Genomic characterization of non-small cell lung cancer - clinical and molecular implications

Karlsson, Anna F

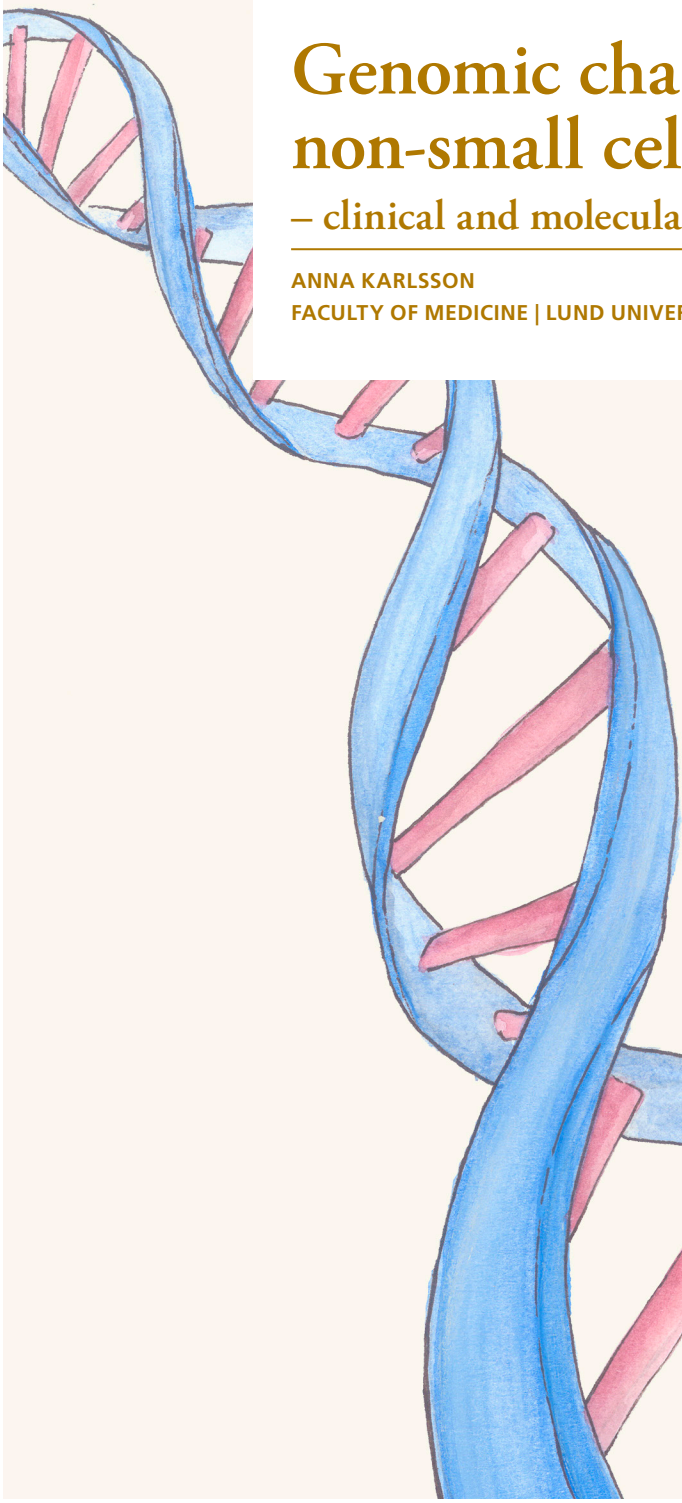
2018

Document Version:
Publisher's PDF, also known as Version of record

Genomic characterization of non-small cell lung cancer

– clinical and molecular implications

ANNA KARLSSON
FACULTY OF MEDICINE | LUND UNIVERSITY



Genomic characterization of non-small cell lung cancer – clinical and molecular implications

Genomic characterization of non-small cell lung cancer – clinical and molecular implications

Anna Karlsson

LUND
UNIVERSITY

DOCTORAL DISSERTATION
by due permission of the Faculty of Medicine, Lund University, Sweden.
To be publicly defended at the lecture hall in the Radiotherapy building,
Klinikgatan 5, Skåne University Hospital, Lund, Sweden.
October 19, 2018 at 9.00 am.

Faculty opponent
Odd-Terje Brustugun, M.D, Ph.D.
Department of Cancer Genetics
Institute for Cancer Research
The Norwegian Radium Hospital
Oslo, Norway.

Organization	Document name
LUND UNIVERSITY	DOCTORAL DISSERTATION
	Date of issue October 19, 2018
Author(s) Anna Karlsson	Sponsoring organization

Title and subtitle

Genomic characterization of non-small cell lung cancer – clinical and molecular implications

Abstract

Lung cancer accounts for 1.6 million cancers annually and 1.3 million deaths per year make the disease the deadliest type of cancer. Smoking is the most prominent cause. High mortality rates are partly due to late diagnosis which limits the therapy options. Improved understanding of molecular and biological mechanisms of tumor development could guide current therapy selection or pave way for novel therapies. The aim of this thesis is to characterize non-small cell lung cancer (NSCLC) genomes, stratifying tumors and patient groups with the intent of improved therapeutic options. A framework for treatment predictive mutation testing was established and, in parallel, the NanoString technology was evaluated for fusion gene detection. In an expansion of the fusion gene detection method, we included the possibility for simultaneous prediction of NSCLC histology, i.e. a multicomponent assay, applicable to tissue amounts used in standard clinical diagnostics. As accurate distinction of the histological subtypes is crucial for clinical management of lung cancer, the World Health Organization (WHO) administer and continuously update guidelines for histological classification. The mutational and transcriptional findings of this thesis work support their 2015 revision of histological subgroups. In addition, comprehensive global methylation analysis of lung cancer was performed, resulting in the detection of five epigenetically distinct groups of lung tumors associated with histology, gene expression, copy number variation and survival. In summary, this thesis has focused on genomic characterization of NSCLC, contributing to molecular findings and clinical implications.

Key words: Non-small cell lung cancer, histological classification, methylation, gene expression, gene fusion.

Classification system and/or index terms (if any)

Supplementary bibliographical information		Language: English
ISSN and key title: 1652-8220		ISBN: 978-91-7619-690-8
Recipient's notes	Number of pages 170	Price
	Security classification	

Signature  Date 2018-09-13

Genomic characterization of non-small cell lung cancer – clinical and molecular implications

Anna Karlsson
Faculty of Medicine
Division of Oncology and Pathology
Department of Clinical Sciences, Lund

Supervisor: Associate Professor, Johan Staaf, Ph.D
Faculty of Medicine
Division of Oncology and Pathology
Department of Clinical Sciences, Lund

Co-supervisor: Associate Professor, Maria Planck, M.D, Ph.D
Faculty of Medicine
Division of Oncology and Pathology
Department of Clinical Sciences, Lund
Department of Respiratory Medicine and Allergology
Skåne University Hospital

Co-supervisor: Associate Professor, Markus Ringnér, Ph.D
Faculty of Medicine
Division of Molecular Cell Biology
Department of Biology, Lund

Coverphoto by Lina Falk and Annette Salomonsson

Faculty of Medicine
Department of Oncology and Pathology

ISBN 978-91-7619-690-8
ISSN 1652-8220

Printed in Sweden by Media-Tryck, Lund University
Lund 2018

To my family ♥

”Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less.”

-Marie Curie

Content

Populärvetenskaplig sammanfattning

Globalt drabbas årligen 1,6 miljoner människor av lungcancer och 1,3 miljoner dör i sin sjukdom, vilket gör lungcancer till den dödligaste formen av cancer. Sjukdomsorsaken tillskrivs till största del rökning (75–80% av fallen) men även andra miljö- och genetiska faktorer kan orsaka sjukdomen. Lungcancer delas in i två huvudgrupper baserat på tumörcellernas egenskaper: småcellig lungcancer (SCLC) och icke-småcellig lungcancer (NSCLC). De flesta tumörer är av typen NSCLC (75%) och denna grupp delas in i ytterligare undergrupper baserat på utseende och karaktäristika, så kallade histologier. Tidig upptäckt, dvs när tumören är liten till storlek och ej spridit sig, är den viktigaste faktorn för överlevnad. Då avlägsnas tumören med kirurgi och patienten behandlas därefter eventuellt med kemoterapi (cytostatika). Tyvärr upptäcks alltför många tumörer i sent stadium när tumören är för stor att avlägsna med kirurgi och/eller spridit sig (metastaserat) till andra delar av kroppen (avancerad sjukdom). Behandlingen ges då i så kallat palliativt syfte, dvs för symtomlindring, tumörkontroll och i bästa fall förlängning av överlevnaden. Det enda alternativet i dessa fall var tidigare behandling med exempelvis kemoterapi och strålbehandling för smärtlindring av tumörbörda och metastasering. Vid avancerad sjukdom är det idag även viktigt att utreda huruvida tumören besitter vissa genetiska förändringar, mutationer, i särskilda gener som visat sig vara tumördrivande. En specifik sådan gen, *EGFR*, är muterad i ca 10-15% av alla lungtumörer. Dessutom inträffar relativt ofta så kallade genfusioner i lungcancer, där en del av en gen smälter ihop med en del av en annan gen. Produkten av denna nya gen (fusionsgen) kan vara starkt tumördrivande. Hos ca 1-5% av lungcancerpatienterna detekteras fusionsgener som involverar *ALK*, *RET* eller *ROS1*. Både *EGFR* mutationer och fusionsgener är förändringar i tumören som denna är starkt beroende av för att kunna existera. Idag finns det riktade läkemedel utvecklade mot dessa, för tumören, livsuppehållande förändringar. För patienter med *EGFR* mutation eller genfusion innebär möjligheten att få målriktad behandling oftast högre livskvalité samt en längre sjukdomsfri överlevnad även om resistensutveckling ofta sker över tid. Metoder för att detektera dessa specifika förändringar kallas ofta behandlingsprediktiv mutationstestning.

