDH1* mRNA and protein expression. Cells were harvested and mRNA expression and protein levels were analyzed 24h post-treatment. *GAPDH* and ACTINB were used as housekeeping genes for normalization and loading control. The band intensities were quantified with ImageJ.

4.3.2.5 MEKK3 and E-cadherin are direct targets of miR-661

To link the direct targeting by miR-661 to the target genes, a co-transfection was carried out with mimic 661 and a 3´UTR luciferase reporter containing the binding sites for miR-661 in HEK293T cells (Illustrated in Figure 20A-B). A significant effect on luciferase activity was obtained for *CDH1* and *MAP3K3*. The relative luciferase activity for the *CDH1* was reduced to 57% by miR-661. A reversed effect was seen after site directed mutagenesis at the seed regions, indicating that the seed sequences are functional binding sites for the miRNA (Figure 20C). The relative luciferase activity for *MAP3K3* 3´UTR reporter was reduced to 70% by miR-661 and the effect was reversed when three seed regions were mutated (Figure 20D), suggesting that *MAP3K3* is deregulated in lung cancer by miR-661. Reliable and reproducible data was not obtained *for DIRAS3, RIPK2* and *GAS7* (data not shown).

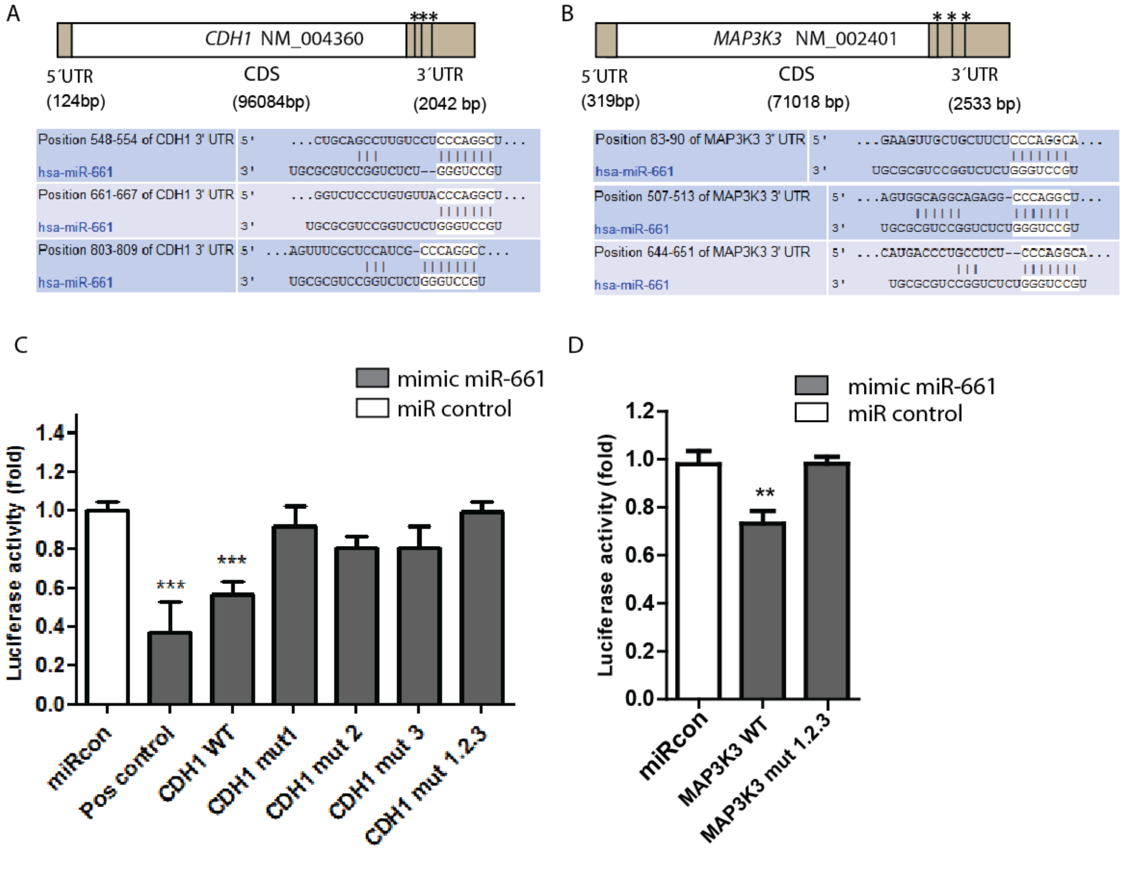


Figure 20. miR-661 directly targets *CDH1* and *MAP3K3*. A-B. Schematic figure of the three 8-mer seed regions for miR-661 at the *CDH1* 3´UTR and *MAP3K3* 3´UTR (Image adapted from TargetScan, version 6.2). *CDH1* and *MAP3K3* 3´UTRs containing three seed regions for miR-661 were cloned into pmiR luciferase vector. **C-D.** HEK293T cells were transfected with *CDH1* 3´UTR or *MAP3K3* 3´UTR pMiR luciferase reporters and co-transfected with mimic miR-661 or miRNA control (miRcon).. The positive control is a 4 time repeat of the mature miR-661 sequence. Statistical analysis employed the unpaired two sided T-test (**p*<0.05, ***p* <0.001, ****p* <0.0001). The graphs show average values with standard deviations which were obtained from least 6 biological replicates for the luciferase assay.

4.3.3 miR-661 is induced during TGFβ1 induced EMT

For a cell to lose intercellular junctions and the apical basal polarity and to gain migratory and invasive properties, downregulation of E-cadherin is one of the major hallmarks [169, 170]. TGF β is a multifunctional growth factor which can act as an inducer of invasion and metastasis in epithelial cancer through activation of SMAD proteins which in turn upregulates transcription regulators e.g. TWIST1 and SLUGs that represses *CDH1* by binding to its promoter [171]. TGFβ1 treatment was carried out in order to stimulate EMT in lung cancer cells to determine the effect on miR-661 expression during EMT. The treatment was carried out in the A549 cell line which expresses E-cadherin and is classified as a lowly invasive cell line [172]. miR-661 expression was induced after 12, 24 and 144h during the transition from epithelial to mesenchymal (Figure 21). *SNAI1*, *Fibronectin 1* and *Collagen (COL1A1)* served as positive controls for the mesenchymal phenotype and E-cadherin served as a control for the loss of epithelial characteristics. The results indicates that miR-661 may play a role during EMT by, together with other regulators, targeting E -Cadherin.

.

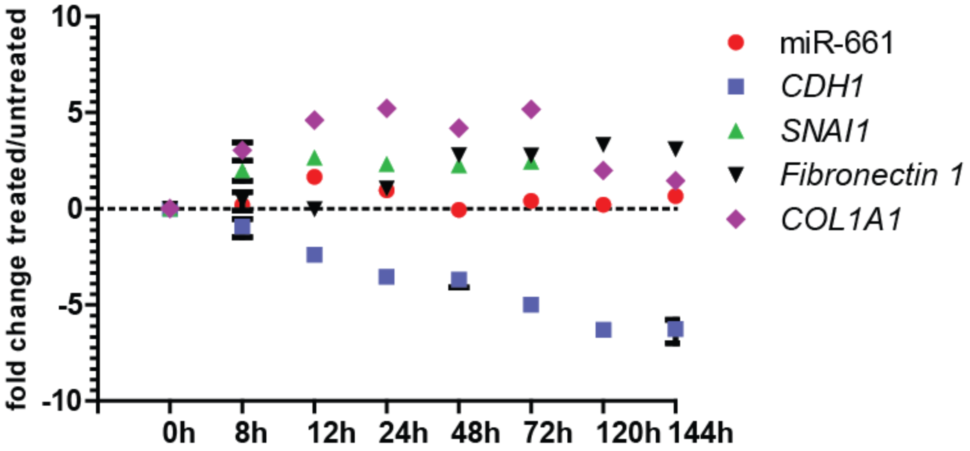


Figure 21. Upregulation of miR-661 in TGF β1 stimulated A549 cells. A549 cells were treated with 5ng/ml TGFβ 1 during 144 h. RNA expression of miR 661 (red), *CDH1* (blue), *SNAI1* (green), *Fibronectin 1* (black) and *COL1A1* (purple) was measured at 0, 8, 12, 24, 48, 72, 120, and 144h. The mRNA expression was normalized to *GAPDH*. The graph shows the average fold change from triplicates in mRNA expression comparing treated with non-treated cells. The error bars show the standard deviation from three replications. The experiment was repeated three times.

4.4 Chromosome 11p15.5

4.4.1 Expression analysis of genes and miRNAs on chromosome 11p15.5

The two prediction algorithms detected 11p15.5 as a gain in cases. In order to elucidate the mechanism by which this may affect lung tumorigenesis, an expression analysis was performed using RT qPCR in two separate sample sets; set 1 for miRNA expression and set 2 for gene expression (See clinical characteristics in Table 6). We extended the gene expression to the maximal gain region for the 11p15.5 CNV and obtained expression results for seven genes and one miRNA (Figure 22). The most striking change was seen for miR-210 where all the tumors were highly upregulated (Figure 22G). Upregulation was also seen for *PKP3* in 63% of the tumors (Figure 22B). *SIGIRR, PTDSS2, RNH1, HRAS* and *DEAF1* were significantly downregulated (Wilcoxon matched pairs signed rank test *p*< 0.0001) (Figure 22 C-F and I) and *IRF7* showed no significant difference between tumor and normal (Figure 22H). Reliable expression data could not be obtained for *B4GALNT4, ANO9, RASSF7 and PHRF1* (data not shown).

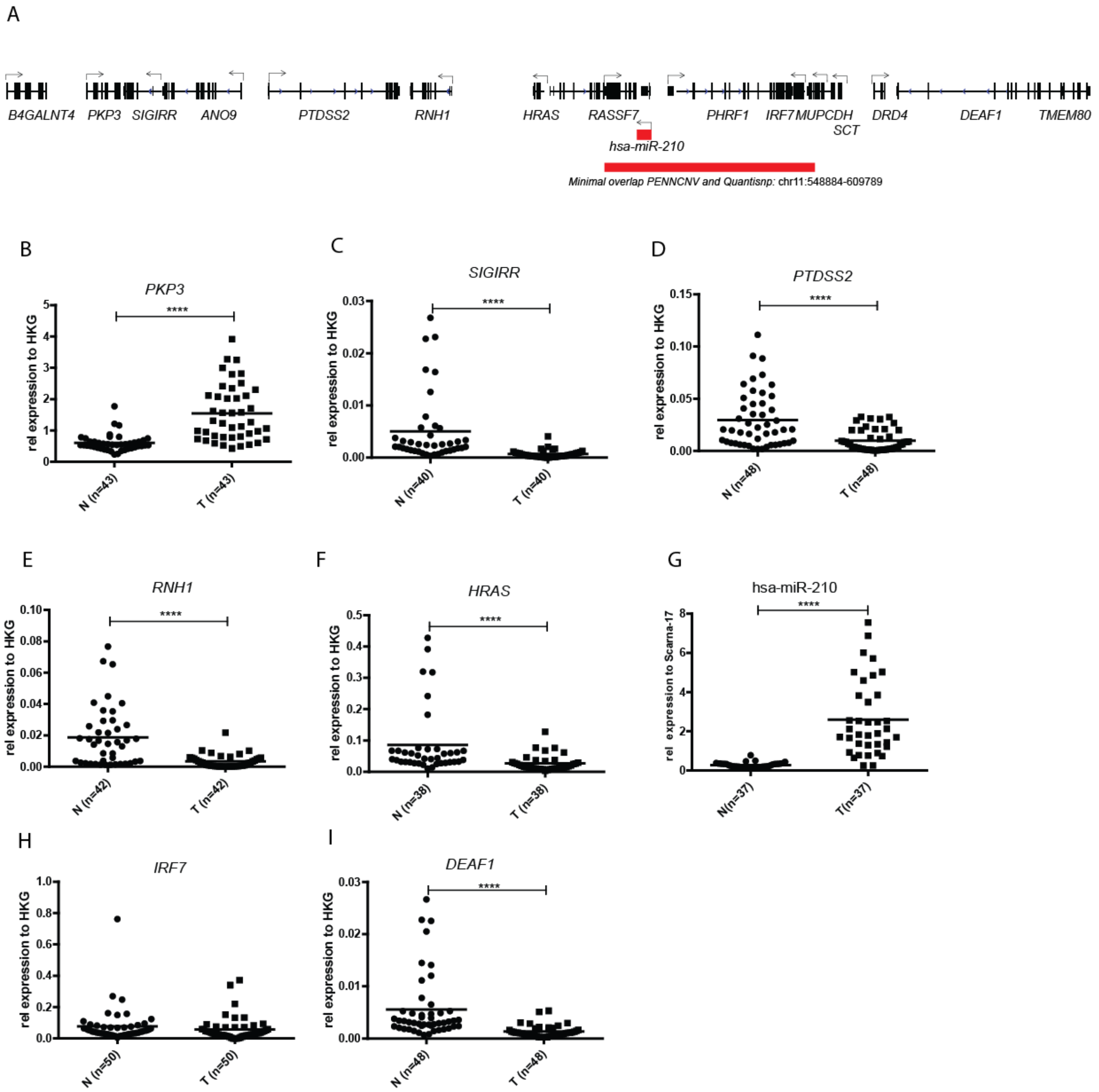


Figure 22. mRNA and miRNA expression of genes on chromosome 11p15.5 in normal lung (N) and lung tumor (T) A. Schematic view of the genomic location of genes on 11p15.5. The detected minimal overlapped CNV is shown in red. **B-I.** mRNA expression analysis of *PKP3*, *SIGIRR, PTDSS2, RNH1, HRAS, IRF7* and *DEAF1* was carried out in set 2 and miR-210 expression analysis was carried out in set 1. Wilcoxon matched pairs signed rank test was carried out to determine significance.

4.4.1.1 PKP3 methylation and survival analysis

PKP3 encodes a member of the armadillo-like proteins Plakophilins which are expressed in desmosomes of epithelial cells. *PKP3* have been reported to be upregulated in NSCLC and to be associated with disease progression [173, 174]. However, how this gene is regulated is unknown. In order to determine whether DNA methylation is regulating the *PKP3* gene in lung cancer, quantitative methylation analysis was carried out in 46 tumor-normal pairs at the transcription start site of *PKP3*. The amplicon was designed to cover a CpG island located in the first exon and overlapped with histone modification marks associated with active transcription (H3K4me3) (ENCODE project [161]) (Figure 23A). The overall amplicon methylation showed a significant reduction in tumor tissue where 55% of the tumors showed >10% hypomethylation (Figure 23B-C). The average methylation over the amplicon in tumor was inversely correlated with expression in tumor which suggests that PKP3 is epigenetically regulated in lung cancer (Spearman r= -0.49, *p*= 0.0009) (Figure 22D). The hypomethylation was confirmed in an independent sample set (Figure S4, set 3, Table 6). To determine whether methylation state was associated with worse outcome, we separated the sample set (n=80) into two groups; low (< 39%) and high (> 39%) methylation based on median methylation in normal tissue. The five year survival for low methylation state was 52.5 % and for high methylation state 71.4%. We did not observe a significant difference between the states (*p* =0.18, HR 1.67, 95% C.I. 0.79-3.50) (Figure 23E). However, the trend indicates that low methylation state may contribute to worse outcome.