Detta avhandlingsarbete består av fem delarbeten. Det samlade målet för avhandlingen är att studera mönster av genuttryck och genförändringar i tumörer

från patienter diagnostiserade med NSCLC och att relatera dessa till patientöverlevnad och till nuvarande eller framtida behandlingsmöjligheter.

Studie I i denna avhandling beskriver ett ramverk uppsatt på avdelningen för Onkologi och Patologi vid Lunds universitet där behandlingsprediktiva mutationer hos lungcancerpatienter diagnostiserade i Södra Sjukvårdsregionen analyserades. Studien beskriver den kliniska validering av den nya tekniken (s.k. targeted NGS) som utprövades parallellt med den vanliga rutinmässiga diagnostiken vid tidpunkten. Vid valideringstidens slut, när den nya tekniken ansågs lämplig för behandlingsprediktiv mutationstestning av lungcancerpatienter, infördes den i klinisk rutin. Med den nya tekniken kartlades även parallellt förändringar i andra gener som i dagsläget inte används som kliniskt beslutsunderlag av behandlande läkare. Studie I rapporterar frekvenser av samtliga förändringar som hittades, sammankopplat med histologiska undergrupper under det första året av klinisk drift vilket gav en mer komplett bild av den tumörgenetiska profilen av NSCLC i sjukvårdsregionen. Upptäckta mutationer rapporterades till avdelningen för klinisk patologi för vidare rapportering, bedömning och kliniskt svar till behandlande läkare. I Studie I utvecklades även en ny teknik för parallell analys av genfusioner baserat på analys av tumör RNA via NanoString teknik. NanoString är en snabb, enkel och robust teknik utvecklad för att med hög känslighet kunna analysera små mängder av nedbrutet RNA, vilket typiskt är fallet med RNA utvunnit ur fixerat kliniskt rutinmaterial. Den NanoString baserade fusionsgensdetektionen rapporterades inte till kliniken som en klinisk rutin, men samtliga fusioner som detekterades med rutindiagnostiken kunde även detekteras med NanoString.

I Studie II gjordes en uppgradering av NanoString metoden från Studie I genom att inkludera fler gener (utöver fusionsgener), till exempel gener som används som markörer för att fastställa de histologiska undergrupperna av NSCLC. Målet med Studie II var att utveckla en formel som kan förutsäga vilken histologi en tumör besitter baserat på RNA uttrycket av gener associerade med lungcancerhistologi, en så kallad prediktor. Den utvecklade RNA prediktorn för histologi visade sig prestera mycket bra, dvs var i hög grad överensstämmande med den histologi som bestämts via patologiska undersökningar. Prediktorn kunde med hög säkerhet förutsäga histologi i patientkohorter där genuttrycket kartlagts med en annan teknik än den som prediktorn själv utvecklats på, dvs den var plattformsoberoende. Med NanoString tekniken är det alltså möjligt att via en enda analys samtidigt kunna fastställa både NSCLC histologi samt fusionsgens status: två kliniskt mycket relevanta frågeställningar i lungcancer.

Två mindre vanligt förekommande NSCLC histologier är storcelliga tumörer (LCC) samt storcelliga neuroendokrina tumörer (LCNEC). Trots att relativt få patienter utvecklar LCC och LCNEC tumörer så representerar dessa två tumörtyper en viktig differentialdiagnos. I Studie III och IV karakteriseras dessa

tumörtyper på genomisk nivå dels ingående som en egen grupp (Studie IV) och dels i en kontext av andra NSCLC tumörer (Studie III). Studie IV påvisade att tumörer av LCNEC histologi är starkt skilda från LCC tumörer avseende genförändringar. LCC tumörerna visade sig även vara en heterogen grupp av tumörer vilka innehöll tumörer av andra subtyper när proteinmarkörer som idag används kliniskt för histologibestämning testades. Vid tidpunkten för Studie IV ändrade världshälsoorganisationen WHO sina riktlinjer för hur LCC tumörer ska klassas, där de dels separerar LCNEC från LCC gruppen vilket även vi noterade i studie IV. Dessutom infördes proteinmarkör baserad testning för att öka känsligheten i histologibestämningen. Det senare medför att många tidigare LCC tumörer nu blir omklassade till andra histologiska subtyper, vilket kraftigt reducerar den kvarvarande LCC gruppen. I Studie IV kunde vi införa dessa nya riktlinjer och rapportera genförändringar kontra de nya riktlinjerna för de i dag använda histologiklasserna. Resultatet av de nya riktlinjerna var även i enlighet med våra egna resultat i Studie III där genuttrycksmönster (RNA) av LCC/LCNEC tumörer analyserades i en kontext av alla histologiska undergrupper av lungcancer. LCNEC tumörer grupperades med SCLC tumörer baserat på genuttryck som en separat grupp med tydligt uttryck av neuroendokrina gener, medan LCC tumörer som uttryckte proteinmarkörer karakteristiska för andra NSCLC undergrupper grupperades med respektive tumörer. För kvarvarande LCC tumörer enligt de nya riktlinjerna (vilka inte ska uttrycka några proteinmarkörer karakteristiska för andra undergrupper) såg vi att de grupperade tillsammans som en separat molekylär undergrupp.

Metylering är en biologisk process som celler i kroppen använder sig av framförallt under fosterutvecklingen för att styra genuttryck under korta perioder. Via specifik metylering av en del av genen kan uttrycket exempelvis stängas ner. Denna biologiska process är även vanlig i cancer där effekten av förändringar i metyleringsstatus dessvärre kan leda till olika fördelar för tumören. I Studie V undersöktes metyleringsmönster hos NSCLC tumörer av olika histologiska undergrupper, och det visade sig att NSCLC tumörer kan delas in i kategorier baserat på skillnader i metyleringsmönster (metyleringsdrivna undergrupper) vilka överlappade med histologiska undergrupper på en högre nivå. De metyleringsdrivna undergrupperna kunde även sammankopplas med överlevnad inom en specifik histologisk undergupp, adenocarcinom. Patienter med tumörer som uppvisade ett visst metyleringsmönster hade ökad överlevnad jämfört med patienter tillhörande en grupp av skilda metyleringsmönster.

Sammantaget visar detta avhandlingsarbete på vikten av genomisk karakterisering av lungcancer, vilken redan har klinisk betydelse idag men som sannolikt kommer öka ännu mer i framtiden. Valmöjligheterna avseende behandling av lungcancer ökar ständigt i takt med ökad kunskap om lungcancerbiologin. Att korrekt undergruppera NSCLC tumörer kommer sannolikt innebära ökade

behandlingsmöjligheter för patienten i framtiden. Detta avhandlingsarbete visar olika exempel på hur kartläggning av olika tumörbiologiska processer och ökad genomisk förståelse av NSCLC kan ha en framtida klinisk betydelse för lungcancerpatienter.

Abbreviations

AC	Adenocarcinoma
AEC	Alveolar Epithelial Cell
AMP	Anchored Multiplex PCR
AUC	Area Under the Curve
CNS	Central Nerve System
CT	Computed Tomography
ctDNA	Circulating tumor DNA
ddPCR	Digital droplet PCR
DNA	Deoxyribonucleic Acid
EBUS	Endobronchial Ultrasound
FDA	Food and Drug Administration
FFPE	Formalin Fixed Paraffin Embedded
FISH	Fluorescence in-situ Hybridization
GEP	Gene Expression Phenotype
GLAD	Gain and Loss Analysis of DNA
GO	Gene Ontology
H&E	Hematoxylin and Eosin
IHC	Immunohistochemistry
IVT	In Vitro Transcription
LCC	Large Cell Carcinoma
LCNEC	Large Cell Neuroendocrine Carcinoma
LOH	Loss of Heterozygosity
miRNA	microRNA
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
NOS	Not Otherwise Specified
NSCLC	Non-Small Cell Lung Cancer
OS	Overall Survival
PCA	Principle Component Analysis
PCR	Polymerase Chain Reaction
PFS	Progression Free Survival
PET	Positron Emission Tomography
RIN	Ribosomal Integrity Number
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristics

SAM	Significance of Microarray Analysis
SBRT	Stereotactic Body Radiation Therapy
SBS	Sequencing By Synthesis
SCLC	Small Cell Lung Cancer
SqCC	Squamous Cell Carcinoma
SSP	Single Sample Predictor
TAT	Turnaround Time
TKI	Tyrosine Kinase Inhibitor
TNM	Tumor Node Metastasis
VAF	Variant Allele Frequency
WGS	Whole Genome Sequencing
WHO	World Health Organization

Studies included in the thesis

The studies are referred to in the text by their Roman numerals.

I. Clinical framework for next generation sequencing based analysis of treatment predictive mutations and multiplexed gene fusion detection in non-small cell lung cancer. *Lindquist KE*, **Karlsson A***, Levéen P, Brunnström H, Reuterswärd C, Holm K, Jönsson M, Annersten K, Rosengren F, Jirström K, Kosieradzki J, Ek L, Borg Å, Planck M, Jönsson G, Staaf J*. Oncotarget. 2017 May 23;8(21):34796-34810. *These authors contributed equally.

II. A combined gene expression tool for gene fusion detection and histological prediction in non-small cell lung cancer from archival tissue. ***Karlsson A**, Cirenajwis H, Ericson Lindquist K, Brunnström H, Reuterswärd C, Jönsson M, Micke P, Patthey A, Behndig AF, Johansson M, Planck M, Staaf J.* Submitted manuscript.

III. Gene Expression Profiling of Large Cell Lung Cancer Links Transcriptional Phenotypes to the New Histological WHO 2015 Classification. ***Karlsson A**, Brunnström H, Micke P, Veerla S, Mattsson J, La Fleur L, Botling J, Jönsson M, Reuterswärd C, Planck M, Staaf J.* J Thorac Oncol. 2017 Aug;12(8):1257-1267.

IV. Mutational and gene fusion analyses of primary large cell and large cell neuroendocrine lung cancer. ***Karlsson A**, Brunnström H, Ericson Lindquist K, Jirström K, Jönsson M, Rosengren F, Reuterswärd C, Cirenajwis H, Borg Å, Jönsson P, Planck M, Jönsson G, Staaf J.* Oncotarget. 2015 May 23;8(21):34796-34810.

V. Genome-wide DNA methylation analysis of lung carcinoma reveals one neuroendocrine and four adenocarcinoma epitypes associated with patient outcome. ***Karlsson A**, Jönsson M, Lauss M, Brunnström H, Jönsson P, Borg Å, Jönsson G, Ringnér M, Planck M, Staaf J*. Clin Cancer Res. 2014 Dec 1;20(23):6127-40.

Related studies not included in thesis

Relation between smoking history and gene expression profiles in lung adenocarcinomas. *Staaf J, Jönsson G, Jönsson M,* ***Karlsson A****, Isaksson S, Salomonsson A, Pettersson HM, Soller M, Ewers SB, Johansson L, Jönsson P, Planck M*. BMC Med Genomics. 2012 Jun 7;5:22.

Histological specificity of alterations and expression of KIT and KITLG in non-small cell lung carcinoma. *Salomonsson A, Jönsson M, Isaksson S,* ***Karlsson A****, Jönsson P, Gaber A, Bendahl PO, Johansson L, Brunnström H, Jirström K, Borg Å, Staaf J, Planck M*. Genes Chromosomes Cancer. 2013 Nov;52(11):1088-96.

Landscape of somatic allelic imbalances and copy number alterations in human lung carcinoma. *Staaf J, Isaksson S,* ***Karlsson A****, Jönsson M, Johansson L, Jönsson P, Botling J, Micke P, Baldetorp B, Planck M.* Int J Cancer. 2013 May 1;132(9):2020-31.

Genomic and transcriptional alterations in lung adenocarcinoma in relation to smoking history. ***Karlsson A****, Ringnér M, Lauss M, Botling J, Micke P, Planck M, Staaf J.* Clin Cancer Res. 2014 Sep 15;20(18):4912-24.

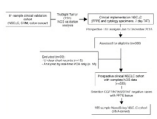
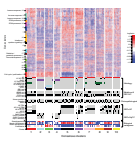
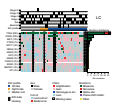
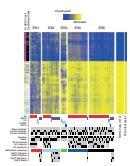
Epigenetic modifications in KDM lysine demethylases associate with survival of early-stage NSCLC. *Wei1 Y, Liang J, Zhang R, Guo Y, Shen S, Su L, Lin X, Moran S, Helland Å, Bjaanæs MM,* ***Karlsson A****, Planck M, Esteller M, Fleischer T, Staaf J, Zhao Y, Chen F, Christiani DC*. Clin Epigenetics. 2018 Apr 2;10:41.

A multi-omic study reveals BTG2 as a reliable prognostic marker for early-stage non-small cell lung cancer. *Shen S, Zhang R, Guo Y, Loehrer E, Wei Y, Zhu Y, Yuan Q, Moran S, Fleischer T, Bjaanæs MM,* ***Karlsson A****, Planck M, Staaf J, Helland Å, Esteller M, Su L, Chen F, Christiani DC*. Mol Oncol. 2018 Apr 15.

DNA Methylation of LRRC3B: A Biomarker for Survival of Early-Stage Non-Small Cell Lung Cancer Patients. *Guo Y, Zhang R, Shen S, Wei Y, Moran S, Fleischer T, Bjaanæs MM,* ***Karlsson A****, Planck M, Su L, Zhu Z, Staaf J, Helland Å, Esteller M, Christiani DC*. Cancer Epidemiol Biomarkers Prev. 2018 Sep 5.

Chapter 10: Clinical application of fusion gene detection using next-generation sequencing and the NanoString technology. ***Karlsson A****, Staaf J*. Tumor Profiling. Methods in Molecular Biology: Springer; 2018.

Thesis at a glance

STUDY	QUESTION	PATIENTS AND METHODS	RESULTS
I 	Is a clinical framework for treatment predictive mutation screening and gene fusion detection possible to establish?	533 patients with advanced disease in clinic routine analysis for treatment predictive mutation testing and gene fusion status using NGS and the NanoString technology.	Framework for treatment predictive mutation testing was successfully established. NanoString technology proved successful in fusion gene detection.
II 	Is it possible to predict NSCLC histology using a single sample predictor and simultaneously retrieve gene fusion status?	68 tumors of AC, SqCC and LCNEC histology using the NanoString technology and AIMS.	A multicomponent assay for fusion gene detection and NSCLC histology prediction was successfully established. SSP was validated in three external cohorts.
III 	Do the WHO2015 revised guidelines translate to the transcriptional landscape of lung cancer, with specific focus on LCC and LCNEC?	Global gene expression analysis of 159 lung cancers using the HT-12 Illumina microarray platform.	Clustering of gene expression data revealed stable clusters closely correlated with histopathological features, validated in external data sets. The WHO revised guidelines with specific focus on LCC an LCNEC translated well to the transcriptional landscape of lung cancer.
IV 	What mutations and gene fusions are frequent in LCC and LCNEC tumors?	41 LCC and 32 LCNEC tumors were screened for mutations using the NGS-based TST panel from Illumina. Gene fusion status was assessed with targeted RNAseq using an ArchedDx panel.	LCC could be further stratified based on mutation status. LCNEC showed similar mutational patterns as described in SCLC. No gene fusions were identified.
V 	Do epitypes exist in lung cancer?	Global methylation analysis of 124 lung cancers using the Illumina 450K methylation microarray platform.	Five distinct epitypes were identified and validated in external datasets. Epitypes were correlated with histopathological features and (in AC tumors) with survival.

Aims

Overall aim

Molecular profiling of lung cancer to stratify tumors and improve clinical management of NSCLC by implementing new technologies in routine diagnostics.

Specific aims

Study I

To build a framework and implement NGS in treatment predictive mutation testing of NSCLC patients diagnosed with advanced disease. Validation of the NanoString technology as a possible screening assay of fusion gene detection.

Study II

Simultaneous gene fusion detection and histology assessment using the NanoString technology.

Study III

Comprehensive global gene expression analysis of lung cancer to investigate whether the transcriptional landscape translate into the revised WHO guidelines regarding histological classification, with a special focus on LCC.

Study IV

Investigating mutations and gene fusion in LCC and LCNEC tumors.

Study V

Comprehensive global methylation analysis of lung cancer with the intent to classify tumors on the basis of differentiating methylation patterns.

Introduction and background

The normal lung

Lung function

The primary organs of the respiratory system are the lungs, which are responsible for a variety of life sustaining functions. Foremost functions include delivering oxygen into the bloodstream and carbon dioxide exchange retrieved from deoxygenated blood from the heart. The lungs serve as a protective filter against pollutants, small blood clots and infections, and the lungs are involved in maintaining homeostasis[1]. The trachea is branched into two bronchi, forming the lower respiratory tract consisting of the right and left lung. The right lung, being larger in size than the left, consists of three lobes compared with only two lobes in the left lung. The bronchial tree continues intralobulart into bronchioles, ending in far distal alveoli where gas exchange takes place[1, 2] (Figure 1).

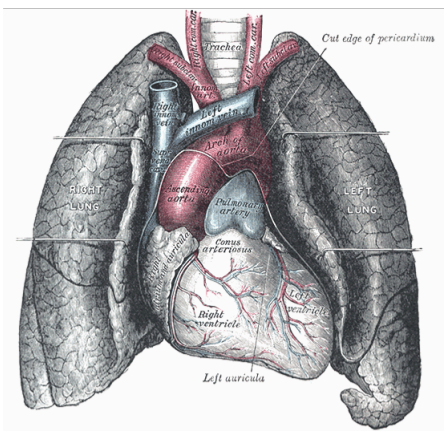


Figure 1. Anatomy of the human lungs. The human right and left human lung are part of the respiratory system, sectioned into lobes and are branched from the trachea to bronchi, bronchioles and alveoli[3].