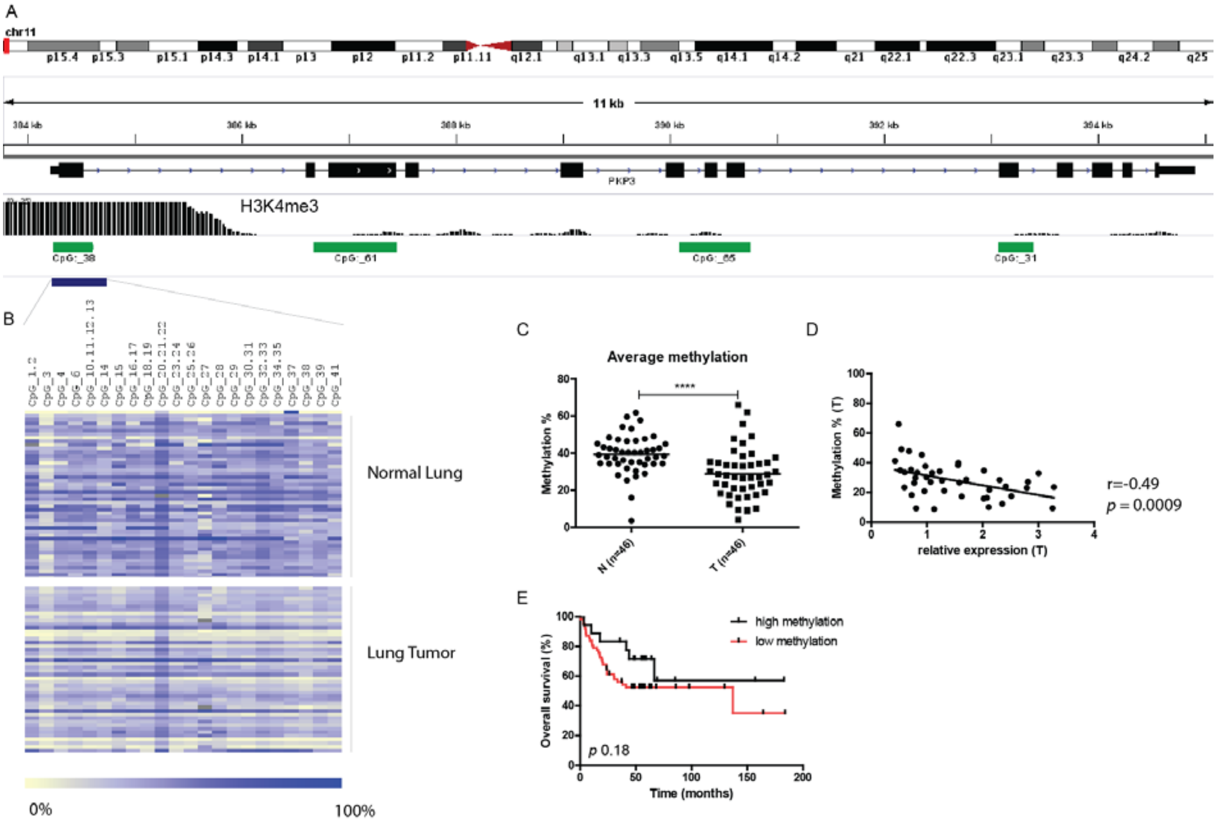


Figure 23. *PKP3* is deregulated by hypomethylation in lung cancer. Quantitative methylation analysis was carried out in 46 normal (N) and 46 tumor (T) tissues at the transcription start site of *PKP3*. **A.** Schematic overview of the *PKP3* locus. The amplicon analyzed is shown in blue overlapping with the active transcription mark H3K4me3 from cell lines indicated in black bars (ENCODE project) and CpG island shown in green. **B**. The heatmap shows methylation ranging from 0 (yellow) to 100% (dark blue) per CpG site or CpG unit in columns. Each row represents one tissue sample. The first 46 rows show the normal lung tissue and the following 46 rows show the matched tumor tissue. **C.** Average % methylation over the amplicon is significantly hyp0methylated in tumor (p < 0.0001, Wilcoxon matched pairs signed rank test). **D**. The average % methylation in tumor tissue is plotted against the *PKP3* mRNA expression in tumor (Spearman r =-0.49, ***p*= 0.0009). **E**. Kaplan Meier survival analysis of 80 patients grouped by PKP3 methylation.

4.4.1.2 miRNA 210 targets RUNX3 in lung cancer

Several studies have shown that miR-210 is induced during hypoxia and has been shown to be upregulated in several cancers [175-178]. In the search for unique targets of miR-210, we combined lung cancer expression data of the top hit targets identified by TargetScan, miRanda and miRDB, and further investigated the runt- related transcription factor *RUNX3*. This gene encodes a transcription factor which is known to form a complex with Smads, the transducer of TGF β signaling and is required for the TGF β mediated induction of p21, a negative regulator of the cell cycle, and of *BIM*, a proapoptotic gene [179]'[180]. It has also been shown to act as a tumor suppressor in lung adenocarcinoma [181]. mRNA expression was determined in 36 tumor-normal pairs and 100% showed downregulation in tumor (p< 0.00001, Wilcoxon matched pairs signed rank test) (Figure 24A). The expression of *RUNX3* showed a significant inverse correlation with miR-210 expression in tumor, suggesting that the *RUNX3* gene is a direct target of miR-210 (Spearman r = -0.50, p= 0.005) (Figure 24B). To strengthen this hypothesis, we ectopically overexpressed miR210 in lung cancer cell lines and observed a significant reduction of the endogenous mRNA level of *RUNX3* (Figure 24C). In order to elucidate the direct pairing of the miR-210 to the seed sequence in the 3´UTR, we performed a luciferase reporter assay with the cloned 3´UTR including the seed region for miR-210. The luciferase activity was reduced to 50% when miR210 mimic was co transfected. These results suggest a role of miR-210 in the regulation of *RUNX3*.

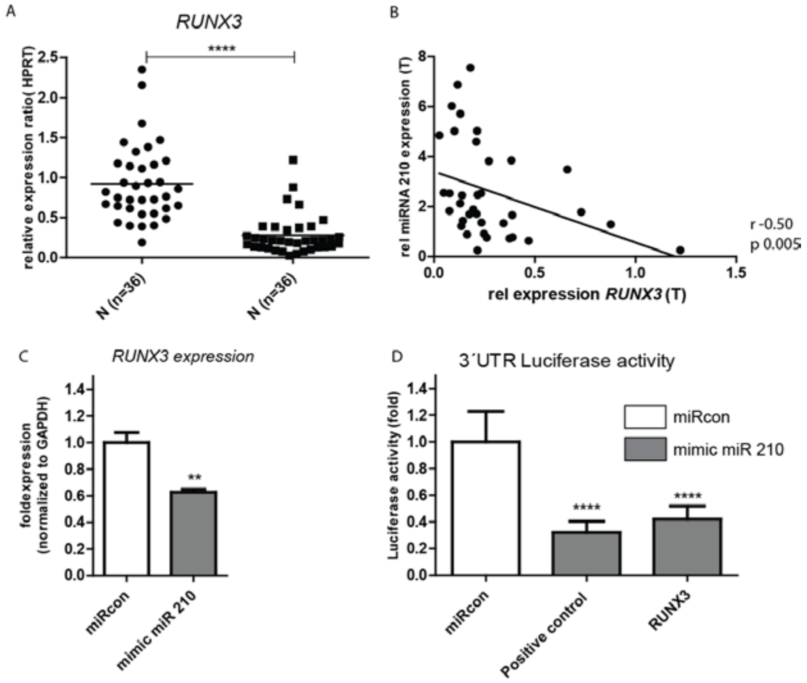


Figure 24. *RUNX3* is a target of miR-210. A. *RUNX3* mRNA expression was quantified in normal 36 (N) and tumor (T) tissues and was significantly downregulated in T (p< 0.0001, Wilcoxon matched pairs signed rank test). **B**. The relative expression of miR-210 in tumor was inversely correlated with relative *RUNX3* mRNA expression in tumor (Spearman r = -0.50, *p*= 0.005). **C.** miR-210 was ectopically overexpressed in H1299 cells and *RUNX3* mRNA expression was significantly reduced to 62% upon overexpression (p 0.001, two sided, unpaired T test). **D**. The 3´UTR including the seed region for miR 210 was cloned into a pMIR luciferase reporter assay and co tranfected with either miRcontrol or mimic mir210. The luciferase activity shows the average fold change from least 6 biological repeats and the error bars show standard deviations. A two sided unpaired T test was used to determine significance (*p* <0.05).

5. Discussion

Genome wide association studies have identified several risk loci in lung cancer that increase the susceptibility of developing the disease. However, the functionality of these variations and how they are regulated is not well known. In the current study, we identified and investigated germline CNVs associated with early-onset lung cancer risk and the potential function and epigenetic regulation of a subset of genes and miRNAs located in two copy number gain regions on chromosome 8q24.3 and 11p15.5.

5.1 CNV detection on GWA data for early-onset lung cancer

Two CNV detection algorithms, PennCNV and QuantiSNP, were applied on the SNP array data from a previously performed GWA study in early onset lung cancer. We focused on the CNVs identified by both methods to decrease the false positive CNVs. To use more than one algorithm for detection of CNVs is useful as the discrepancy between different software tools has been shown to be high [182, 183]. With this strategy 25 CNVs were detected with both tools (Table S7). To determine which of these CNVs were associated with lung cancer risk, Fisher´s exact test was performed on each SNP in the CNV region and a combined p-value for the region was obtained using Stouffer's test [156]. Ten CNVs were identified to be associated with early-onset risk using both algorithms (Table 7).

The current study focused on 8q24.3 and 11p15.5 based on the following criterias: 1) CNV was only present in cases, and 2) CNVs harbors miRNAs and genes with potential tumor-relevant functions. This is not to exclude that the other regions e.g. those without annotated genes (CNVs on 8q24.23, 6q12, and 19p12), are not important. In fact, most SNPs associated with disease risk map to non-coding regions in the genome [184]. Recently, a comprehensive integrative approach was carried out combining epigenome data, transcription factor binding sites and breast cancer associated SNPs to reveal the functional relationship between them. The study found that risk SNPs were enriched for FOXA1 and ESR1 transcription factor binding sites and H3K4me1 histone modifications, affecting the binding affinity to chromatin [185]. SNPs on 8q24 associated with prostate cancer risk have also been found in FOXA1 binding sites [186]. Moreover, we identified a loss on 1q21.1 partly overlapping with a CNV that has previously been associated with neuroblastoma [131]. The previous study identified this region in 15% of the cases and 9% of the controls. In contrast to their study, we identified a higher proportion of controls (3.4% cases and 7.6% of the controls with PennCNV and 4.4% cases and 5.7% of the controls with QuantiSNP). The fact that this region was more common in controls in our study may indicate that this particular CNV is not a suitable risk marker. Furthermore, our study identified two copy number loss regions on 11q11, both harboring variants of the olfactory receptor family (*OR4P4*, *OR4S2* and *OR4C11*). Interestingly, these regions are known as common CNVs in the genome [71, 112, 118, 119, 159]. Surprisingly, they appear significantly associated with early-onset lung cancer risk in our study ($p< 10^{-18}$ and p=0.003). This finding highlights the importance of using an integrated approach combining GWA data with functional analysis to identify loci that have an impact on the disease. It has also been suggested that some common CNV regions appear only present in a fraction of blood cells, e.g. T cell receptors or immunoglobulin related genes as the results may be effected by different cell type composition between cases and controls [72, 142, 187]. Moreover, the 10 CNVs found to be associated with early-onset lung cancer did not replicate in a study carried out on more elderly lung cancer patients (unpublished data, data not shown), suggesting that these CNVs may be specific susceptibility loci in early-onset lung cancer. Further replication studies in larger populations focusing on early-onset lung cancer populations will be necessary to support our findings. Taken together, among the 10 CNVs associated with early-onset lung cancer, the copy number gain regions observed only in cases for 8q24.3 and 11p15.5 appeared to be the most promising CNV regions to further investigate for functionality and relevance for lung tumorigenesis.

5.2 Method optimization for copy number analysis with the TyperAssay

The initial aim was to carry out a protocol optimization of the allele specific copy number (ACN) TyperAssay application from Sequenom in order to establish a cost effective, quantitative, high-throughput method with which several regions can be analyzed simultaneously in one sample. Such a method would be a very useful tool for replication studies of germline CNVs associated with risk. For the protocol optimization, we used samples with known copy numbers (homozygous deletion, 1 copy, and 2 copies) for *GSTM1* based on results from a qPCR based method published previously in our group [144]. *GSTM1* copy number was analyzed and could be confirmed quantitatively by the TyperAssay (Figure 7A). We observed that one of the parameters for reliability of the assay lies in the optimal determination of the competitor concentration which we showed was crucial for the quantification of the exact copy number (Figure 7B). An alternative for the synthetic competitor has been proposed by Williams *et al.* in 2008, using chimpanzee gDNA as the competing strand. The advantage of this alternative is the use of only one extra competing DNA for all regions therefore decreasing the complexity of the multiplex PCR [188].

5.2.1 CNV validation with TyperAssay

To validate the CNV gains on 8q24.3 and 11p15.5 detected by QuantiSNP and PennCNV, we designed assays for SNPs located in the overlapped region. The results from the TyperAssay protocol could not confirm the copy number gain in the tested samples (Figure 8 and Figure 9). Very recently, the first comprehensive report using the MassARRAY platform for CNV analysis was published [189]. In contrast to our protocol, this study bases normalization only on an endogenous control (2N) and does not include the competitor sequence. The advantage of that study was the use of samples where copy numbers for both loss and gain were known i.e. included robust controls for both variants were available. A positive control for duplications was missing in our assay. Another paper published by Gaudam *et al.,* report a similar strategy as ours using the competitor PCR. However, for copy number determination, that study normalized the EC50 values for each sample to the EC50 value for the 2N control (three 2N controls included). They used the method for validation of deletions, duplications and 2N regions identified from the Affymetrix 50k Chip array. The overall validation concordance was 35% for duplications, 10% for deletions and 66% concordance for 2N regions [190], indicating that in this study copy number gain regions were better validated than regions with loss.

A second attempt to validate the CNVs on 8q24.3 and 11p15.5 was to use FISH on lymphocytes from a subset of the patients with detected CNVs. The weakness of the setup of this assay was the probe size (Figure 10A). The probe size for FISH is optimal with at least 100kb to reach a sufficient fluorescent signal. However, to determine a copy number gain which may be in tandem and have a size of 200kb, it is not possible to distinguish two separate dots (Figure 11). The probes we used were between 80kb and 100 kb and the in most cases expected CNV size was around 100-200kb. We observed an increased fluorescent signal on one allele in a subset of cells and hypothesized that these cells may have an extra copy (Figure 11). With quantification of the signal intensities, significance could not be reached. To improve this method for our particular purpose, custom-designed probes hybridizing to a smaller region is desirable for signal separation on the alleles. Taken together, neither TyperAssay CNV analysis nor FISH analysis, allowed a final conclusion on whether the CNV gains on 8q24.3 or 11p15.5 are present. Alternative approaches for technical validation such as next generation sequencing of CNV regions are likely to provide a better understanding of these candidate CNV regions.

5.3 The role of 8q24.3 in lung cancer

In the current study we identified a germline CNV gain on 8q24.3 detected in 27 (QuantiSNP) or 18 cases (PennCNV). This region has previously not been reported to be associated with lung cancer risk. We compared data from the Database for genomic variants (DGV) to determine whether this is a commonly variable region in the population [115]. This region has been reported as a gain with low frequency in four healthy population studies [191-193]. Jakobsson *et al* found this region in 1/443 controls, Pinto *et al* in 1/6 and Park *et al* in 1/31 controls. This indicates that observed gain may specifically be associated with early-onset lung cancer risk.

5.3.1 Aberrant expression and regulation of miR-661

One aim of the current study was to answer the question whether CNVs associated with lung cancer risk harbor genes or miRNAs that play a role in lung cancer progression. Since the algorithms detected this CNV as a gain in cases, we were searching specifically for oncogenic potential of this region. We hypothesized that copy number gain regions harbor upregulated genes. Expression of genes and miRNA in matched lung tumor-normal pairs was determined by with qPCR and the results showed that most genes were downregulated in tumor compared to normal tissue. However, miR-661 was significantly upregulated in the majority of the tumors (Figure 14). Genes and miRNAs in this region are likely tightly regulated on an individual level. In addition, it has been suggested that CNV miRNAs are more susceptible to gene-dosage than other genes due to their functional importance in the cell [73]. miR-661 belongs to the group of non-conserved miRNAs and shares high sequence homology only with Chimpanzee and Rhesus [194]. miR-661 is located within the second intron of the *PLEC1* gene. This gene consists of 8 annotated isoforms in the human genome but the abundance of the transcripts is not well reported. It has been proposed that the transcription of these isoforms may be tissue specific [195]. The expression was analyzed specifically for the isoform 6, in which the miR-661 is located in the first intron. The expression of the isoform 6 was downregulated in lung cancer. This finding suggests that the miR-661 transcript and the host-gene are not co-regulated. This is in accordance with a study which showed that younger intragenic miRNAs more often than old ones, have evolved differently and therefore have their own promoter [196].

The 8q24.3 region has been reported in lung cancer as an amplified region in less than 5% of the cases [197]. Therefore, we hypothesized that other mechanism than a gene dosage dependent effect may play a role in the regulation of miR-661. DNA methylation was analyzed at a 5kb distal CpG island in the same sample set and we found a potential promoter region of miR-661 to be hypomethylated and significantly negatively correlated with miR-661 expression (Figure 15B). Our finding goes in line with other studies that have shown that miRNAs in the genome are frequently deregulated by methylation in cancer (reviewed in [80]). Additional studies on the exact transcriptional start site for the primary miR-661 are necessary to be certain that the sequence investigated is in fact the regulatory region. We provide data showing that promoter activity of these sites decreases when the demethylated region is excluded, demonstrating that the sequence where hypomethylation takes place is important for active transcription (Figure 15E).

5.3.2 Clinical relevance of miR-661 hypomethylation

Methylation at the putative promoter of miR-661 was significantly correlated with expression in tumor tissue (Figure 15B and E). However, hypomethylation was observed in a subset of the patients, which made us hypothesize that methylation at this locus may be useful as a prognostic marker. DNA methylation as a prognostic marker has been reported for other cancers, e.g. HPV driven oral squamous cell carcinoma [81]. We observed a significant association with overall survival revealing that low methylation is associated with worse outcome (Figure 16). The survival data is marginally significant, and thus, an extension of the study with a larger study population is necessary to strengthen our finding and its potential use in the clinics for prognosis.

5.3.3 miR-661 targets

miR-661 has more than 2000 predicted targets in the genome and to find those that may affect pathways involved in lung tumorigenesis is challenging. miR-661 has been reported in the context of breast cancer [168, 198], but its role in lung cancer is not known. Two approaches were used to find gene targets for the miR-661 in lung cancer. The first strategy was to combine the top hits from five prediction algorithms (TargetScan 6.2, Diana microT-v5, miRanda, miRDB and miRwalk [163-167]). From among the target genes identified, a selection of genes was further investigated based on their high scores from the prediction algorithms. We investigated the interaction between miR-661 and the four target genes *MAP3K3, DIRAS3, RIPK2,* and *GAS7* (Figure 18). Among them, *MAP3K3* appeared to be a direct target of miR-661 supported by the 3´UTR luciferase reporter system in cell lines and including a site directed mutagenesis approach (Figure 20). This protein plays a role in the canonical activating pathway of NFkB [199]. Although overexpression of *MAP3K3* has been shown to correlate with NFkB activity and tumor progression in breast and ovarian cancer [200], another study has revealed a potential tumor suppressor role of *MAP3K3* by its involvement in cell cycle arrest by suppressing cyclin D1 [201]. Our study shows that *MAP3K3* is significantly downregulated in NSCLC, suggesting that the function of *MAP3K3* may be cell type specific. The results reveal a link between miR-661 upregulation and *MAP3K3* downregulation, which may indicate that *MAP3K3* plays an unknown tumor suppressor role in lung cancer. Further studies on the function of *MAP3K3* in lung cancer are needed to support this hypothesis.

The second approach was to focus on EMT related targets since it was previously shown that miR-661 may be an early regulator in this process [168]. Interestingly, we found the major marker for EMT, E-cadherin, to be a predicted target and hypothesized that upregulation of miR-661 in cells undergoing EMT contributes to invasiveness and metastasis via E-cadherin regulation. In the publication by Vetter *et al*. E-cadherin is used as a negative control and E-cadherin is said to not have any seed sequences for miR-661 in the 3´UTR [168]. The use of an earlier version of the MiRBASE prediction tool may be the reason for this contradiction. Our results demonstrate that ectopic overexpression of miR-661 reduced the mRNA and protein level of E-cadherin significantly in A549 lung cancer cells. Additionally, a direct link between the 3´UTR of E-cadherin and overexpression of miR-661 was supported by a 3´UTR luciferase reporter assay.

The function of E-cadherin in cancer during invasion and metastasis is well characterized [169, 170, 202] and several different mechanisms for its regulation have been reported. For example, downregulation by DNA methylation has been reported in several epithelial cancers [203-209]. The epigenetic regulatory mechanism for *CDH1* which encodes E-cadherin has been shown to be controlled by the transcriptional repressor SNAIL1-G9a-DNMT1 complex where SNAIL1 directs the DNMTs to specific CG sites at the *CDH1* promoter [210]. Other repressors e.g. the ZEB family is also responsible for deregulation of E-cadherin. Additionally, miR-9 and mir-10b target E-cadherin in cancer [88, 89] as well as the mir200 family through indirect regulation via targeting the ZEB protein family [87, 211]. Our study uncovers an additional regulation of E-cadherin in lung cancer by miR-661. Heterogeneous mRNA expression was observed between the patients analyzed and a correlation could not be shown between the *CDH1* expression and miR-661 expression (Figure 19). This may be explained by the fact that downregulation of *CDH1* only occurs in cells undergoing EMT and these cells are located in the leading front of the tumor. Given that the RNA was isolated without selecting for this particular subpopulation of tumor cells this effect is diluted. We addressed this issue by using immunohistochemistry (IHC) to determine whether the E-cadherin expression was reduced specifically at the invasive front (Figure S3). However, E-cadherin was strongly expressed in the epithelial cells in the majority of the samples tested and it was not possible to distinguish quantitatively whether the invasive front showed a different expression pattern. Further supporting results e.g. using additional EMT markers, may be helpful to elucidate an EMT phenotype in these patients.

We further analyzed whether miR-661 expression is affected by inducing EMT with TGF β treatment in A549 cells over time and if the treatment changed the expression of the epithelial marker *CDH1* and the mesenchymal markers, *SNAI1*, *COL1A1* and *Fibronectin 1* (Figure 21). We observed a statistically significant increase of miR-661 expression after 48h and 144h treatment. However, the induction was minor in comparison with the other mesenchymal markers. TGF β1 is known to activate the SMAD proteins which then relocate to the nucleus where they interact with the transcriptional repressor SNAIL1 [171]. Additionally, it has been reported that SNAIL1 is also upregulating Fibronectin 1 [212]. In line with this, we observed that the induction of *Fibronectin 1* expression was at a later time point than the *SNAI1* gene.

.

5.4 The role of 11p15.5 in lung cancer

Both PennCNV and QuantiSNP identified germline copy number gain at 11p15.5 in a small subset of early-onset lung cancer. As described for 8q24.3, expression analysis was performed on tumor and matched normal lung tissue to answer the question whether this predicted copy number gain region harbors potentially relevant genes for tumorigenesis.

5.4.1 *Plakophilin 3 (PKP3)*

In the CNV region on 11p15.5 *PKP3* was found to be upregulated in NSCLC (Figure 22A). Furukawa *et al*. found that *PKP3* upregulation in lung tumors correlated with worse prognosis and could serve as a potential therapeutic target in lung cancer [174]. Additional evidence for *PKP3* upregulation in malignancies comes from studies in breast and prostate cancers [213, 214]. *PKP3* encodes a member of the arm-repeat (armadillo) and plakophilin gene families and is localized to cell desmosomes and in the nuclei, and participate in linking cadherins to intermediate filaments in the cytoskeleton [173, 215]. The mechanism how this gene may be upregulated in lung cancer is not known. Evidence in colon cancer has shown that the ZEB transcriptional regulator is responsible for repression of this gene [216]. We provide data revealing that the *PKP3* promoter DNA methylation is decreased in tumor tissue and that the methylation pattern in tumor tissue inversely correlates with expression (Figure 23). To determine whether the hypomethylation would be suitable as a prognostic factor for survival, we divided the samples in high or low methylation state based on the average methylation at the transcription start site. The five year survival for low methylation state was 52.5 % and for high methylation state 71.4% but revealed no significant difference. Even if the difference is not statistically significant, the trend suggest that hypomethylation of *PKP3* may have a negative impact on overall survival. Further investigation in a larger sample set is required to strengthen our finding.

5.4.2 miR-210

miR-210 is among the top upregulated miRNAs in lung cancer [217] and is a potential diagnostic marker as it is detected and associated with lung cancer in sputum [75, 76]. We could confirm the upregulation of miR-210 in NSCLC in our study (Figure 22G). Several targets for mir-210 have been identified and it is evident that this miRNA is a key player in hypoxia, cell survival and migration [177, 178, 218, 219]. In the current study, *RUNX3* was shown to be a promising candidate target gene of miR-210 (Figure 24). Knockout mouse studies have shown an essential role for *RUNX3* in lung development and a functional inhibitory role on lung epithelial proliferation during late phases of lung development [180, 181]. The results in our study indicate an interaction between miR-210 and *RUNX3* supported by ectopic miRNA overexpression experiments in lung cancer cell lines and a 3´UTR luciferase reporter assay (Figure 24). Further studies to support a direct interaction and effect on *RUNX3* suppression, should include site directed mutagenesis of the seed sequences to determine whether the luciferase activity can be rescued. We further show that *RUNX3* is downregulated in NSCLC and significantly correlates negatively with miR-210 expression (Figure 24). *RUNX3* downregulation was shown in NSCLC before and is supported by several studies showing that *RUNX3* is hypermethylated, especially in lung adenocarcinoma [220-224]. Additionally, downregulation of *RUNX3* has been detected in peripheral whole blood in an early stage lung cancer case-control study, suggesting that RUNX3 may serve as a biomarker for lung cancer [225]. The results in the our study suggest miR-210 as an additional epigenetic regulator of *RUNX3* in lung cancer.

6. Strengths and weaknesses of the study

A limitation of the current study is the absence of a verification of the CNVs in a larger population study in early-onset lung cancer. Thus, a replication study is necessary to verify the associations and to answer questions regarding the possible impact of these copy number variations on early-onset risk. In addition, the 550k SNP array from Illumina is limited regarding CNVs because of the initial selection criteria of SNPs for the arrays where common CNV regions in segmental duplications have been excluded. Other platforms e.g. the Human SNP array 6.0 platforms by Affimetrix or Illumina´s 1 M SNP BeadChip have included common CNV regions and would be better suitable for CNV studies. Furthermore, the current study did not take into account stratification by ethnicity, gender, histology and smoking status which should in the future be considered. Moreover, a limitation of the study was that the GWA study included cases from all histologies but the expression and methylation analyses were carried out in NSCLC due to material not being available for the other subtypes.

To our knowledge, the GWA study on which this work is based is the only one to date that identified CNVs associated with risk of early-onset lung cancer. The advantage of studying this subtype of lung cancer is the strong link to heritability in comparison to older lung cancer cases [17, 21], suggesting that genetic variations may play a bigger role in this subtype in the susceptibility to the disease.

Genome wide association studies have identified several risk loci for almost all cancers and the field is now moving in the direction to identify the functional impact these regions may have on the disease [226]. A particular feature of this study is its width starting with a genome-wide association study as a platform for identifying important risk loci for early-onset lung cancer to further investigate the potential functionality of these regions on the tumorigenesis.

One of the aims of the current study was to establish a robust protocol for quantitative CNV analysis in blood. *GSTM1* deletions, 1 copy and 2 copies could be achieved with the TyperAssay, however, copy number gain on 8q24.3 and 11p15.5 could not be determined. There are several different variables that should be considered for further optimization of the protocol: robust controls for 2N regions for normalization, regions with known 3N or more copies and alternatives for the competitor strand that would be more equal to gDNA and thus increase the sensitivity of the assay. Furthermore, a verification of whether the copy number gains in blood would lead to acquired copy number alterations in tumor tissue is missing in the current study. This link is missing in this study due to lack of available quantitative copy number analysis methods and availability of somatic tissue from cases included in the GWAs.

7. Conclusion and outlook

The current study provides new insights into how GWAs could be used to point towards functionally relevant regions in the genome that may directly or indirectly modulate known oncogenic pathways in lung cancer. We identified ten novel CNVs associated with early – onset lung cancer and focused further functional studies on two copy number gain regions of 8q24.3 and 11p15.5. This study provides evidence that two CNV regions associated with early-onset lung cancer risk harbor aberrantly expressed miRNAs and genes in lung cancer and that epigenetic deregulation is responsible for fine tuning of the expression of specifically miR-661 on 8q24.3 and *PKP3* on 11p15.5. Thus this study underscores the importance of combining analysis of genetic and epigenetic information. Additional functional analyses in lung cancer cell lines of miR-661 and miR-210 revealed their potential oncogenic features in lung cancer by direct targeting of the genes *MAP3K3*, *CDH1* and *RUNX3*, which are all involved in tumor associated pathways with potential tumor suppressor functions.

To identify and further understand the importance of heritable susceptibility loci for lung cancer, meta and pooled analyses for available GWA data are now ongoing and recently published [227]. This approach could also be used in the future to perform CNV analyses. Although there are limitations imposed by the old array technologies used, the vast numbers of GWA resources now available would enable subgroup analyses.

It would be of interest to look in detail at the regions identified in this study without gene-annotations since it has been shown that most SNPs associated with disease risk are located in intergenic regions [184]. These regions may consist of regulatory enhancers, noncoding RNAs, repressors or chromatin structures. Identification of such possible regulatory elements would be an important step in understanding the functional effect of these risk loci.

The role of miR-661 in lung cancer has not been studied before. Further investigations of the function of miR-661 as a potentical oncomiR and its role in EMT would help understanding whether this miRNA may be a potential therapeutic target in lung cancer. *In vivo* models would be useful to determine the possible oncogenic feature and involvement in EMT or other mechanistically relevant pathways for lung cancer. One big challenge in understanding the functional role of a miRNA in the cell, is the identification of relevant target genes. In depth studies of the key pathways regulated by miR-661 could include expression profiling in the absence of or with stable overexpression of miR-661. Another way to identify specific targets in lung cancer cells could be immunoprecipiation of the RISC complex in miR-661 induced system and therefore identify which targets are bound to miR-661.

Another important future approach should be focused on exploring whether germline CNVs leads to acquired copy number changes in somatic tissue and whether a direct dose dependency can be verified. This could be an important step towards better understanding the functionality and the effect of germline CNVs associated with risk of tumorigenesis.

.