Picture from Gray's Anatomy is in the public domain following expiration of its patent.

Lung morphology and progenitor cells

The entire lower respiratory tract is lined with respiratory epithelium. To facilitate gas exchange and host defense, lung epithelial cells in the distal lung are highly specialized and organized[1, 4, 5] (Figure 2). The trachea and bronchi are lined by a pseudostratified epithelium, consisting of basal, ciliated and secretory (Clara) cells. More distal from the bronchi, a simple columnar epithelium with ciliated and Clara cells lines the intralobular bronchioles with interspersed mucus-producing goblet cells. The latter cells clear the lungs of inhaled microorganisms and particulates. The alveolar epithelium consists of type 1 and type 2 alveolar epithelial cells (AEC). AEC2 are cuboidal and are most often found in the most distal part of the alveolus being the primary source of pulmonary surfactants (which decreases surface tension to maximize gas exchange and contributes to host defense)[5, 6]. It has been postulated, but not verified through extensive genetic lineage *in vivo*[5, 7] that AEC2 are progenitor cells for other alveolar epithelial cells. In contrast to AEC2 cells, AEC1 cells are squamous or flat, comprise the vast majority of the lung surface (~95%) and represent the major site for gas exchange and regulation of fluid homeostasis[8]. Importantly, even distal airways of the human lung are lined with pseudostratified epithelium where the basal cells, with their relatively undifferentiated nature are defined as a stem cell population capable of self-renewal and differentiation along the ciliated and secretory lineages[9-12].

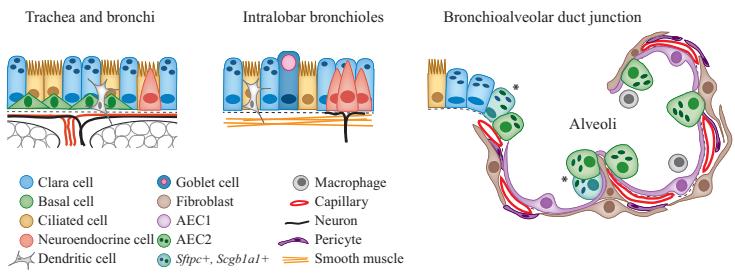


Figure 2. Lung progenitor cells. Lung epithelial cells line the entire respiratory tract. The trachea and bronchi consist mostly of Clara cells, basal cells and ciliated cells with a sparse amount of neuroendocrine cells. Bronchioles have a higher frequency of neuroendocrine cells with mucus-producing goblet cells. The distal airways and alveoli consist mostly of AEC1/AEC2 cells.

Lung cancer

Epidemiology, etiology and risk factors

Lung cancer accounts for 1.6 million deaths annually making it the deadliest form of cancer worldwide with an overall 5-year survival rate of 18.6%[13, 14]. Geographic and gender variations exist, and incidence rates are highest in Central and Eastern Europe for males and Northern America for females [15]. Smoking is the most prominent cause, and it is estimated to initiate up to 75-80% of all lung cancer cases[16]. Smoking duration and dose are steeply related to the risk of developing lung cancer, and the risk between genders appears to be similar, given the same level of tobacco consumption[17]. Due to the fact that cigarette smoke contains chemicals, which have the potential to directly or indirectly damage the respiratory epithelium, accumulation of specific genomic alterations is observed in lung cancers arising in smokers as compared with never-smokers[18]. Besides smoking, several etiological factors have been suggested to promote the disease including environmental tobacco exposure, air pollution, previous lung disease, radon exposure, various occupational carcinogens (asbestos, silica, arsenic) and genetic susceptibility[18-20].

Clinical management

Diagnosis

High mortality rates in lung cancer are primarily due to late diagnosis and can to a great extent be ascribed to initial symptoms being relatively vague, including loss of breath, fatigue, coughing, weight loss and chest pain[1, 21]. If lung cancer is suspected from an initial chest x-ray, this is followed by Computed Tomography (CT) guidance and, in cases of potentially curable disease, a Positron Emission Tomography (PET) scan[21]. Abnormalities detected with these methods are biopsied through bronchoscopy or thoracic puncture for histopathological diagnosis.

Staging

Stage dictates treatment as well as prognosis. The TNM classification for staging (Table 1) includes primary tumor size, the extension of lymph node involvement and presence or absence of tumor spread (metastases)[22-24]. PET-CT can reveal nodal (N) or metastatic spread (M). Endobronchial ultrasound-guided transbronchial needle aspiration (EBUS-TBNA) is used to assess mediastinal lymph node involvement (N) cytologically.

Table 1. TNM classification and staging in lung cancer. TNM classification of lung tumors are directly related to treatment and prognosis[25].

T/M	Label	N0	N1	N2	N3
T1	T1a *≤1*	IA1	IIB	IIIA	IIIB
	T1b *≤1-2*	IA2			
	T1c *>2-3*	IA3			
T2	T2a *>3-4*	IB	IIB	IIIA	IIIB
	T2b *>4-5*	IIA			
T3	T3 *>5-7, Inv, Satell*	IIB	IIIA	IIIB	IIIC
T4	T4 *>7, Inv, Ipsi Nod*	IIIA	IIIA	IIIB	IIIC
M1	M1a-b *Contr Nod, P1 Dissem, Single*	IVA	IVA	IVA	IVA
	M1c Multi	IVB	IVB	IVB	IVB

Lung cancer histology subtypes

Lung cancer is broadly divided into small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC) based on several morphological, histological and protein marker expression differences. NSCLC constitutes ~75% of all lung cancer cases and is further divided into refined subgroups, the three main being adenocarcinoma (AC), squamous cell carcinoma (SqCC) and large cell carcinoma (LCC). The different lung cancer subtypes arise from distinct cells of origin in the respiratory epithelium (Figure 3). Large cell neuroendocrine tumors (LCNEC) arise, like SCLC, from neuroendocrine cells and constitute a minor proportion of the lung cancer cases but represent an important differential diagnosis. SqCC tumors are often localized in the trachea or bronchi, SCLC in the bronchiole, while AC tumors often arise more distally, in the alveoli[26]. Classification of lung tumors are based on guidelines from the World Health Organization (WHO)[27]. Several features of the tumor are considered in the classification scheme, summarized in Figure 3.

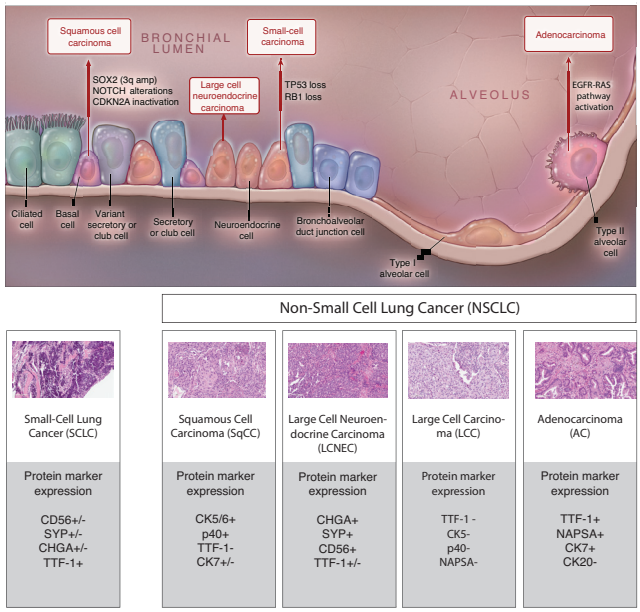


Figure 3. Histological subtypes of lung cancer and cells of origin. Lung cancer is broadly divided into small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC) based on morphological differences. The distinct subtypes arise from different cells of origin in the lung epithelium. NSCLC is further divide into three main subgroups: adenocarcinoma (AC), squamous cell carcinoma (SqCC) and large cell carcinoma (LCC) based on protein marker expression differences. Large cell neuroendocrine carcinoma (LCNEC) constitutes a minority of the lung cancer cases but represent an important differential diagnosis.

Reproduced and modified with permission from [26], Copyright Massachusetts Medical Society. Pictures of hematoxylin & eosin (H&E) stains kindly provided by Dr Hans Brunnström, department of Clinical Pathology, Lund.