References

1. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell 2000;100(1):57-70.
2. Haber DA, Settleman J. Cancer: drivers and passengers. Nature 2007;446(7132):145-146.
3. Simpson AJ. Sequence-based advances in the definition of cancer-associated gene mutations. Curr Opin Oncol 2009;21(1):47-52.
4. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011;144(5):646-674.
5. Ferlay J, Shin HR, Bray F*, et al.* Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int J Cancer 2010;127(12):2893-2917.
6. Rekhtman N. Neuroendocrine tumors of the lung: an update. Arch Pathol Lab Med 2010;134(11):1628-1638.
7. Travis WD. Pathology of lung cancer. Clin Chest Med 2002;23(1):65-81, viii.
8. Devesa SS, Bray F, Vizcaino AP*, et al.* International lung cancer trends by histologic type: male:female differences diminishing and adenocarcinoma rates rising. Int J Cancer 2005;117(2):294-299.
9. Sun S, Schiller JH, Gazdar AF. Lung cancer in never smokers--a different disease. Nat Rev Cancer 2007;7(10):778-790.
10. Stellman SD, Muscat JE, Thompson S*, et al.* Risk of squamous cell carcinoma and adenocarcinoma of the lung in relation to lifetime filter cigarette smoking. Cancer 1997;80(3):382-388.
11. Witschi H. A short history of lung cancer. Toxicol Sci 2001;64(1):4-6.
12. Belinsky SA. Gene-promoter hypermethylation as a biomarker in lung cancer. Nat Rev Cancer 2004;4(9):707-717.
13. Mountain CF, Hermes KE. Surgical treatment of lung cancer. Past and present. Methods Mol Med 2003;75:453-487.
14. Rosenberger A, Illig T, Korb K*, et al.* Do genetic factors protect for early onset lung cancer? A case control study before the age of 50 years. BMC Cancer 2008;8:60.
15. Boffetta P, Kreuzer M, Benhamou S*, et al.* Risk of lung cancer from tobacco smoking among young women from Europe. Int J Cancer 2001;91(5):745-746.
16. Bourke W, Milstein D, Giura R*, et al.* Lung cancer in young adults. Chest 1992;102(6):1723-1729.
17. Kreuzer M, Kreienbrock L, Gerken M*, et al.* Risk factors for lung cancer in young adults. Am J Epidemiol 1998;147(11):1028-1037.
18. Timofeeva MN, Kropp S, Sauter W*, et al.* CYP450 polymorphisms as risk factors for early-onset lung cancer: gender-specific differences. Carcinogenesis 2009;30(7):1161-1169.
19. Sauter W, Rosenberger A, Beckmann L*, et al.* Matrix metalloproteinase 1 (MMP1) is associated with early-onset lung cancer. Cancer Epidemiol Biomarkers Prev 2008;17(5):1127-1135.
20. Timofeeva M, Kropp S, Sauter W*, et al.* Genetic polymorphisms of MPO, GSTT1, *GSTM1*, GSTP1, EPHX1 and NQO1 as risk factors of early-onset lung cancer. Int J Cancer 2010;127(7):1547-1561.
21. Bromen K, Pohlabeln H, Jahn I*, et al.* Aggregation of lung cancer in families: results from a population-based case-control study in Germany. Am J Epidemiol 2000;152(6):497-505.
22. Travis WD. Pathology of lung cancer. Clin Chest Med 2011;32(4):669-692.

23. Pao W, Girard N. New driver mutations in non-small-cell lung cancer. Lancet Oncol 2011;12(2):175-180.
24. Seo JS, Ju YS, Lee WC*, et al.* The transcriptional landscape and mutational profile of lung adenocarcinoma. Genome Res 2012;22(11):2109-2119.
25. Cancer Genome Atlas Research N. Comprehensive genomic characterization of squamous cell lung cancers. Nature 2012;489(7417):519-525.
26. Dutt A, Ramos AH, Hammerman PS*, et al.* Inhibitor-sensitive FGFR1 amplification in human non-small cell lung cancer. PLoS One 2011;6(6):e20351.
27. Hammerman PS, Sos ML, Ramos AH*, et al.* Mutations in the DDR2 kinase gene identify a novel therapeutic target in squamous cell lung cancer. Cancer Discov 2011;1(1):78-89.
28. Travis WD, Brambilla E, Noguchi M*, et al.* International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. J Thorac Oncol 2011;6(2):244-285.
29. Govindan R, Ding L, Griffith M*, et al.* Genomic landscape of non-small cell lung cancer in smokers and never-smokers. Cell 2012;150(6):1121-1134.
30. Riggs AD, Jones PA. 5-methylcytosine, gene regulation, and cancer. Adv Cancer Res 1983;40:1-30.
31. Strahl BD, Allis CD. The language of covalent histone modifications. Nature 2000;403(6765):41-45.
32. Struhl K, Segal E. Determinants of nucleosome positioning. Nat Struct Mol Biol 2013;20(3):267-273.
33. Kelly TK, Miranda TB, Liang G*, et al.* H2A.Z maintenance during mitosis reveals nucleosome shifting on mitotically silenced genes. Mol Cell 2010;39(6):901-911.
34. Heintzman ND, Hon GC, Hawkins RD*, et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 2009;459(7243):108-112.
35. Costello JF, Plass C. Methylation matters. J Med Genet 2001;38(5):285-303.
36. Ehrlich M. DNA methylation in cancer: too much, but also too little. Oncogene 2002;21(35):5400-5413.
37. Hashimshony T, Zhang J, Keshet I*, et al.* The role of DNA methylation in setting up chromatin structure during development. Nat Genet 2003;34(2):187-192.
38. Zhang X, Ho SM. Epigenetics meets endocrinology. J Mol Endocrinol 2011;46(1):R11-32.
39. Irizarry RA, Ladd-Acosta C, Wen B*, et al.* The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet 2009;41(2):178-186.
40. Feinberg AP, Vogelstein B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. Nature 1983;301(5895):89-92.
41. Dunn BK. Hypomethylation: one side of a larger picture. Ann N Y Acad Sci 2003;983:28-42.
42. Feinberg AP, Vogelstein B. Hypomethylation of ras oncogenes in primary human cancers. Biochem Biophys Res Commun 1983;111(1):47-54.
43. Radhakrishnan VM, Jensen TJ, Cui H*, et al.* Hypomethylation of the 14-3-3sigma promoter leads to increased expression in non-small cell lung cancer. Genes Chromosomes Cancer 2011;50(10):830-836.

44. Stewart DJ. Tumor and host factors that may limit efficacy of chemotherapy in non-small cell and small cell lung cancer. Crit Rev Oncol Hematol 2010;75(3):173-234.
45. Esteller M. Epigenetics in cancer. N Engl J Med 2008;358(11):1148-1159.
46. Esteller M, Fraga MF, Guo M*, et al.* DNA methylation patterns in hereditary human cancers mimic sporadic tumorigenesis. Hum Mol Genet 2001;10(26):3001-3007.
47. Cairns P, Polascik TJ, Eby Y*, et al.* Frequency of homozygous deletion at p16/CDKN2 in primary human tumours. Nat Genet 1995;11(2):210-212.
48. Merlo A, Herman JG, Mao L*, et al.* 5' CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers. Nat Med 1995;1(7):686-692.
49. Rusin MR, Okamoto A, Chorazy M*, et al.* Intragenic mutations of the p16(INK4), p15(INK4B) and p18 genes in primary non-small-cell lung cancers. Int J Cancer 1996;65(6):734-739.
50. Belinsky SA, Palmisano WA, Gilliland FD*, et al.* Aberrant promoter methylation in bronchial epithelium and sputum from current and former smokers. Cancer Res 2002;62(8):2370-2377.
51. Vuillemenot BR, Pulling LC, Palmisano WA*, et al.* Carcinogen exposure differentially modulates RAR-beta promoter hypermethylation, an early and frequent event in mouse lung carcinogenesis. Carcinogenesis 2004;25(4):623-629.
52. Pulling LC, Vuillemenot BR, Hutt JA*, et al.* Aberrant promoter hypermethylation of the death-associated protein kinase gene is early and frequent in murine lung tumors induced by cigarette smoke and tobacco carcinogens. Cancer Res 2004;64(11):3844-3848.
53. Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 1993;75(5):843-854.
54. Ambros V, Lee RC, Lavanway A*, et al.* MicroRNAs and other tiny endogenous RNAs in C. elegans. Curr Biol 2003;13(10):807-818.
55. Croce CM. Causes and consequences of microRNA dysregulation in cancer. Nat Rev Genet 2009;10(10):704-714.
56. Chang TC, Wentzel EA, Kent OA*, et al.* Transactivation of miR-34a by p53 broadly influences gene expression and promotes apoptosis. Mol Cell 2007;26(5):745-752.
57. Lee Y, Kim M, Han J*, et al.* MicroRNA genes are transcribed by RNA polymerase II. EMBO J 2004;23(20):4051-4060.
58. Berezikov E, Chung WJ, Willis J*, et al.* Mammalian mirtron genes. Mol Cell 2007;28(2):328-336.
59. Ruby JG, Jan CH, Bartel DP. Intronic microRNA precursors that bypass Drosha processing. Nature 2007;448(7149):83-86.
60. Yi R, Qin Y, Macara IG*, et al.* Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. Genes Dev 2003;17(24):3011-3016.
61. Bhayani MK, Calin GA, Lai SY. Functional relevance of miRNA sequences in human disease. Mutat Res 2012;731(1-2):14-19.
62. Vasudevan S, Tong Y, Steitz JA. Switching from repression to activation: microRNAs can up-regulate translation. Science 2007;318(5858):1931-1934.
63. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 2004;116(2):281-297.

64. Takamizawa J, Konishi H, Yanagisawa K*, et al.* Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival. Cancer Res 2004;64(11):3753-3756.

65. Johnson SM, Grosshans H, Shingara J*, et al.* RAS is regulated by the let-7 microRNA family. Cell 2005;120(5):635-647.

66. Johnson CD, Esquela-Kerscher A, Stefani G*, et al.* The let-7 microRNA represses cell proliferation pathways in human cells. Cancer Res 2007;67(16):7713-7722.

67. Xiong S, Zheng Y, Jiang P*, et al.* MicroRNA-7 inhibits the growth of human non-small cell lung cancer A549 cells through targeting BCL-2. Int J Biol Sci 2011;7(6):805-814.

68. Hayashita Y, Osada H, Tatematsu Y*, et al.* A polycistronic microRNA cluster, miR-17-92, is overexpressed in human lung cancers and enhances cell proliferation. Cancer Res 2005;65(21):9628-9632.

69. Oglesby IK, McElvaney NG, Greene CM. MicroRNAs in inflammatory lung disease--master regulators or target practice? Respir Res 2010;11:148.

70. Li C, Nguyen HT, Zhuang Y*, et al.* Comparative profiling of miRNA expression of lung adenocarcinoma cells in two-dimensional and three-dimensional cultures. Gene 2012;511(2):143-150.

71. Redon R, Ishikawa S, Fitch KR*, et al.* Global variation in copy number in the human genome. Nature 2006;444(7118):444-454.

72. Wong KK, deLeeuw RJ, Dosanjh NS*, et al.* A comprehensive analysis of common copy-number variations in the human genome. Am J Hum Genet 2007;80(1):91-104.

73. Marcinkowska M, Szymanski M, Krzyzosiak WJ*, et al.* Copy number variation of microRNA genes in the human genome. BMC Genomics 2011;12:183.

74. Felekkis K, Voskarides K, Dweep H*, et al.* Increased number of microRNA target sites in genes encoded in CNV regions. Evidence for an evolutionary genomic interaction. Mol Biol Evol 2011;28(9):2421-2424.

75. Yu L, Todd NW, Xing L*, et al.* Early detection of lung adenocarcinoma in sputum by a panel of microRNA markers. Int J Cancer 2010;127(12):2870-2878.

76. Xing L, Todd NW, Yu L*, et al.* Early detection of squamous cell lung cancer in sputum by a panel of microRNA markers. Mod Pathol 2010;23(8):1157-1164.

77. Boeri M, Verri C, Conte D*, et al.* MicroRNA signatures in tissues and plasma predict development and prognosis of computed tomography detected lung cancer. Proc Natl Acad Sci U S A 2011;108(9):3713-3718.

78. Kumar MS, Erkeland SJ, Pester RE*, et al.* Suppression of non-small cell lung tumor development by the let-7 microRNA family. Proc Natl Acad Sci U S A 2008;105(10):3903-3908.

79. Baer C, Claus R, Frenzel LP*, et al.* Extensive promoter DNA hypermethylation and hypomethylation is associated with aberrant microRNA expression in chronic lymphocytic leukemia. Cancer Res 2012;72(15):3775-3785.

80. Baer C, Claus R, Plass C. Genome-wide epigenetic regulation of miRNAs in cancer. Cancer Res 2013;73(2):473-477.

81. Kostareli E, Holzinger D, Bogatyrova O*, et al.* HPV-related methylation signature predicts survival in oropharyngeal squamous cell carcinomas. J Clin Invest 2013;123(6):2488-2501.

82. Cao J, Song Y, Bi N*, et al.* DNA Methylation-Mediated Repression of miR-886-3p Predicts Poor Outcome of Human Small Cell Lung Cancer. Cancer Res 2013;73(11):3326-3335.

83. Heller G, Weinzierl M, Noll C*, et al.* Genome-wide miRNA expression profiling identifies miR-9-3 and miR-193a as targets for DNA methylation in non-small cell lung cancers. Clin Cancer Res 2012;18(6):1619-1629.

84. Brueckner B, Stresemann C, Kuner R*, et al.* The human let-7a-3 locus contains an epigenetically regulated microRNA gene with oncogenic function. Cancer Res 2007;67(4):1419-1423.

85. Fabbri M, Garzon R, Cimmino A*, et al.* MicroRNA-29 family reverts aberrant methylation in lung cancer by targeting DNA methyltransferases 3A and 3B. Proc Natl Acad Sci U S A 2007;104(40):15805-15810.

86. Thiery JP, Acloque H, Huang RY*, et al.* Epithelial-mesenchymal transitions in development and disease. Cell 2009;139(5):871-890.

87. Gregory PA, Bert AG, Paterson EL*, et al.* The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. Nat Cell Biol 2008;10(5):593-601.

88. Ma L, Teruya-Feldstein J, Weinberg RA. Tumour invasion and metastasis initiated by microRNA-10b in breast cancer. Nature 2007;449(7163):682-688.

89. Ma L, Young J, Prabhala H*, et al.* miR-9, a MYC/MYCN-activated microRNA, regulates E-cadherin and cancer metastasis. Nat Cell Biol 2010;12(3):247-256.

90. Nakata S, Sugio K, Uramoto H*, et al.* The methylation status and protein expression of CDH1, p16(INK4A), and fragile histidine triad in nonsmall cell lung carcinoma: epigenetic silencing, clinical features, and prognostic significance. Cancer 2006;106(10):2190-2199.

91. Thiery JP. Epithelial-mesenchymal transitions in tumour progression. Nat Rev Cancer 2002;2(6):442-454.

92. Thiery JP, Sleeman JP. Complex networks orchestrate epithelial-mesenchymal transitions. Nat Rev Mol Cell Biol 2006;7(2):131-142.

93. Sato M, Shames DS, Hasegawa Y. Emerging evidence of epithelial-to-mesenchymal transition in lung carcinogenesis. Respirology 2012;17(7):1048-1059.

94. Al-Saad S, Al-Shibli K, Donnem T*, et al.* The prognostic impact of NF-kappaB p105, vimentin, E-cadherin and Par6 expression in epithelial and stromal compartment in non-small-cell lung cancer. Br J Cancer 2008;99(9):1476-1483.

95. Yoo JY, Yang SH, Lee JE*, et al.* E-cadherin as a predictive marker of brain metastasis in non-small-cell lung cancer, and its regulation by pioglitazone in a preclinical model. J Neurooncol 2012;109(2):219-227.

96. Hasegawa Y, Takanashi S, Kanehira Y*, et al.* Transforming growth factor-beta1 level correlates with angiogenesis, tumor progression, and prognosis in patients with nonsmall cell lung carcinoma. Cancer 2001;91(5):964-971.

97. Hung JJ, Yang MH, Hsu HS*, et al.* Prognostic significance of hypoxia-inducible factor-1alpha, TWIST1 and Snail expression in resectable non-small cell lung cancer. Thorax 2009;64(12):1082-1089.

98. Yanagawa J, Walser TC, Zhu LX*, et al.* Snail promotes CXCR2 ligand-dependent tumor progression in non-small cell lung carcinoma. Clin Cancer Res 2009;15(22):6820-6829.

99. Amos CI, Xu W, Spitz MR. Is there a genetic basis for lung cancer susceptibility? Recent Results Cancer Res 1999;151:3-12.

100. Hung RJ, McKay JD, Gaborieau V*, et al.* A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. Nature 2008;452(7187):633-637.

101. Amos CI, Wu X, Broderick P*, et al.* Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nat Genet 2008;40(5):616-622.
102. McKay JD, Hung RJ, Gaborieau V*, et al.* Lung cancer susceptibility locus at 5p15.33. Nat Genet 2008;40(12):1404-1406.
103. Wang Y, Broderick P, Webb E*, et al.* Common 5p15.33 and 6p21.33 variants influence lung cancer risk. Nat Genet 2008;40(12):1407-1409.
104. Spitz MR, Amos CI, Dong Q*, et al.* The CHRNA5-A3 region on chromosome 15q24-25.1 is a risk factor both for nicotine dependence and for lung cancer. J Natl Cancer Inst 2008;100(21):1552-1556.
105. Thorgeirsson TE, Geller F, Sulem P*, et al.* A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature 2008;452(7187):638-642.
106. Paliwal A, Vaissiere T, Krais A*, et al.* Aberrant DNA methylation links cancer susceptibility locus 15q25.1 to apoptotic regulation and lung cancer. Cancer Res 2010;70(7):2779-2788.
107. Scherf DB, Sarkisyan N, Jacobsson H*, et al.* Epigenetic screen identifies genotype-specific promoter DNA methylation and oncogenic potential of CHRNB4. Oncogene 2013;32(28):3329-3338.
108. Brenner DR, Brennan P, Boffetta P*, et al.* Hierarchical modeling identifies novel lung cancer susceptibility variants in inflammation pathways among 10,140 cases and 11,012 controls. Hum Genet 2013;132(5):579-589.
109. Shi J, Chatterjee N, Rotunno M*, et al.* Inherited variation at chromosome 12p13.33, including RAD52, influences the risk of squamous cell lung carcinoma. Cancer Discov 2012;2(2):131-139.
110. Spitz MR, Gorlov IP, Dong Q*, et al.* Multistage analysis of variants in the inflammation pathway and lung cancer risk in smokers. Cancer Epidemiol Biomarkers Prev 2012;21(7):1213-1221.
111. Hu Z, Chen J, Tian T*, et al.* Genetic variants of miRNA sequences and non-small cell lung cancer survival. J Clin Invest 2008;118(7):2600-2608.
112. Conrad DF, Pinto D, Redon R*, et al.* Origins and functional impact of copy number variation in the human genome. Nature 2010;464(7289):704-712.
113. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat Rev Genet 2006;7(2):85-97.
114. Freeman JL, Perry GH, Feuk L*, et al.* Copy number variation: new insights in genome diversity. Genome Res 2006;16(8):949-961.
115. Iafrate AJ, Feuk L, Rivera MN*, et al.* Detection of large-scale variation in the human genome. Nat Genet 2004;36(9):949-951.
116. Sharp AJ, Locke DP, McGrath SD*, et al.* Segmental duplications and copy-number variation in the human genome. Am J Hum Genet 2005;77(1):78-88.
117. Valsesia A, Stevenson BJ, Waterworth D*, et al.* Identification and validation of copy number variants using SNP genotyping arrays from a large clinical cohort. BMC Genomics 2012;13:241.
118. Kidd JM, Cooper GM, Donahue WF*, et al.* Mapping and sequencing of structural variation from eight human genomes. Nature 2008;453(7191):56-64.
119. Korbel JO, Urban AE, Affourtit JP*, et al.* Paired-end mapping reveals extensive structural variation in the human genome. Science 2007;318(5849):420-426.
120. McCarroll SA, Kuruvilla FG, Korn JM*, et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet 2008;40(10):1166-1174.

121. Grozeva D, Kirov G, Ivanov D*, et al.* Rare copy number variants: a point of rarity in genetic risk for bipolar disorder and schizophrenia. Arch Gen Psychiatry 2010;67(4):318-327.

122. Zhang F, Gu W, Hurles ME*, et al.* Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet 2009;10:451-481.

123. Lupski JR. Genome structural variation and sporadic disease traits. Nat Genet 2006;38(9):974-976.

124. Li FP, Fraumeni JF, Jr. Rhabdomyosarcoma in children: epidemiologic study and identification of a familial cancer syndrome. J Natl Cancer Inst 1969;43(6):1365-1373.

125. Li FP, Fraumeni JF, Jr. Soft-tissue sarcomas, breast cancer, and other neoplasms. A familial syndrome? Ann Intern Med 1969;71(4):747-752.

126. Li FP, Fraumeni JF, Jr., Mulvihill JJ*, et al.* A cancer family syndrome in twenty-four kindreds. Cancer Res 1988;48(18):5358-5362.

127. Malkin D. p53 and the Li-Fraumeni syndrome. Biochim Biophys Acta 1994;1198(2-3):197-213.

128. Malkin D. Li-fraumeni syndrome. Genes Cancer 2011;2(4):475-484.

129. Shlien A, Tabori U, Marshall CR*, et al.* Excessive genomic DNA copy number variation in the Li-Fraumeni cancer predisposition syndrome. Proc Natl Acad Sci U S A 2008;105(32):11264-11269.

130. Demichelis F, Setlur SR, Banerjee S*, et al.* Identification of functionally active, low frequency copy number variants at 15q21.3 and 12q21.31 associated with prostate cancer risk. Proc Natl Acad Sci U S A 2012;109(17):6686-6691.

131. Diskin SJ, Hou C, Glessner JT*, et al.* Copy number variation at 1q21.1 associated with neuroblastoma. Nature 2009;459(7249):987-991.

132. Young RP, Hopkins RJ, Hay BA*, et al. GSTM1* null genotype in COPD and lung cancer: evidence of a modifier or confounding effect? Appl Clin Genet 2011;4:137-144.

133. Houlston RS. Glutathione S-transferase M1 status and lung cancer risk: a meta-analysis. Cancer Epidemiol Biomarkers Prev 1999;8(8):675-682.

134. Butler MW, Hackett NR, Salit J*, et al.* Glutathione S-transferase copy number variation alters lung gene expression. Eur Respir J 2011;38(1):15-28.

135. Johansson I, Lundqvist E, Bertilsson L*, et al.* Inherited amplification of an active gene in the cytochrome P450 CYP2D locus as a cause of ultrarapid metabolism of debrisoquine. Proc Natl Acad Sci U S A 1993;90(24):11825-11829.

136. Rodriguez-Antona C, Gomez A, Karlgren M*, et al.* Molecular genetics and epigenetics of the cytochrome P450 gene family and its relevance for cancer risk and treatment. Hum Genet 2010;127(1):1-17.

137. Liu B, Yang L, Huang B*, et al.* A functional copy-number variation in MAPKAPK2 predicts risk and prognosis of lung cancer. Am J Hum Genet 2012;91(2):384-390.

138. Yang L, Liu B, Huang B*, et al.* A functional copy number variation in the WWOX gene is associated with lung cancer risk in Chinese. Hum Mol Genet 2013;22(9):1886-1894.

139. Gunderson KL, Steemers FJ, Lee G*, et al.* A genome-wide scalable SNP genotyping assay using microarray technology. Nat Genet 2005;37(5):549-554.

140. Steemers FJ, Gunderson KL. Illumina, Inc. Pharmacogenomics 2005;6(7):777-782.

141. Valsesia A, Mace A, Jacquemont S*, et al.* The Growing Importance of CNVs: New Insights for Detection and Clinical Interpretation. Front Genet 2013;4:92.
142. Wang K, Li M, Hadley D*, et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 2007;17(11):1665-1674.
143. Colella S, Yau C, Taylor JM*, et al.* QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res 2007;35(6):2013-2025.
144. Timofeeva M, Jager B, Rosenberger A*, et al.* A multiplex real-time PCR method for detection of *GSTM1* and GSTT1 copy numbers. Clin Biochem 2009;42(6):500-509.
145. Landegent JE, Vanommen GJB, Baas F*, et al.* High-Sensitivity Insitu Hybridization of Human Single-Copy Genes Using 2-Acetyl-Aminofluorene (Aaf)-Modified DNA Probes and Reflection Contrast Microscopy. Cytogenetics and Cell Genetics 1985;40(1-4):677-677.
146. Zeng FY, Ren ZR, Huang SZ*, et al.* Array-MLPA: Comprehensive detection of deletions and duplications and its application to DMD patients. Human Mutation 2008;29(1):190-197.
147. Wichmann HE, Gieger C, Illig T*, et al.* KORA-gen--resource for population genetics, controls and a broad spectrum of disease phenotypes. Gesundheitswesen 2005;67 Suppl 1:S26-30.
148. Organization WH. Histological Typing of Lung Tumours. 2004, 3rd ed. Lyon:IARC
149. Ding C, Cantor CR. A high-throughput gene expression analysis technique using competitive PCR and matrix-assisted laser desorption ionization time-of-flight MS. Proceedings of the National Academy of Sciences 2003;100(6):3059-3064.
150. Siebert PD, Larrick JW. PCR MIMICS: competitive DNA fragments for use as internal standards in quantitative PCR. Biotechniques 1993;14(2):244-249.
151. International Human Genome Sequencing C. Finishing the euchromatic sequence of the human genome. Nature 2004;431(7011):931-945.
152. Kristensen LS, Hansen LL. PCR-based methods for detecting single-locus DNA methylation biomarkers in cancer diagnostics, prognostics, and response to treatment. Clin Chem 2009;55(8):1471-1483.
153. Fan F, Wood KV. Bioluminescent assays for high-throughput screening. Assay Drug Dev Technol 2007;5(1):127-136.
154. Klug M, Rehli M. Functional analysis of promoter CpG methylation using a CpG-free luciferase reporter vector. Epigenetics 2006;1(3):127-130.
155. Pappa G, Strathmann J, Lowinger M*, et al.* Quantitative combination effects between sulforaphane and 3,3'-diindolylmethane on proliferation of human colon cancer cells in vitro. Carcinogenesis 2007;28(7):1471-1477.
156. Whitlock MC. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. J Evol Biol 2005;18(5):1368-1373.
157. Abel U, Berger J, Wiebelt H. CRITLEVEL: an exploratory procedure for the evaluation of quantitative prognostic factors. Methods Inf Med 1984;23(3):154-156.
158. Cox DR. Regression models and life tables (with discussion). J R Stat Soc 1972;34:187-200.
159. Sebat J, Lakshmi B, Troge J*, et al.* Large-scale copy number polymorphism in the human genome. Science 2004;305(5683):525-528.

160. Yu SH, Graf WD, Shprintzen RJ. Genomic disorders on chromosome 22. Current Opinion in Pediatrics 2012;24(6):665-671.

161. Consortium EP. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 2004;306(5696):636-640.

162. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 2013;14(2):178-192.

163. Garcia DM, Baek D, Shin C*, et al.* Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. Nat Struct Mol Biol 2011;18(10):1139-1146.

164. Vlachos IS, Kostoulas N, Vergoulis T*, et al.* DIANA miRPath v.2.0: investigating the combinatorial effect of microRNAs in pathways. Nucleic Acids Res 2012;40(Web Server issue):W498-504.

165. Betel D, Koppal A, Agius P*, et al.* Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biol 2010;11(8):R90.

166. Wang X. miRDB: a microRNA target prediction and functional annotation database with a wiki interface. RNA 2008;14(6):1012-1017.

167. Dweep H, Sticht C, Pandey P*, et al.* miRWalk--database: prediction of possible miRNA binding sites by "walking" the genes of three genomes. J Biomed Inform 2011;44(5):839-847.

168. Vetter G, Saumet A, Moes M*, et al.* miR-661 expression in SNAI1-induced epithelial to mesenchymal transition contributes to breast cancer cell invasion by targeting Nectin-1 and StarD10 messengers. Oncogene 2010;29(31):4436-4448.

169. Frixen UH, Behrens J, Sachs M*, et al.* E-cadherin-mediated cell-cell adhesion prevents invasiveness of human carcinoma cells. J Cell Biol 1991;113(1):173-185.

170. Vleminckx K, Vakaet L, Jr., Mareel M*, et al.* Genetic manipulation of E-cadherin expression by epithelial tumor cells reveals an invasion suppressor role. Cell 1991;66(1):107-119.

171. Massague J. TGFbeta signalling in context. Nat Rev Mol Cell Biol 2012;13(10):616-630.

172. Kim AN, Jeon WK, Lim KH*, et al.* Fyn mediates transforming growth factor-beta1-induced down-regulation of E-cadherin in human A549 lung cancer cells. Biochem Biophys Res Commun 2011;407(1):181-184.

173. Bonne S, van Hengel J, Nollet F*, et al.* Plakophilin-3, a novel armadillo-like protein present in nuclei and desmosomes of epithelial cells. J Cell Sci 1999;112 (Pt 14):2265-2276.

174. Furukawa C, Daigo Y, Ishikawa N*, et al.* Plakophilin 3 oncogene as prognostic marker and therapeutic target for lung cancer. Cancer Res 2005;65(16):7102-7110.

175. Esquela-Kerscher A, Slack FJ. Oncomirs - microRNAs with a role in cancer. Nat Rev Cancer 2006;6(4):259-269.

176. Chan YC, Banerjee J, Choi SY*, et al.* miR-210: the master hypoxamir. Microcirculation 2011;19(3):215-223.

177. Puissegur MP, Mazure NM, Bertero T*, et al.* miR-210 is overexpressed in late stages of lung cancer and mediates mitochondrial alterations associated with modulation of HIF-1 activity. Cell Death Differ 2011;18(3):465-478.

178. Chan SY, Loscalzo J. MicroRNA-210: a unique and pleiotropic hypoxamir. Cell Cycle 2010;9(6):1072-1083.

179. Chi XZ, Yang JO, Lee KY*, et al.* RUNX3 suppresses gastric epithelial cell growth by inducing p21(WAF1/Cip1) expression in cooperation with transforming growth factor {beta}-activated SMAD. Mol Cell Biol 2005;25(18):8097-8107.

180. Lee JM, Kwon HJ, Bae SC*, et al.* Lung tissue regeneration after induced injury in Runx3 KO mice. Cell Tissue Res 2010;341(3):465-470.

181. Lee KS, Lee YS, Lee JM*, et al.* Runx3 is required for the differentiation of lung epithelial cells and suppression of lung cancer. Oncogene 2010;29(23):3349-3361.

182. Winchester L, Yau C, Ragoussis J. Comparing CNV detection methods for SNP arrays. Brief Funct Genomic Proteomic 2009;8(5):353-366.