Genomic landscape of lung cancer

Tumor development

The transition of a normal cell into a cancer cell relies on acquisition of certain capabilities depicted as hallmarks by Hanahan and Weinberg[28, 29]. These capabilities include: 1) sustaining proliferative signaling, 2) evading growth suppressors, 3) avoiding immune destruction, 4) enabling replicative immortality, 5) tumor-promoting inflammation, 6) activating invasion and metastasis, 7) inducing angiogenesis, 8) genome instability and mutation, 9) resisting cell death and 10) deregulation of cellular energetics. These hallmarks are crucial for tumor development and involve both cells of origin as well as the capability to create a tumor-promoting milieu, i.e. microenvironment. The most prominent cause of lung cancer development is smoking. Cigarette smoke is a complex mixture of carcinogenic compounds that causes inflammation, damage to the respiratory

epithelium and influences flow of air and blood[30, 31]. One particular effect of smoking and the different carcinogenic substances is the large number of somatic DNA mutations inferred through creation of DNA adducts in affected cells[32]. In fact, lung cancer represents, second to UV-induced melanoma the tumor type with the highest number of mutations, referred to as mutational load or tumor mutational burden[33]. In principle, the DNA mutations are a random process, and the DNA damages eventually lead to a tumorigenic process. However, smoking is not the sole cause of lung cancer, as a proportion of lung cancer patients (20-25%) are never smokers[19]. Several distinctions in mutational, transcriptional and methylation profiles have been noted through numerous studies comparing patients with a smoking history to never smokers. As to histological subtypes, SCLC, LCNEC, LCC, and SqCC typically affect smokers whereas never smokers´ tumors most often are of AC histology. Besides mutations, early events in NSCLC development include loss of heterozygosity (LOH) of chromosomal regions containing encoding sites for tumor suppressor genes such as Ras association domain family member 1 (*RASSF1*), nuclear fusion protein (*FUS1*), fragile histidine triad (*FHIT*), *p16* and tumor protein p53 (*TP53*)[34, 35].

Profiling of lung cancer using high-dimensional data approaches

As an effect of the evolution of high-dimensional and high-throughput technologies, mutational, transcriptional and epigenetic profiling of lung cancer, in numerous studies, have revealed distinct profiles. These profiles are associated with distinct histological subtypes, key oncogenic drivers, and etiology (e.g. smoking)[18, 19, 36-52]. Multiple studies have revealed a clear separation of AC, SqCC, and SCLC cases into gene expression subclusters, which are driven by specific transcriptional programs. Within both adenocarcinoma and SqCC, different gene expression phenotypes (GEPs) have been proposed, most widely adapted concerning tumors of AC histology being the terminal respiratory unit (TRU, formerly bronchioid), the proximal-inflammatory (PI, formerly squamoid), and the proximal-proliferative (PP, formerly magnoid) transcriptional subtypes[41, 42, 47, 48, 53-58]. In spite of the tremendous number of studies with the intent to map the landscape of lung cancer, no GEPs have yet been established (in contrary to, e.g. breast cancer). Regarding the genomic landscape of the disease, pinpointing genetic drives of tumorigenesis and characterizing tumor heterogeneity[26] has contributed to the development of targeted treatments.

Oncogene addiction and therapeutic targets

In total, about 50% of NSCLC cases are associated with mutations in specific proto-oncogenes (typically different tyrosine kinases) including the epidermal growth factor receptor (*EGFR*), Kirsten rat sarcoma viral oncogene homolog (*KRAS*), tyrosine-protein kinase Met (*MET*), proto-oncogene B-raf (*BRAF*) and human epidermal growth factor receptor 2 (*HER2*) as well as gene fusions

involving anaplastic lymphoma kinase (*ALK*), proto-oncogene tyrosine-protein kinase receptor Ret (*RET*), c-ros oncogene 1 (*ROS1*) and members of the neurotrophic tyrosine kinase (*NTRK*) and fibroblast growth factor receptor (*FGFR*) families. Alterations in these genes through activating mutations/gene rearrangements represent highly desirable therapeutic targets due to: 1) the concept of oncogene addiction, i.e. genetic alterations that governs the oncogenic potential of malignant cells that are crucial for tumor survival, and 2) that these alterations are mutually exclusive to other driver events. These potential targets for therapy are all associated with AC histology and the majority of mutations occur in the tyrosine kinases *EGFR* and *KRAS*. *EGFR* mutations are almost exclusively

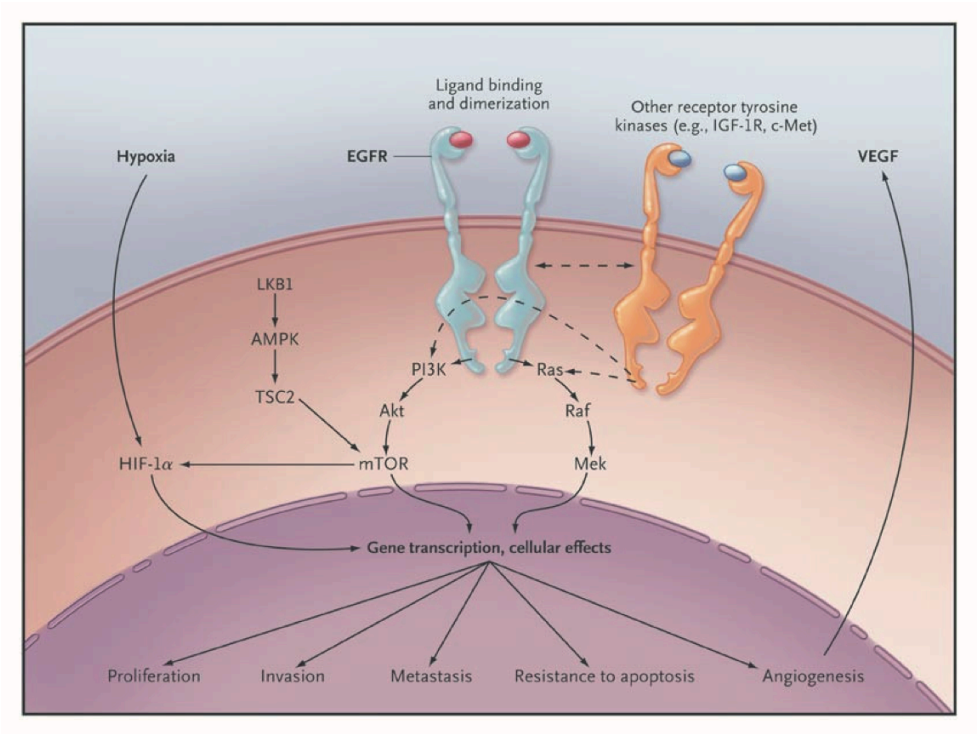


Figure 4. EGFR cell signalling pathways and receptor. In the presence of a ligand, such as a growth factor, dimerization of the EGF receptor trigges a cascade of intracelluar signalling through the Ras/Raf/Mek pathway and the PI3K/Akt/mTOR pathway. This leads to gene transcription and tumor promoting effects including proliferation, survival, invasiveness, metastatic spread, and tumor angiogenesis through pathways that are either dependent on or independent of the hypoxia inducible factor (HIF). These pathways also may be modulated by other receptor tyrosine kinases, such as insulin-like growth factor 1 receptor (IGF-1R) and cMET, and by the LKB1–amp-activated protein kinase (AMPK) pathway, which is involved in energy sensing and cellular stress.

Reproduced with permission from [34], Copyright Massachusetts Medical Society.

associated with a non-smoking history, while *KRAS* mutations are more frequent in tumors from smokers[34, 43, 44, 59-61]. Serving as a prototypical example of a tyrosine kinase oncogene, the EGFR gene is activated by binding of growth factors to its ligand, triggering a cascade of intracellular processes that involves proliferation and DNA synthesis through the RAS/MAPK, PI3K/Akt and STAT pathways (Figure 4). In 2004, two independent research groups identified somatic mutations in the tyrosine kinase (TK) domain of *EGFR* [62, 63]. Exons 18-24 of the *EGFR* gene encodes for the TK domain and somatic mutations verified to have an activating effect on the TK domain affects mainly exons 18-21. A majority of these mutations (80-90%) are of two variants: small deletions in exon 19 or point mutations (substitutions) in exon 21 (typically the L858R mutation). Activating mutations of the EGFR receptor leads to dimerization of the transmembrane receptor and activation of the ras-pathways even in the absence of a binding ligand, triggering a signaling cascade resulting in tumor promoting activities (Figure 4). Tyrosine kinase inhibitors (TKIs) are therapeutic agents that interacts with the TK domain of the receptor leading to inhibition of the tumor promoting activities[34]. Targeted therapy using TKIs have shown a remarkable prolonged survival in patients with mutations/gene rearrangement associated with oncogene addiction. Unfortunately, patients treated with TKIs eventually relapse due to development of resistance mutations either within the targeted gene itself (e.g. T790M or C797S mutations in *EGFR*) or other genes (amplification of another oncogene like *MET*)[61]. Those resistance mutations either alters, e.g. the 3D protein structure (hindering binding of a drug, e.g. T790M), or allows rewiring of the signal network of tumor promoting pathways. Although the vast majority of NSCLC mutations affect the *KRAS* gene, no effective targeted therapy has yet been developed. SCLC and LCNEC tumors are characterized by high genetic instability and inactivating mutations affecting the genes *TP53* and *RB1*. No targetable drivers have yet been identified in SCLC and LCNEC tumors[27, 64].