183. Zhang D, Qian Y, Akula N*, et al.* Accuracy of CNV Detection from GWAS Data. PLoS One 2011;6(1):e14511.

184. Frazer KA, Murray SS, Schork NJ*, et al.* Human genetic variation and its contribution to complex traits. Nat Rev Genet 2009;10(4):241-251.

185. Cowper-Sal lari R, Zhang X, Wright JB*, et al.* Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. Nat Genet 2012;44(11):1191-1198.

186. Jia L, Landan G, Pomerantz M*, et al.* Functional enhancers at the gene-poor 8q24 cancer-linked locus. PLoS Genet 2009;5(8):e1000597.

187. Tuzun E, Sharp AJ, Bailey JA*, et al.* Fine-scale structural variation of the human genome. Nat Genet 2005;37(7):727-732.

188. Williams NM, Williams H, Majounie E*, et al.* Analysis of copy number variation using quantitative interspecies competitive PCR. Nucleic Acids Research 2008;36(17).

189. Gao Y, Chen X, Wang J*, et al.* A novel approach for copy number variation analysis by combining multiplex PCR with matrix-assisted laser desorption ionization time-of-flight mass spectrometry. Journal of biotechnology 2013.

190. Gautam P, Jha P, Kumar D*, et al.* Spectrum of large copy number variations in 26 diverse Indian populations: potential involvement in phenotypic diversity. Human Genetics 2012;131(1):131-143.

191. Jakobsson M, Scholz SW, Scheet P*, et al.* Genotype, haplotype and copy-number variation in worldwide human populations. Nature 2008;451(7181):998-1003.

192. Park H, Kim JI, Ju YS*, et al.* Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. Nat Genet 2010;42(5):400-405.

193. Pinto D, Darvishi K, Shi X*, et al.* Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat Biotechnol 2011;29(6):512-520.

194. Zhang Z, Schwartz S, Wagner L*, et al.* A greedy algorithm for aligning DNA sequences. Journal of Computational Biology 2000;7(1-2):203-214.

195. Fuchs P, Zorer M, Rezniczek GA*, et al.* Unusual 5 ' transcript complexity of plectin isoforms: novel tissue-specific exons modulate actin binding activity. Human Molecular Genetics 1999;8(13):2461-2472.

196. He CJ, Li ZJ, Chen P*, et al.* Young intragenic miRNAs are less coexpressed with host genes than old ones: implications of miRNA-host gene coevolution. Nucleic Acids Research 2012;40(9):4002-4012.

197. Staaf J, Isaksson S, Karlsson A*, et al.* Landscape of somatic allelic imbalances and copy number alterations in human lung carcinoma. Int J Cancer 2013;132(9):2020-2031.

198. Reddy SD, Pakala SB, Ohshiro K*, et al.* MicroRNA-661, a c/EBPalpha target, inhibits metastatic tumor antigen 1 and regulates its functions. Cancer Res 2009;69(14):5639-5642.

199. Yang J, Lin Y, Guo Z*, et al.* The essential role of MEKK3 in TNF-induced NF-kappaB activation. Nat Immunol 2001;2(7):620-624.

200. Samanta AK, Huang HJ, Bast RC, Jr.*, et al.* Overexpression of MEKK3 confers resistance to apoptosis through activation of NFkappaB. J Biol Chem 2004;279(9):7576-7583.

201. Ellinger-Ziegelbauer H, Kelly K, Siebenlist U. Cell cycle arrest and reversion of Ras-induced transformation by a conditionally activated form of mitogen-activated protein kinase kinase kinase 3. Mol Cell Biol 1999;19(5):3857-3868.

202. Derksen PW, Liu X, Saridin F*, et al.* Somatic inactivation of E-cadherin and p53 in mice leads to metastatic lobular mammary carcinoma through induction of anoikis resistance and angiogenesis. Cancer Cell 2006;10(5):437-449.

203. Chang HW, Chow V, Lam KY*, et al.* Loss of E-cadherin expression resulting from promoter hypermethylation in oral tongue carcinoma and its prognostic significance. Cancer 2002;94(2):386-392.

204. Chiles MC, Ai L, Fan CY*, et al.* E-cadherin promoter hypermethylation in preneoplastic and neoplastic skin lesions. Laboratory Investigation 2003;83(1):89a-89a.

205. Corn PG, Heath EI, Heitmiller R*, et al.* Frequent hypermethylation of the 5 ' CpG island of E-cadherin in esophageal adenocarcinoma. Clinical Cancer Research 2001;7(9):2765-2769.

206. Di Croce L, Pelicci PG. Tumour-associated hypermethylation: silencing E-cadherin expression enhances invasion and metastasis. European Journal of Cancer 2003;39(4):413-414.

207. Garinis GA, Menounos PG, Spanakis NE*, et al.* Hypermethylation-associated transcriptional silencing of E-cadherin in primary sporadic colorectal carcinomas. Journal of Pathology 2002;198(4):442-449.

208. Graff JR, Herman JG, Lapidus RG*, et al.* E-Cadherin Expression Is Silenced by DNA Hypermethylation in Human Breast and Prostate Carcinomas. Cancer Research 1995;55(22):5195-5199.

209. Horikawa Y, Sugano K, Shigyo M*, et al.* Hypermethylation of a E-cadherin (CDH1) promoter region in high grade transitional cell carcinoma of the bladder comprising carcinoma in situ (CIS). Journal of Urology 2003;169(4):187-187.

210. Dong C, Wu Y, Yao J*, et al.* G9a interacts with Snail and is critical for Snail-mediated E-cadherin repression in human breast cancer. J Clin Invest 2012;122(4):1469-1486.

211. Park SM, Gaur AB, Lengyel E*, et al.* The miR-200 family determines the epithelial phenotype of cancer cells by targeting the E-cadherin repressors ZEB1 and ZEB2. Genes Dev 2008;22(7):894-907.

212. Stanisavljevic J, Porta-de-la-Riva M, Batlle R*, et al.* The p65 subunit of NF-kappa B and PARP1 assist Snail1 in activating fibronectin transcription. Journal of Cell Science 2011;124(24):4159-4169.

213. Breuninger S, Reidenbach S, Sauer CG*, et al.* Desmosomal plakophilins in the prostate and prostatic adenocarcinomas: implications for diagnosis and tumor progression. Am J Pathol 2010;176(5):2509-2519.

214. Demirag GG, Sullu Y, Yucel I. Expression of Plakophilins (PKP1, PKP2, and PKP3) in breast cancers. Med Oncol 2012;29(3):1518-1522.

215. Schmidt A, Jager S. Plakophilins--hard work in the desmosome, recreation in the nucleus? Eur J Cell Biol 2005;84(2-3):189-204.

216. Aigner K, Descovich L, Mikula M*, et al.* The transcription factor ZEB1 (deltaEF1) represses Plakophilin 3 during human cancer progression. FEBS Lett 2007;581(8):1617-1624.

217. Guan P, Yin Z, Li X*, et al.* Meta-analysis of human lung cancer microRNA expression profiling studies comparing cancer tissues with normal tissues. J Exp Clin Cancer Res 2012;31:54.

218. Fasanaro P, D'Alessandra Y, Di Stefano V*, et al.* MicroRNA-210 modulates endothelial cell response to hypoxia and inhibits the receptor tyrosine kinase ligand Ephrin-A3. J Biol Chem 2008;283(23):15878-15883.

219. Fasanaro P, Greco S, Lorenzi M*, et al.* An integrated approach for experimental target identification of hypoxia-induced miR-210. J Biol Chem 2009;284(50):35134-35143.

220. Li QL, Kim HR, Kim WJ*, et al.* Transcriptional silencing of the RUNX3 gene by CpG hypermethylation is associated with lung cancer. Biochem Biophys Res Commun 2004;314(1):223-228.

221. Omar MF, Ito K, Nga ME*, et al.* RUNX3 downregulation in human lung adenocarcinoma is independent of p53, EGFR or KRAS status. Pathol Oncol Res 2012;18(4):783-792.

222. Sato K, Tomizawa Y, Iijima H*, et al.* Epigenetic inactivation of the RUNX3 gene in lung cancer. Oncol Rep 2006;15(1):129-135.

223. Yanagawa N, Tamura G, Oizumi H*, et al.* Inverse correlation between EGFR mutation and FHIT, RASSF1A and RUNX3 methylation in lung adenocarcinoma: relation with smoking status. Anticancer Res 2011;31(4):1211-1214.

224. Yu GP, Ji Y, Chen GQ*, et al.* Application of RUNX3 gene promoter methylation in the diagnosis of non-small cell lung cancer. Oncol Lett 2012;3(1):159-162.

225. Rotunno M, Hu N, Su H*, et al.* A gene expression signature from peripheral whole blood for stage I lung adenocarcinoma. Cancer Prev Res (Phila) 2011;4(10):1599-1608.

226. Freedman ML, Monteiro AN, Gayther SA*, et al.* Principles for the post-GWAS functional characterization of cancer risk loci. Nat Genet 2011;43(6):513-518.

227. Timofeeva MN, Hung RJ, Rafnar T*, et al.* Influence of common genetic variation on lung cancer risk: meta-analysis of 14 900 cases and 29 485 controls. Hum Mol Genet 2012;21(22):4980-4995.

Supplemental data

Table S1. Primers for quantitative allele specific CNV analysis-TyperAssay

SNP_ID	Primer F	Primer R	Extension primer
13255347	ACGTTGGATGCGCTGTCACGCGCTGCCTG	ACGTTGGATGAGAGAAAGGGCGGCTTGGAG	CCGAGACCCCTGCCCGC
4977177	ACGTTGGATGATGCTGTGGAATCTGACTGG	ACGTTGGATGTCAGGTGGAGGCCCGGAGAA	AAGAGTGTGGGGAAGTGA
1062099	ACGTTGGATGGACCCGGTCCTGATTTTAAC	ACGTTGGATGAGACAGCGGGTGGATCCTC	TTTTAACAGTAGACTTGAGAAG
6715929	ACGTTGGATGCCTCATGAAACAAGCTACCC	ACGTTGGATGAAAAATCACAGCTGCGGAAG	ACAAAAACTGGCTTTGC
737497	ACGTTGGATGCCAGCATCCCCTTCCCATAA	ACGTTGGATGCATTCGTTCATGTGACAGTATTCT	TGAGTGCCCGGTCTCCTC
746707	ACGTTGGATGTGCTCGCACGTGGATAGATG	ACGTTGGATGAGGTGACAGTACTGATGAGG	CCTCACACACCAGCAGGCA

Table S2. List of expression primers for qPCR with Roche UPL system and SYBRgreen

primer ID	Primer F	Primer R	purpose	UPL
IRF7	AGCTGTGCTGGCGAGAAG	CATGTGTGTGTGCCAGGAA	11p15.5	Sybr
exp PKP3 new	AGCCTGGAGGAGAAGGCTAAT	AGTGCTGGCTATCCCAAGATACT	11p15.5	Sybr
DIRAS3	TTCTAGGCTGCTTGGTTCGT	TGCACAAGTTCTCCCACACT	target	18
CDH1	GGTCTGTCATGGAAGGTGCT	GATGGCGGCATTGTAGGT	target	5
MAP3K3	GACACTCACGGACCTTAGCC	GTTCAATGCCTCCTGTTCGT	target	70
HPRT F exp	TGACCTTGATTTATTTTGCATACC	CGAGCAAGACGTTCAGTCCT	HKG	73
GAPDH F exp	AGCCACATCGCTCAGACAC	GCCCAATACGACCAAATCC	HKG	60
RIPK2_2	CTTGGTGTAAATTACCTGCACAA	ATGCGCCACTTTGATAAACC	target	63
GAS7_2	CCCCAGAGAAGGTTAGCTGTT	GTGGAAGGATGACCGTCTG	target	7
*RUNX3*_1	TCAGCACCACAAGCCACTT	AATGGGTTCAGTTCCGAGGT	target	71
GRINA	GGAGATCGTGTACGCCTCA	CTCAGGGACAGCTGCTTGTT	8q24.3	38
SIGIRR RT 1 F	CTCAGAGCCATGCCAGGT	CCTCAGCACCTGGTCTTCA	11p15.5	55
PTDSS2_2 RT F	GCCATTTTCCAGACCTCATC	GAGAAACAGCTCGTAGACCACA	11p15.5	21
OPLAH F 1	TCTGTCCTTCAAACTTGTCCAG	CTGCAGTTGGGGACCTTG	8q24.3	63
EXOSC4 F 1	TACATTGAGCAGGGCAACAC	TGCTGAAGGTCGCTGAACTA	8q24.3	76
GPAA1 F 1	GACACTGCTGGCGATTTATG	GGGCCTGTGTGCTTACCA	8q24.3	43
SHARPIN F 1	CTGCCCAGTCCACTCCAG	GGGTGCTACACATCTCACAGC	8q24.3	4
MAF1 F 2	TTCTTTAGCTGCCGTTCCAT	CTCCATGTCCAGCTCGTTG	8q24.3	39
RNH1 F 2	AGCAAAAAGGGGTGTCTCAG	ATGGTGGAGGTGAAGAGTGG	11p15.5	13
HRAS F	GGCATCCCCTACATCGAGA	CTCACGCACCAACGTGTAGA	11p15.5	88
DEAF1 RT F	GGGAGGCTATGAGCGAGTG	TGCTGGTGATCCTTCCAGT	11p15.5	39
SPATC1_1 F	CATCCCAGAGAAGATCATCCA	GAGCCTCTGGCACAGCTT	8q24.3	42
PLEC1_61	CCCTGTGGTGCCTGCTAC	ACACGATCCCGCTCATCT	8q24.3	Sybr
SNAI1	GCTGCAGGACTCTAATCCAGA	ATCTCCGGAGGTGGGATG	TGFbeta	11
Fibronectin1	GGAAAGTGTCCCTATCTCTGATACC	AATGTTGGTGAATCGCAGGT	TGFbeta	33
COL1A1	GGGATTCCCTGGACCTAAAG	GGAACACCTCGCTCTCCA	TGFbeta	67

Table S3. List of primers for BT PCR for massCLEAVE^{TM} assay

Amplicon name	Primer F	Primer R (T7 promoter tag)	CNV	PCR	other
CDHR5_1	aggaagagagGTTTTTTTTTGTTTAAGTAGG	cagtaatacgactcactatagggagaaggctCTTCAAAATAAAAACCCCAAC	11p15.5	54°C	IRF7
PKP3_38	aggaagagagGTGAAGATAGTTGGGTTTGGAG	cagtaatacgactcactatagggagaaggctCTAACCAAACTCAATCTTTAAAAAAC	11p15.5	58°C	CGI 38
PLEC1	aggaagagagGGGTTTGGTTTGGTTAGGGTT	cagtaatacgactcactatagggagaaggctCAACTTACACCCCCATATACCC	8q24.3	60°C	CG161
mir661_5_new_2	aggaagagagAGGTTATGTGTTGGAGGAGGGT	cagtaatacgactcactatagggagaaggctCCACAAATCAACTACACCCTA	8q24.3	56-58°C	CG161
mir661_6	aggaagagagGTTTAAGTTGGTTAAGTATTTAG	cagtaatacgactcactatagggagaaggctCACAAATCAACTACACCCTAAC	8q24.3	56-58°C	CG161
mir661_7	aggaagagagGGTATATGGGGGTGTAAGTTG	cagtaatacgactcactatagggagaaggctCTACCTCTTAAAATACCCCCCACT	8q24.3	56-58°C	CG161

Table S4. List of primers for cloning

Primer ID	Sequence 5´to 3´	Method
61_F3 (HINDIII)	TTGGCCAAGCTTTGGCGGAGGTGGGCGATG	promoter cloning
61_F1 (HINDIII)	TTGGCCAAGCTTCGAAGGCAAGGGCAGTGTTG	promoter cloning
61_R1(BAMHI)	ACCTGAGGGATCCAGGGACCTTGAAGGATGTGTTTA	promoter cloning
CDH1_F1_newpMIR	TGATCACGCGTTCACCCAGCACCTTGCAG	3´UTR cloning
CDH1_R2 HINDIII	ATGATCAAGCTTAATTCAGGAGTGAGAGTTGA	3´UTR cloning
MAP3K3 F pMIR	ATGATCACGCGTGCTCTCACGGCCACACAGCTG	3´UTR cloning
MAP3K3 R pMIR	ATGATCAAGCTTCTGGGTACAGCATAAGAGTGAC	3´UTR cloning
pMIR seq primer F	ACGACGGCCAGTGCCAAGCTA	sequencing
pMIR seq primer R	GATCCTCATAAAGGCCAAGAAG	sequencing

Table S5. CNVs detected with PennCNV.