Gene rearrangements as oncogene drivers

Since the discovery[65] of recurrent fusion events involving the Echinoderm Microtubule Associated Protein Like 4 (*EML4*) gene and the proto-oncogene *ALK,* which result in a highly oncogenic protein underlying tumor development in lung cancer of primarily AC histology, novel gene fusions and alternate splicing events are continuously discovered. The oncogenic potential of a fusion gene relies on the kinase domain of the proto-oncogene being conserved. For the fusion gene to have a favorable tumorigenic effect, protein synthesis must occur. As transcriptional direction moves from the 5' end to the 3' end of the transcribed gene, the oncogenic potential of the fusion gene is observed as elevated expression of the 3' end of the proto-oncogene (containing the kinase domain) compared with the un-transcribed 5' end (Figure 5A)[66, 67]. Multiple recurrent fusion genes have been mapped and associated with oncogenic potential driving NSCLC tumor

development. In today's clinical practice, fusion gene status of *ALK* and *ROS1* are included in routine treatment predictive testing, with the potential option of targeted TKI treatment using crizotinib or more recent drugs if a gene fusion is detected[68]. Another emerging target is the *MET* exon 14 skipping alteration[69]. Due to mutation of the *MET* gene, aberrant splicing can occur causing intron retention or exon skipping that results in an oncogenic driver of tumorigenesis (Figure 5B)[70]. Several ongoing clinical trials indicate the use of TKIs such as crizotinib or cabozantinib in patients with an established *MET* exon 14 skipping event due to *MET* mutation[71-74].

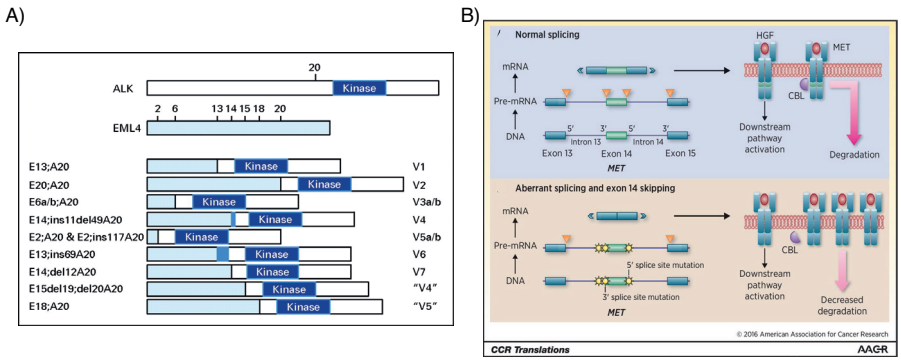


Figure 5. A) *EML4-ALK* fusion gene. Schematic representation of *EML4-ALK* gene fusions NSCLC. Multiple *EML4-ALK* variants (V1 to V7) have been identified in NSCLCs. All involve the intracellular tyrosine kinase domain of *ALK* starting at a portion encoded by exon 20. *EML4*, however, is variably truncated. **B) *MET* exon 14 skipping due to mutations of the *MET* gene.** *MET* mutations (yellow) that disrupt the branch point and/or 3′ splice site of intron 13, and the 5′ splice site of intron 14 result in aberrant splicing and exon 14 skipping. MET exon 14 is thus excluded in mRNA that is later translated into a protein product lacking the Y1003 residue. Loss of this region leads to decreased MET receptor ubiquitination by CBL. Decreased degradation results in oncogenesis driven by increased levels of MET.

Reproduced and modified with permission from[66], American Society of Clinical Oncology and[74], American Association for Cancer Research, Copyright Clearance Center.

DNA methylation and epigenetics

Epigenetics is a wide term that includes DNA methylation, chromatin remodeling and microRNA (miRNA, short noncoding RNA) regulation. As a result of gene mutation, the DNA is altered, which gives rise to a modified or absent protein during translation. On the contrary, DNA methylation results in gene silencing but the structure of the DNA remains the same, which leads to the absence of a protein during translation that cannot be ascribed to a modification of the DNA. This is due to the fact that methylated cytosine (C_m) cannot be discriminated from an unmethylated cytosine (C) structurally[75, 76]. CpG sites (CpGs) are DNA regions where a C is followed by guanin (G). CpG islands are enriched for CpG sites (300-3000 bp long) that often exists in the promoter region of genes. About 70% of

human promoter regions contain CpG islands[77, 78]. Hypermethylation of key CpGs results in gene silencing, i.e. loss of expression, while hypomethylation can cause over-expression of the gene. In cancer, loss of expression due to hypermethylation has been estimated to be ten times more frequent than due to mutation[79]. Smoking induced methylation of *p16* (*CDKN2A* gene) and *FHIT* are thought of as early events in certain lung cancers, observed in pre-malignant squamous-cell lesions of the lung (i.e. hyperplasia or metaplasia) whereas *p16* methylation in AC is a rare event that occurs at a much later stage in tumor development[34, 80-86]. Integrative studies using genome-wide methylation and global gene expression have revealed phenotypes associated with mutations of oncogenic drivers[51], histological subtypes[40], smoking habits[87] related to tumor aggressiveness and prognosis and identification of differentially methylated regions that were predictive of treatment efficacy[52]. The Cancer Genome Altas (TCGA)[88] comprehensive mapping of lung cancers of AC histology identified subgroups labeled as CpG island methylator phenotypes (CIMPs)[42]. CIMPs refers to an exceptional high frequency of hypermethylated CpGs[89] and TCGA categorized three CIMPs identified as CIMP-high, CIMP-low and CIMP-intermediate where the CIMP-high phenotype could be further stratified on the basis of inflammatory processes, mutation rates and GEPs. An extended TCGA analysis of NSCLC tumors identified nine subtypes with two AC subtypes associated with CIMP stratified by activation of the immune checkpoint pathway (i.e. immunotherapy candidates)[90]. Overall, genome-wide methylation studies add important information to pre-existing knowledge about the transcriptional and mutational landscape of lung cancer, and may contribute to a more enhanced tumor classification and potentially improved clinical management.

Lung cancer treatment and therapy options

Treatment options for lung cancer rely highly on tumor stage, histology and molecular markers (mutation status of *EGFR* and *BRAF*, fusion gene status of *ALK* and *ROS1,* and protein expression of PDL1). Patients with low stage tumors are treated with a curative intent. Primary choice is surgery with or without the addition of adjuvant chemotherapy. Unfortunately, due to frequent late stage diagnosis and/or (often smoking-related) co-morbidity, most patients are treated with a palliative intent. Figure 6 outlines treatment options for lung cancer patients[21, 91].

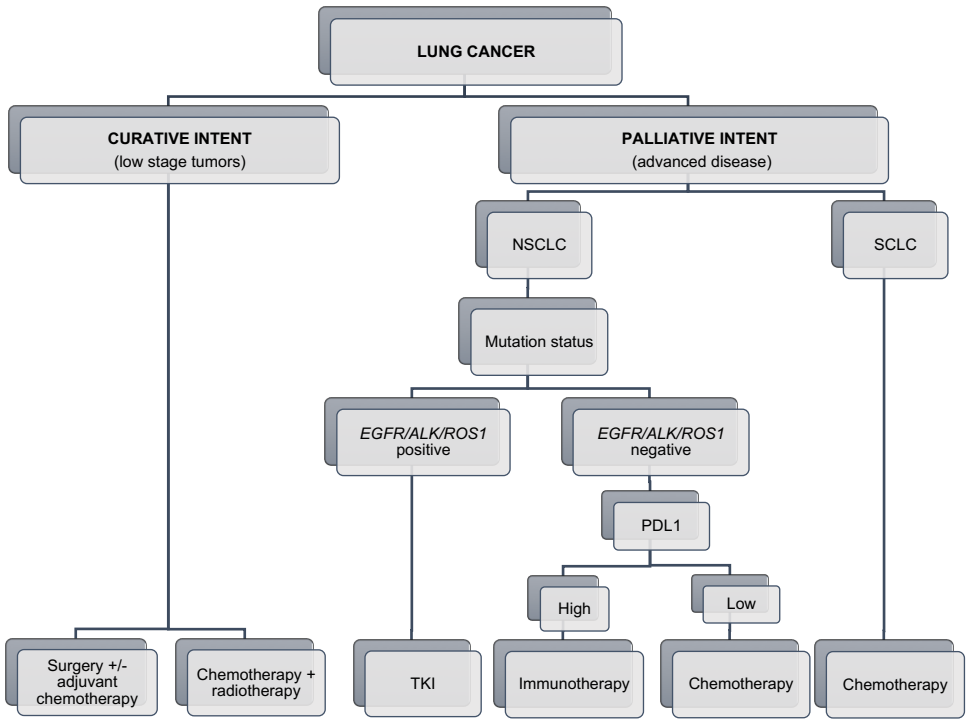


Figure 6. Lung cancer treatment and therapy options. Therapy decisions are made based on several criteria including disease stage, tumor histology and molecular status, patient performance status, age and co-morbidities.

Surgery

Surgery is the primary curative option for lung cancer patients diagnosed with NSCLC stage I, II or possibly IIIA. The standard procedure in lung cancer resection is lobectomy, i.e. removal of the tumor-infested lobe. Bi-lobectomy can be performed if right-sided lung tumor growth has affected more than one lobe. In selected cases, a pulmectomy (resection of an entire lung) is required. During surgical resection, lymph node sampling or dissection is performed in order to assess pathological N-stage as part of TNM staging. Patients with resected tumor stage IB-III are offered adjuvant chemotherapy[59, 91].

Chemotherapy

Chemotherapy is used in lung cancer treatment with both a curative and palliative intent. Differences between strategies exist. The patient's performance status (Table 2) is a prominent factor for therapy decisions.

Table 2. WHO criteria for patient performance status[92].