Chr	Start	End	Start SNP	End SNP	#cases	#controls	CNV cases	CNV controls
chr1	147305744	147478120	rs11579261	rs12409037	22	42	Gain:5 Loss:17	Gain:5 Loss:37
chr1	243703662	243713984	rs6428923	rs12121903	0	12	Gain:0 Loss:0	Gain: 0 Loss:12
chr2	41092148	41099005	rs12617846	rs2373974	41	57	Gain:0 Loss:41	Gain:0 Loss:57*1
chr2	242565979	242593982	rs12987376	rs10189267	11	27	Gain:0 Loss:11	Gain:0 Loss:27
chr3	152997280	153028731	rs17204697	rs1042201	0	14	Gain:0 Loss:0	Gain:0 Loss:14
chr4	161276893	161290832	rs1796466	rs10084880	11	17	Gain:0 Loss:11	Gain:0 Loss:17
chr5	28847546	28877702	rs2548010	rs2652689	14	14	Gain:14 Loss:0	Gain: 14 Loss:0
chr5	97087517	97107276	rs12658613	rs10515261	23	30	Gain:0 Loss:23	Gain:0 Loss:30
chr6	67093085	67105019	rs11758713	rs1634207	39	67	Gain:0 Loss:39	Gain:0 Loss:67
chr6	124496015	124507628	rs11758638	rs2093502	0	10	Gain:0 Loss:0	Gain:11 Loss:0
chr6	168078929	168138806	rs3800533	rs3778667	0	12	Gain:0 Loss:0	Gain:12 Loss:
chr7	141419097	141429438	rs4329195	rs10265585	63	48	Gain:0 Loss:53	Gain:0 Loss:48
chr7	38323070	38323848	rs17171329	rs11765884	0	16	Gain:0 Loss:0	Gain:0 Loss:16
chr7	76270269	76395148	rs38635	rs3912067	0	11	Gain:0 Loss:0	Gain:9 Loss:2
chr8	5583199	5591903	rs2527118	rs1635664	36	54	Gain:0 Loss:36	Gain:0 Loss:54
chr8	145090342	145223898	rs11786896	rs2070688	18	0	Gain:18 Loss:0	Gain:0 Loss:0
chr8	137898044	137913669	rs2582447	rs2681674	0	22	Gain:0 Loss:0	Gain:0 Loss:22
chr10	47098898	47110350	rs4926057	rs12775238	27	0	Gain:25 Loss:2	Gain:0 Loss:0
chr10	135125348	135182921	rs2252728	rs10776686	13	19	Gain:12 Loss:21	Gain:19 Loss:0
chr10	67748487	67785209	rs4297361	rs2893986	0	12	Gain:0 Loss:0	Gain:0 Loss:12
chr11	548884	609789	rs2061586	rs2246614	10	0	Gain:10 Loss:0	Gain:0 Loss:0
chr11	55139733	55179162	rs573732	rs11230571	57	0	Gain:0 Loss:57	Gain:0 Loss:0
chr11	55127597	55139733	rs2456022	rs573732	0	104	Gain:0 Loss:0	Gain:0 Loss:104
chr12	7892179	7936264	rs16916683	rs11502980	14	10	Gain:13 Loss:1	Gain:9 Loss:1
chr12	31271893	31276546	rs11051344	rs12831069	23	31	Gain:23 Loss:0	Gain:31 Loss:0
chr12	19360345	19431361	rs10743315	rs2961370	0	11	Gain:0 Loss:0	Gain:11 Loss:0
chr14	21831090	21832903	rs10483271	rs17198328	11	0	Gain:0 Loss:11	Gain:0 Loss:0

chr14	21834952	21838610	rs12588739	rs3811260	0	13	Gain:0 Loss:0	Gain:0 Loss:13
chr18	64898548	64905367	rs11876036	rs13381870	21	24	Gain:1 Loss:20	Gain:0 Loss:24
chr19	1189899	1201109	rs2301759	rs3746106	10	0	Gain:10 Loss:0	Gain:0 Loss:0
chr19	20422200	20473895	rs10408291	rs2021399	22	38	Gain:0 Loss:22	Gain:0 Loss:38
chr19	48197824	48205499	rs11881408	rs11668932	17	22	Gain:4 Loss:13	Gain:4 Loss:18
chr19	59423491	59435029	rs17207328	rs17239607	0	20	Gain:0 Loss:0	Gain:20 Loss:0
chr20	52081230	52088118	rs1557853	rs290469	21	29	Gain:0 Loss:21	Gain:0 Loss:29
chr22	21392612	21401228	rs6003245	rs12484427	14	27	Gain:14 Loss:0	Gain:27 Loss:0
chr22	24083777	24165514	rs5996921	rs1207587	17	26	Gain:10 Loss:7	Gain:15 Loss:9
chr22	17257787	17355587	rs2543958	rs2518805	0	12	Gain:0 Loss:0	Gain:12 Loss:0
chr22	21457585	21554058	rs11912861	rs2282667	0	17	Gain:0 Loss:0	Gain:17 Loss:0

Table S6. CNVs detected with QuantiSNP.

Chr	Start	End	StartSNP	EndSNP	CNV	#SNP	#cases	#controls	CNV cases	CNV controls
chr1	147305744	147478120	rs11579261	rs12409037	1	9	35	37	Gain:3 Loss:32	Gain:9 Loss:28
chr1	195089653	195163711	rs16840607	rs4915318	1	10	11	8	Gain:2 Loss:9	Gain:3 Loss:5
chr2	41092148	41099005	rs12617846	rs2373974	1	9	56	59	Gain:0 Loss:56	Gain:0 Loss:59
chr2	89743465	89877778	rs17091423	rs13003799	1	9	21	18	Gain:2 Loss:19	Gain:0 Loss:18
chr2	110228954	110339819	rs17842653	rs13386516	3	12	5	12	Gain:3 Loss:2	Gain:6 Loss:6
chr3	152997280	153028731	rs17204697	rs1042201	1	11	11	16	Gain:0 Loss:11	Gain:0 Loss:16
chr4	161276893	161290832	rs1796466	rs10084880	1	19	22	22	Gain:0 Loss:22	Gain:1 Loss:21
chr4	162172873	162207355	rs4690999	rs1523553	3	9	5	10	Gain:5 Loss:0	Gain:10 Loss:0
chr5	8756615	8800106	rs10073742	rs10434659	1	8	9	12	Gain:0 Loss:9	Gain:0 Loss:12
chr5	97074222	97107276	rs2914928	rs10515261	1	13	34	36	Gain:0 Loss:34	Gain:0 Loss:36
chr5	28842013	28912873	rs2548005	rs457561	3	17	10	6	Gain:10 Loss:0	Gain:6 Loss:0
chr6	67093085	67105019	rs11758713	rs1634207	1	7	55	84	Gain:0 Loss:55	Gain:0 Loss:84
chr6	168216329	168240295	rs2171983	rs10046330	3	10	10	0	Gain:10 Loss:0	Gain:0 Loss:0
chr7	8810973	8826141	rs12702782	rs10486260	1	10	11	10	Gain:0 Loss:11	Gain:0 Loss:10
chr7	38285864	38302045	rs2240826	rs2191311	1	13	12	11	Gain:0 Loss:12	Gain:0 Loss:11
chr7	61075979	61752449	rs13247259	rs238258	3	9	20	0	Gain:20 Loss:0	Gain:0 Loss:0
chr7	141420759	141429438	rs4281037	rs10265585	1	7	66	60	Gain:0 Loss:66	Gain:1 Loss:59
chr8	5583199	5591903	rs2527118	rs1635664	1	6	56	51	Gain:0 Loss:56	Gain:0 Loss:51
chr8	137898044	137913669	rs2582447	rs2681674	1	19	15	29	Gain:0 Loss:15	Gain:1 Loss:28
chr8	145195417	145247517	rs4977177	rs13264654	3	7	27	0	Gain:27 Loss:0	Gain:0 Loss:0
chr10	47013328	47110350	rs11259779	rs12775238	3	17	47	38	Gain:43 Loss:4	Gain:34 Loss:4
chr10	67748487	67785209	rs4297361	rs2893986	1	13	14	12	Gain:0 Loss:14	Gain:0 Loss:12
chr11	548884	609789	rs2061586	rs2246614	3	7	13	0	Gain:11 Loss:1	Gain:0 Loss:0
chr11	50599126	51077585	rs1592593	rs4323853	3	3	42	14	Gain:42 Loss:0	Gain:14 Loss:0
chr11	55127597	55139733	rs2456022	rs573732	1	2	0	117	Gain:0 Loss:0	Gain:0 Loss:117
chr11	55139733	55193702	rs573732	rs17498926	1	8	80	2	Gain:0 Loss:80	Gain:0 Loss:2
chr11	81174591	81194909	rs4409862	rs12293984	1	13	2	15	Gain:0 Loss:2	Gain:1 Loss:14

chr12	7899399	7990569	rs2889504	rs1473164	3	12	21	15	Gain:16 Loss:5	Gain:12 Loss:3
chr12	19360345	19442103	rs10743315	rs2565666	3	17	4	13	Gain:4 Loss:0	Gain:13 Loss:0
chr12	31257563	31276546	rs10771812	rs12831069	3	10	35	39	Gain:35 Loss:0	Gain:39 Loss:0
chr12	31915523	31941353	rs1419311	rs1150971	3	14	10	11	Gain:10 Loss:0	Gain:11 Loss:0
chr12	62269256	62351565	rs11175055	rs12231958	3	6	14	9	Gain:11 Loss:3	Gain:5 Loss:4
chr14	21767514	21792564	rs10483269	rs10148895	1	6	10	0	Gain:0 Loss:10	Gain:0 Loss:0
chr14	21834952	21838610	rs12588739	rs3811260	1	3	10	24	Gain:0 Loss:10	Gain:0 Loss:24
chr14	21898729	21914810	rs2331662	rs4982619	1	3	0	15	Gain:0 Loss:0	Gain:1 Loss:14
chr18	64898548	64905367	rs11876036	rs13381870	1	13	31	30	Gain:0 Loss:31	Gain:1 Loss:29
chr19	20422200	20473895	rs10408291	rs2021399	1	9	25	33	Gain:0 Loss:25	Gain:0 Loss:33
chr19	48093776	48160500	rs10405494	rs17279415	1	8	0	28	Gain:0 Loss:0	Gain:3 Loss:25
chr20	52081230	52088118	rs1557853	rs290469	1	7	37	38	Gain:35 Loss:2	Gain:0 Loss:38
chr22	17257787	17355587	rs2543958	rs2518805	3	21	5	12	Gain:2 Loss:2	Gain:12 Loss:0
chr22	24017514	24215704	rs5752118	rs1930966	3	34	27	30	Gain:19 Loss:8	Gain:20 Loss:10

Table S7. Overlapped CNVs between QuantiSNP and PennCNV

QuantiSNP					PennCNV				
Chr	**Start**	**End**	**#pval**	**pval_comb**	**Chr**	**Start**	**End**	**#pval**	***p*val_comb**
chr1	1.47E+08	1.47E+08	9	3.2E-07	chr1	1.47E+08	1.47E+08	9	1.0E-12
chr2	41092148	41099005	9	9.9E-01	chr2	41092148	41099005	9	2.5E-05
chr3	1.53E+08	1.53E+08	11	8.0E-02	chr3	1.53E+08	1.53E+08	11	2.9E-03
chr4	1.61E+08	1.61E+08	19	1.0E+00	chr4	1.61E+08	1.61E+08	19	2.7E-03
chr5	97074222	97107276	13	<1e-18	chr5	97087517	97107276	13	2.8E-02
chr5	28842013	28912873	17	<1e-18	chr5	28847546	28877702	17	1.0E+00
chr6	67093085	67105019	7	3.7E-10	chr6	67093085	67105019	7	8.9E-13
chr7	1.41E+08	1.41E+08	7	9.2E-01	chr7	1.41E+08	1.41E+08	7	9.8E-03
chr8	5583199	5591903	7	9.7E-01	chr8	5583199	5591903	7	4.3E-06
chr8	1.38E+08	1.38E+08	19	2.2E-16	chr8	1.38E+08	1.38E+08	19	<1e-18
chr8	1.45E+08	1.45E+08	6	<1e-18	chr8	1.45E+08	1.45E+08	12	<1e-18
chr10	47013328	47110350	17	2.3E-01	chr10	47098898	47110350	17	5.4E-01
chr10	67748487	67785209	17	1.0E+00	chr10	67748487	67785209	17	3.5E-03
chr11	548884	609789	10	<1e-18	chr11	548884	609789	10	4.4E-16
chr11	55127597	55139733	1	1.9E-03	chr11	55127597	55139733	2	4.1E-10
chr11	55139733	55193702	7	<1e-18	chr11	55139733	55179162	7	<1e-18
chr12	7899399	7990569	21	1.1E-01	chr12	7892179	7936264	21	9.1E-01
chr12	19360345	19442103	17	<1e-18	chr12	19360345	19431361	17	<1e-18
chr12	31257563	31276546	10	8.5E-01	chr12	31271893	31276546	10	1.1E-02
chr14	21834952	21838610	3	7.7E-05	chr14	21834952	21838610	3	1.0E+00
chr18	64898548	64905367	13	1.0E+00	chr18	64898548	64905367	13	9.2E-01
chr19	20422200	20473895	9	4.1E-02	chr19	20422200	20473895	9	1.7E-08
chr20	52081230	52088118	7	1.0E+00	chr20	52081230	52088118	7	1.0E+00
chr22	17257787	17355587	21	<1e-18	chr22	17257787	17355587	21	<1e-18
chr22	24017514	24215704	34	<1e-18	chr22	24083777	24165514	34	1.7E-08

[a]Pval comb : P value combined was obtained according to reference [156]

Table S8. Significant CNVs associated with lung cancer risk (QuantiSNP)