GRADE	EXPLANATION OF ACTIVITY
0	Fully active, able to carry on all pre-disease performance without restriction
1	Restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature, e.g., light house work, office work
2	Ambulatory and capable of all selfcare but unable to carry out any work activities. Up and about more than 50% of waking hours
3	Capable of only limited selfcare, confined to bed or chair more than 50% of waking hours
4	Completely disabled. Cannot carry on any selfcare. Totally confined to bed or chair

Adjuvant chemotherapy

Several studies have shown a significant increase in survival rates associated with adjuvant platinum-based chemotherapy post-surgery - standard for NSCLC is cisplatin in combination with vinorelbin[91, 93, 94].

First-line chemotherapy with palliative intent

In advanced disease, chemotherapy remains the only choice if the tumor is *EGFR* mutation negative or *ALK*/*ROS1* fusion negative and has a low PDL1 expression. First-line chemotherapy is typically administered as a combination of one platinum substance (cisplatin or carboplatin) and one of the following: docetaxel, gemcitabine, paklitaxel, pemetrexed or vinorelbin. Studies have shown both a slightly higher efficacy and a slightly higher toxicity with cisplatin than with carboplatin[91, 95, 96]. Tumor histology affects choice of chemotherapy as tumors of AC and LCC histology have shown a higher sensitivity towards pemetrexed whereas SqCC have not[97]. For the rare LCNEC cases, the optimal platinum combination is still under debate, but, currently, LCNEC is often treated as SCLC (see below), rather than as NSCLC, in first line[98].

Tumor progression or relapse

Few studies exist on which chemotherapy regimen is preferred in case of NSCLC progression or relapse. In case of relapse, it has been debated whether to use the primary chosen therapy strategy or not, but no guidelines exist. At tumor progression, immunotherapy (see below) has recently become the preferred alternative in second-line treatment, but if immunotherapy is contraindicated (due

to e.g. uncontrolled brain metastases or autoimmune diseases), chemotherapy (preferably a regimen new to the patient) is chosen.

Small cell lung cancer treatment

In stage I-III SCLC, chemotherapy (preferably cisplatin plus etoposide) is combined with curative radiotherapy, whereas patients with stage IV disease receive chemotherapy with palliative intent.

Radiotherapy

Radiotherapy for lung cancer is used in many different situations. After lung cancer surgery, in case of incomplete resection, curatively intended radiotherapy can be added to the treatment. In case of non-resectable, locally advanced tumors, radiotherapy with curative intent is typically given in combination with chemotherapy. For low stage tumors in medically inoperable patients, high-dose stereotactic body radiation therapy (SBRT) can be a curative option [99-103]. Furthermore, radiotherapy is frequently given with a palliative intent, e.g. as pain relief (typically in cases with skeletal metastases) or to handle tumor burden at critical locations such as central airway/circulation or CNS[91].

Targeted therapies

Targeted therapies in lung cancer include tyrosine kinase inhibitors (TKIs) of the *EGFR-, BRAF-, ALK-,* or *ROS1*-receptors, angiogenesis inhibitors (targeting the *VEGF* receptor), or immunotherapy using anti-PD1 or anti-PDL1 agents. Treatment predictive mutation testing is an important tool in clinical management because tumors positive for *EGFR*/*BRAF* mutation or *ALK/ROS1* gene fusion events make the tumor sensitive to TKIs and are therefore used as first-line treatments in advanced stage disease. Thus, results from treatment predictive testing is needed before patients can start therapy, stressing the need of rapid, sensitive, and accurate molecular methods in clinical diagnostics. Tumors that are *EGFR*/*ALK*/*ROS1* negative, are tested for PD1/PDL1 protein expression to guide the use of immunotherapy.

Tyrosine kinase inhibitors

Randomized studies have shown a remarkable improvement in survival rates by the use of TKIs compared to chemotherapy[104, 105]. The importance of activating *EGFR* mutations in lung cancer patients was discovered in 2004[62, 63]. Erlotinib, gefitinib (1st generation EGFR inhibitors) or afatinib (2nd generation) are now routinely used as first-line monotherapy in advanced disease for patients harboring these specific alterations. Despite rapid responses, treatment resistance generally occurs for these TKIs, often in the form of secondary mutations of *EGFR* (e.g. the T790M resistance mutation) within 9-14 months[106] using 1st generation TKIs. In response to this, a 3rd generation TKI, osimertinib, has been developed. Gene fusions involving the *ROS1* gene make the tumor sensitive towards crizotinib, with similar often dramatic responses as for *EGFR* mutated patients (and resistance development)[107]. *ALK* fusion positive tumors have previously been treated with crizotinib but recently alectinib is generally the first-line monotherapy choice due to increased overall survival (OS) rates but also less risk of CNS metastasis. Identification of *BRAF* V600 mutation indicates use of dabrafenib and/or trametinib TKI first-line monotherapy. Research on resistance mutations and TKIs are constantly evolving, paving way for an increased number of more efficacious targeted therapies, as well as a better understanding of how and when drugs should be used.

Anti-angiogenesis

Studies have shown that a proportion of patients with tumors of AC histology and non-eligible for TKI treatment have an increased therapy response and progression free survival (PFS) when treated with bevacizumab in addition to chemotherapy. Bevacizumab is a monoclonal antibody that blocks the *VEGF* receptor, but no treatment predictive test has yet been developed to indicate anti-angiogenesis therapy[91, 108-113].

Immunotherapy

Immunotherapy using anti-PD1/anti-PDL1 (nivolumab, pembrolizumab/ atezolizumab) has emerged as a novel therapeutic option for NSCLC. Tumors that express PDL1 in >50% of tumor cells (using immunohistochemistry) can be treated with pembrolizumab in first-line. In second-line, all three approved drugs are used. The field of immunotherapy in lung cancer is evolving rapidly. Several clinical trials, which combine immunotherapy with chemotherapy compared with using only one of the mentioned immunotherapies, are ongoing[114, 115].

Material

Biobanks

Several internal biobanks, evident from Figure 7, have been available for selection of tumor material, including both fresh frozen tissue from early-stage resected tumors as well as archived tumor tissue from patients with advanced disease. All patients included have been informed by a written consent. All studies have been reviewed and approved by the Regional Ethical Review Board in Lund, Sweden according to the Helsinki declaration.

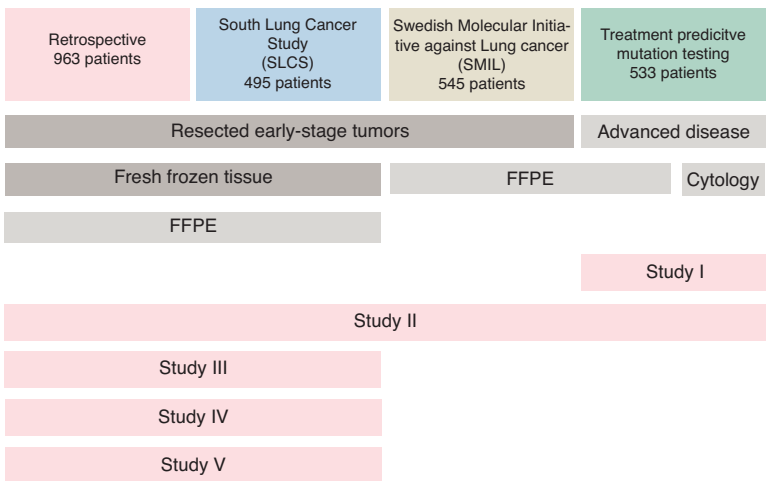


Figure 7. Internal biobanks constituting the basis of this thesis work. Three cohorts of fresh frozen tissue were available from patients diagnosed at an early stage and submitted to surgery. One cohort with patients diagnosed at an advanced stage was available.

Cell lines

Cell lines with genetic alterations associated with lung cancer were cultivated and harvested prior to nucleic acids extraction. Specific cell lines include HCC78 (*SLC34A2-ROS1* fusion), LC-2/ad (*CCDC6-RET* fusion), NCI H228 (*EML4-ALK* fusion) and KARPAS 299 (*ALK-NPM1* fusion). Extracted RNA was used as positive control for methods, which detect fusion genes associated with lung cancer.

Methods

Extraction of nucleic acids

Downstream data and results all rely on the quality of the input. No results can be better than the starting material. Therefore, the choice of extraction method for the very basis of research is essential[116]. Studies in this thesis have been based on different tissue origins: formalin-fixed paraffin embedded (FFPE) tissue (Studies I, II, IV), cytology specimens (Study I) and fresh frozen tumor tissue (Studies II, III-V). Since protocols differ depending on tissue origin, several methods for extraction of nucleic acids have been used. For FFPE tissue, a column-based commercial method was used (Allprep DNA/RNA FFPE Kit, Qiagen) that has proven superior in multiple studies[117-119]. For cytology specimens, merely DNA was extracted from cytology slides at the regional clinical pathology department using a commercial kit (QIAamp DNA Micro Kit, Qiagen) prior to the downstream treatment predictive mutational screening performed in Study I. Studies II-V are entirely (Studies III and V) or partially (Studies II and IV) based on fresh frozen tumor tissue obtained from surgically dissected tumors of early stage lung cancer patients. Extraction of RNA and DNA was performed using a column-based modified commercial method (AllPrep DNA/RNA Mini Kit, Qiagen)[120].