Cytoband	CNV coordinates (hg18)[a]	CNV type	QuantiSNP n cases	QuantiSNP n controls	QuantiSNP *p* value[b]
1q21.1	chr1:147305744-147478120	loss	Gain: 3 Loss: 32	Gain: 9 Loss: 28	$3.2x10^{-07}$
8q24.23	chr8:137898044-137913669	loss	Gain:0 Loss: 15	Gain: 1 Loss:28	$2.2x10^{-16}$
8q24.3	chr8:145195417-145247517	gain	Gain: 27 Loss: 0	Gain: 0 Loss: 0	$<10^{-18}$
6q12	chr6:67093085- 67105019	loss	Gain: 0 Loss:55	Gain: 0 Loss: 84	$3.7x10^{-10}$
11p15.5	chr11:548884-609789	gain	Gain: 11 Loss:1	Gain: 0 Loss: 0	$<10^{-18}$
11q11	chr11:55139733-55193702	loss	Gain: 0 Loss: 80	Gain: 0 Loss: 2	$<10^{-18}$
11q11	chr11:55127597-55139733	loss	Gain:0 Loss: 0	Gain: 0 Loss: 117	$1.93x10^{-3}$
12p12.3	chr12:19360345-19442103	gain	Gain: 4 Loss: 0	Gain: 13 Loss: 0	$<10^{-18}$
19p12	chr19:20422200-20473895	loss	Gain: 0 Loss:25	Gain: 0 Loss:33	$4.08x10^{-2}$
22q11.21	chr22:17257787-17355587	gain	Gain: 2 Loss: 3	Gain: 12 Loss: 0	$<10^{-18}$

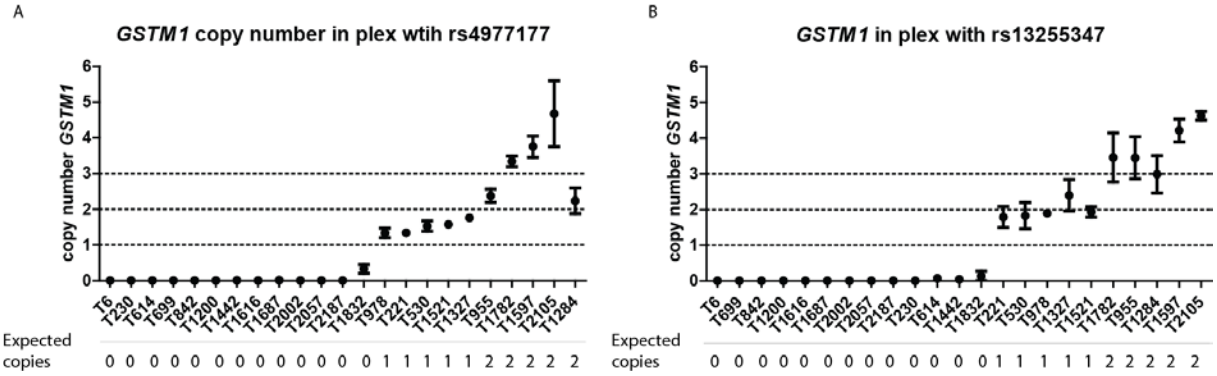


Figure S1. *GSTM1* copy numbers in 3 plex assay for 8q24.3. A. Copy number determination in 3 plex with 2N control and rs4977177. B. Copy number determination in 3 plex with 2N control and rs13255347.

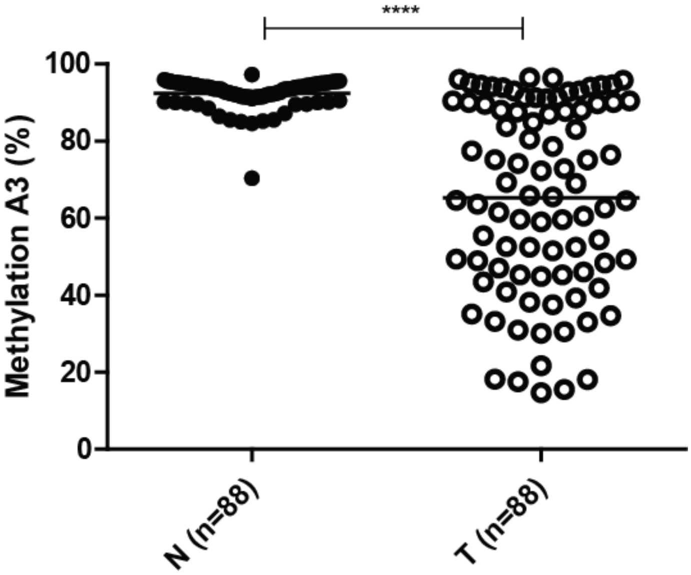


Figure S2. Validation of methylation at the putative promoter of miR-661. Average % methylation of amplicon A3 was carried out in normal lung (N) and lung tumor (T) of patients from set 2 and 3 (n=88). Statistical analysis was carried out with Wilcoxon matched pairs signed rank test with a significance cut off at *p*<0.05.

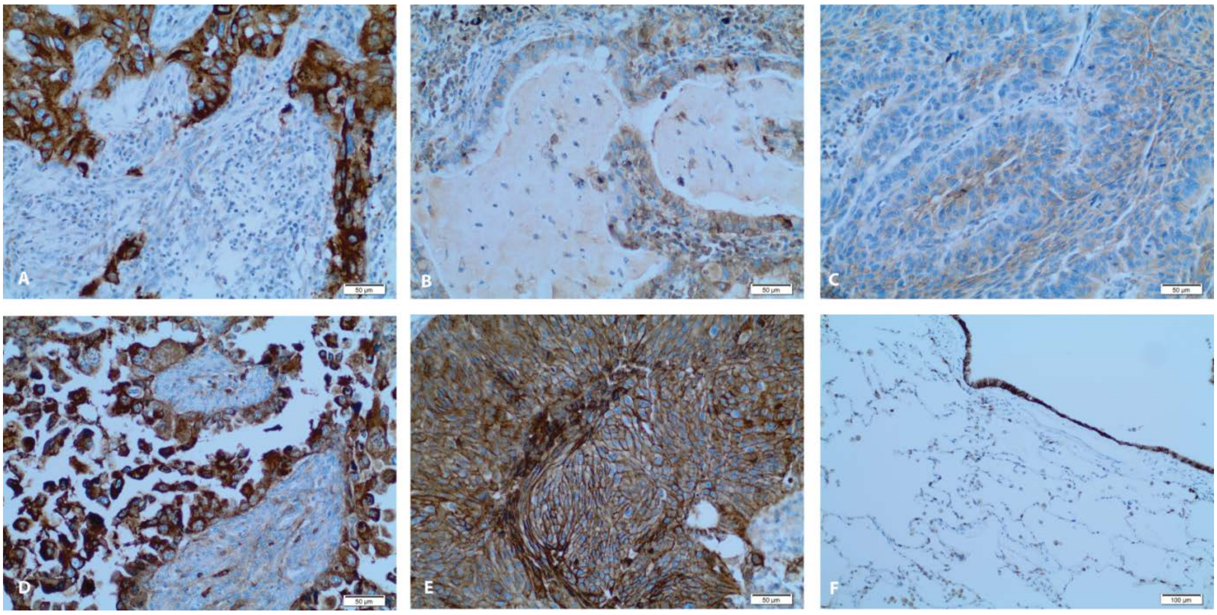


Figure S3. E-cadherin immunohistochemical staining of primary AdC, SCC and bronchial and aveolar cells. A. Example of the invasion front in SCC **B-C**. NSCLC with low expression of E-cadherin. **D**. AdC case with high expression of E-cadherin. E. SCC with high expression of E-cadherin. **F**. Expression of E-cadherin in alveolar and bronchus (normal lung tissue). IHC was performed by NCT and evaluation was made by pathologist Arne Warth, Institute of Pathology, University Hospital Heidelberg, Germany .

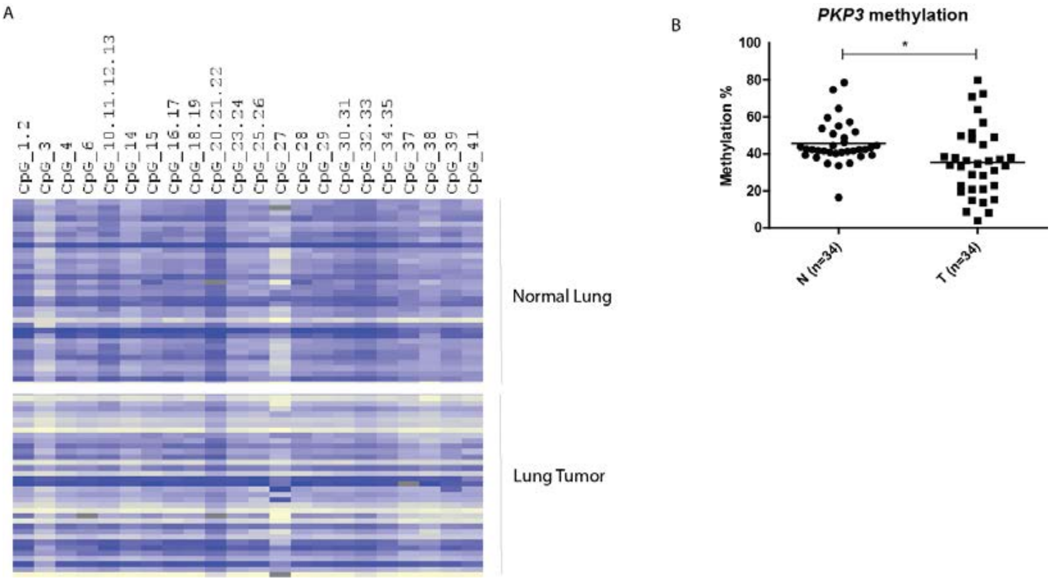


Figure S4. Validation of methylation at the TSS of PKP3 in sample set 3. A. Methylation values are shown in the heatmap ranging from yellow (0%) to blue (100%) for each CpG or CpG unit (columns) for 34 tumor-normal pairs (rows). **B**. Average % methylation over the amplicon showed in A in N (normal lung) and T (lung tumor) Statistical analysis was carried out with Wilcoxon matched pairs signed rank test with a significance cut off at p<0.05.

Bibliography

Publications:

Jacobsson H, Hotz-Wagenblatt A, Beckmann L, Sauter W, Rosenberger A, Muley T, Meister M, Illig T, Senturk N, Warth A, Dienemann H, Chang-Claude J, Wichmann H-E, Bickeböller H, Plass C, Risch A, Post-GWA functional characterization of copy number gain identifies methylation dependent upregulation of miR-661 in NSCLC **(submitted to *Journal of National Cancer Institute*, July 2013).**

Scherf DB, Sarkisyan N, Jacobsson H, Claus R, Bermejo J L, Peil B, Gu L, Muley T, Meister M, Dienemann H, Plass C, Risch A (2012) Epigenetic screen identifies genotype- specific promoter DNA methylation and oncogenic potential of CHRNB4**. Oncogene. 2013 Jul 11;32(28):3329-38. doi: 10.1038/onc.2012.344.**

Risch A, Sarkisyan N, Scherf DB, Jacobsson H, Hagmann W, Plass C, Epigenetic epidemiology in cancer (2012) K.B. Michels (ed.), *Epigenetic Epidemiology*, **doi 10.1007/978-94-007-2495-2_13, Springer Science+Business Media B.V. 2012.**

Abstracts:

Jacobsson H, Hotz-Wagenblatt A, Beckmann L, Sauter W, Muley T, Meister M, Dienemann H, Wichmann E, Bickeböller H, Plass C, Risch A, Epigenetic regulation of genes in CNV risk loci for lung cancer.
Poster presentation at the Wellcome Trust conference "Epigenomics in common diseases", ***Baltimore, USA, October 2012.***

Jacobsson H, Hotz-Wagenblatt A, Beckmann L, Sauter W, Muley T, Meister M, Dienemann H, Wichmann E, Bickeböller H, Plass C, Risch A, Epigenetic regulation of potential CNV risk loci in lung cancer.
*Invited speaker at the 10th annual retreat of the international graduate program of Molecular Biology and Medicine of the lung****, Giessen, Germany August 2012.***

Jacobsson H, Claus R**,** Hotz-Wagenblatt a, Beckmann L, Sauter W, Muley T, Meister M, Dienemann H, Wichmann E, Bickeböller H, Plass C, Risch A, The impact of copy number variations and aberrant methylation on lung cancer risk
PhD poster presentation at DKFZ, ***Heidelberg, Germany December 2011.***

Jacobsson H, Claus R**,** Hotz-Wagenblatt a, Beckmann L, Sauter W, Muley T, Meister M, Dienemann H, Wichmann E, Bickeböller H, Plass C, Risch A, The impact of CNVs and epigenetic patterns on lung cancer risk
KICancer retreat, Karolinska institutet, ***Stockholm Sweden, September 2011.***

Acknowledgement

This work would not have been possible without several people. I am very happy to thank the following people…

- Christoph Plass for giving me the opportunity to work in the Epigenomics and Cancer Risk factor division at the DKFZ and for constructive discussions during our Friday meetings.

- Angela Risch for supervision, providing the project and for the input, good discussions and ideas during these years.

- David Scherf for a good friendship and for all the fun discussions and great support during good and bad times. Every division should have a David!

- The lung cancer group; Wolfgang Hagmann, Christian Faltus, Narek Sarkysian, Marion Bähr, Svitlana Melnik and Marina Laplana. A special thanks to Svitlana Melnik for great discussions and suggestions during the time we worked together. Thank you, Marina Laplana, for bringing the Catalan flavor to the group and for the critical input during my thesis writing. Thanks also to Marion Bähr for technical support. Additional thanks to Valentin Frank for proofreading the reference list!

- The Human Genetics department at the Heidelberg University. Especially to Anna Jauch and Brigitte Scholl who introduced the FISH method to me and Katrin Hinderhöfer and Bianca Maas for performing MLPA.

- The TAC committee members Odilia Popanda, Jenny Chang-Claude and Anna Jauch for good discussions and suggestions during the meetings.

- Agnes Hotz-Wagenblatt and Lars Beckmann who performed the initial analysis of the CNV detection.

-The Thoraxklinik, especially Michael Meister and Thomas Muley for a good collaboration.

-The division of Epigenomics and Cancer Risk Factors for simply creating a good and inspiring working environment. Special thanks go to

- Rainer Claus for helping me, supporting and struggling with me during my first year with the CNV analysis method.

- Anders Lindroth for helping me maintaining my mother tongue (!) and for being a great support scientifically and generally.

- Yoon Jung Park for being such a motivating and supporting person both with work and daily life related issues.

- Dieter Weichenhan for supporting the northern European population and for giving great experimental suggestions and honest input.

- Chris Oakes for showing interest and giving input on the *MAP3K3* story and for fun times outside the lab.

- Christoph Weigel for nice discussions about the TGFbeta story.

- Khelifa Arab for always showing a big interest in my project and providing good ideas.

- Conny Bär for introducing me to the fascinating world of miRNAs.

-Marta Faryna and Joseph Huang for keeping up with me in the “glass box” and for sharing the chewing gum supply!

-My lunch-friends Miriam Sonnet, Tania Witte and Céline Dutruel with whom I had great times with at work and outside. Thanks Miriam for helping me writing the German summary!

-Lea Geiselhart who has been a great help the last months during thesis and application writing.

-My dear friends in Heidelberg. Without you, Aoife Ward, those years in Germany would have been very different and probably not that crazy fun! Thank you so much for keeping my motivation up in both work and daily life. Thank you Jennifer Lee for a great friendship with all what that means. Also, without Jen, I would probably not have run any half-marathons!

- My Johannes. Thank you for your endless support, for keeping my mood up and for being there for me during this “rollercoaster” period. Baby bala bala, you deserve a gold medal for standing Koopa during these last months!

-Karin och Frida. Jag är otroligt lyckligt lottad som har så fantastiska vänner. Tack för ert mentala stöd under denna period och speciellt för att ni alltid trott pa mig. Det är gott det!

-Min familj: Olof, Sofia med familj, Ulla och pappa för all kärlek och stöd. Ett stort och speciellt tack till dig, pappa, för din förmåga att alltid finnas till hands, för ditt intresse för min forskning och för intressanta och långa forskningsinspirerande diskussioner.