Quality control of nucleic acids

Prior to all downstream analyses, quantification and quality control of nucleic acids were performed. Quantification was performed either using the NanoDrop Fluorometer (ThermoFisher Scientific) or the Qubit system (ThermoFisher Scientific). Quality control of nucleic acids from fresh frozen tissue and FFPE varied due to the fact that FFPE DNA/RNA are of lower quality (more degraded). In Studies I and IV, DNA was subjected to a q-PCR assay using primers and reference supplied by Illumina that assess the amplifiability of the material. A delta Ct was calculated and quality of input DNA was based on the calculated value. Input amount to downstream analysis was based on the delta Ct value according to the manufacturer's instructions. All RNA extracted from fresh frozen

tumor tissue was evaluated on the Agilent Bioanalyzer[121], calculating a Ribosomal Integrity Number (RIN) value to assure intact RNA. RNA extracted from FFPE tissue was evaluated on the Agilent Bioanalyzer, calculating the DV^{200} to assure that the majority of the fragmented RNA was of a minimum length of 200 nucleotides.

Next-generation sequencing

One of the most helpful and used tools within molecular biology and clinical molecular diagnostics is sequencing. The so-called Sanger sequencing technique has been used since the 1970s[122] and in 2005 this first-generation method was revolutionarily evolved. The capability to sequence multiple samples at single-nucleotide resolution revealed massive options for research and daily clinical diagnostics and the method is termed next-generation sequencing (NGS). There are multiple options when it comes to NGS and although whole genome sequencing (WGS) would supply all information necessary, this is merely an option for a variety of reasons including time, economical resources, data management, sample quality, data storage and clinical relevance. Anything that is, or can be transformed into, DNA can be subjected to NGS. Prior to sequencing, a library preparation is performed. Depending on your intentions with NGS data, the library preparation differs. If starting material is RNA, this needs to be converted into cDNA. If using targeted sequencing, i.e. selection of regions/genes of interest to be sequenced, these are tracked out from the sample DNA/cDNA by hybridization to a pre-designed probe pool (Studies I and IV) or by gene-specific primers (Study IV). During library preparation, sequencing adapters and index (for unique sample identification) are added and the library is amplified using PCR followed by purification of the PCR product. The purified, index tagged and adapter ligated PCR product is pooled with PCR products treated similarly prior to sequencing. Pooled samples are then sequenced in a parallel manner where the adapter sequences attach to the surface of the sequencing flow cell. Clusters are bridge amplified and generated clusters are subjected to fluorescently labeled nucleotides that are incorporated base by base. As each base is incorporated, fluorescence emission is recorded. This process is called sequencing by synthesis (SBS) by Illumina and the process is repeated for as many times (cycles) creating a length (read) as defined by the user and chemicals used. After sequencing, data is aligned i.e. mapped towards a reference genome (Figure 8).

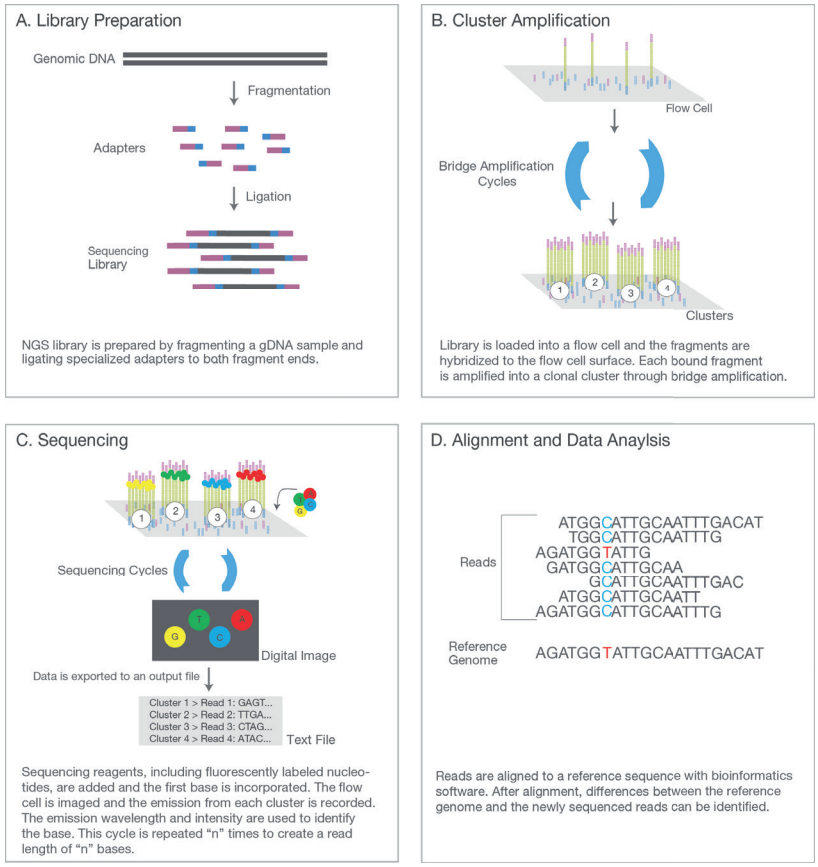


Figure 8. Concept of Illumina NGS and data generation. Fragmented DNA is ligated with adapters that hybridize to the flow cell surface. The bound fragments are amplified to form clusters and fluorescently labeled nucleotides are incorporated. Through imaging, the emission from the incorporated fluorscently labeled nucleotides is recorded, and the wavelength is used to identify the base. This process is repeated for a given number of cycles to create a given read length. Reads are aligned to a reference genome to identify similarities or differences between the sequenced reads and the reference genome.

Mutation detection using NGS

The TruSight Tumor 26 library preparation kit from Illumina was used to detect mutations in Studies I and IV. This focused panel has been developed to identify low-frequency variations across 26 solid tumor related genes. The assay has been optimized to use FFPE derived DNA and it utilizes two pools with gene specific primers during library preparation. In fact, one sample is sequenced twice and data is merged from both probe pools. By this action, formalin induced nucleotide remodeling can be flagged as artifacts and not a true mutation as only nucleic acid changes present in both pools are considered true mutations. All libraries were sequenced on a MiSeq instrument and the data were aligned to the Human UCSC

hg19 reference genome. Mutation detection was performed in the Illumina supplied software VariantStudio where variants were called and labeled.

Fusion gene detection using NGS

Targeted RNA sequencing (RNAseq) using NGS was employed in Study IV to identify therapy targetable gene fusions frequently involved in lung cancer development (e.g. *ALK*, *RET*, *ROS1*) and in Study I to identify a novel fusion not detected with the method used initially. The commercially available Archer DX FusionPlex assay[123] is based on a proprietary Anchored Multiplex PCR (AMP™) target enrichment chemistry to detect fusions of all genes in a single sequencing assay, even without prior knowledge of fusion partners or breakpoints. The kit also detects selected insertions and point mutations in *ALK* and *RET*, including those reported in cell-based assays to convey crizotinib resistance[124]. The kit has been developed to handle low amounts of highly degraded RNA (i.e. from archived material). After the initial cDNA conversion, which in some cases is not successful in low-quality RNA samples, library preparation is performed using AMP and library is sequenced as described previously. Data analysis to detect fusions is performed using a software provided by Archer DX, resulting in aligned and annotated fusion transcripts.

Multicomponent analysis

The NanoString technology

The NanoString technology[125] is based on the dual hybridization of a capture and a molecularly barcoded reporter probe complementary to a contiguous target sequence. A capture probe consists of a target-specific, biotinylated probe while a reporter probe consists of a target-specific probe linked to fluorescently-labeled tags that serves as a barcode in the multiplex assay. The capture probe and the reporter probe are referred to as a probe set (Figure 9). Probe sets are directly hybridized to RNA transcripts with no prior cDNA synthesis or enzymatic steps. On removal of excess probes, the hybridization complex is immobilized to a streptavidin-coated surface and aligned. The sequence-specific, fluorescently labeled reporter barcodes are digitally imaged and counted. The number of unique reporter barcodes specific to a target sequence is proportional to the number of transcripts present[126-129]. The technology was developed to overcome the challenges of highly degraded RNA used as input and is therefore ideal for FFPE samples used in Studies I and II. The NanoString technology is referred to as a

multicomponent assay as it simultaneously derives gene expression data and fusion gene status in one single assay for multiple user-defined targets (probe set).

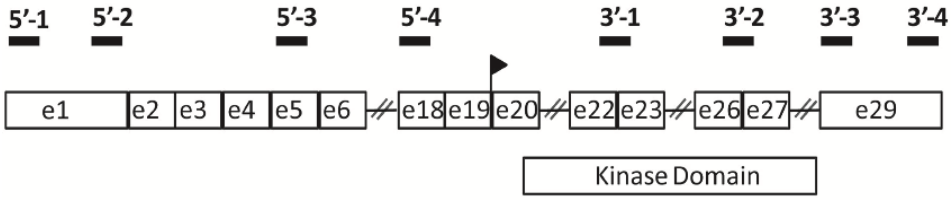


Figure 9. NanoString probe design. Probes span the entire gene of interest.

Reproduced with permission from[126], The Journal of Molecular Diagnostics, Elsevier and Copyright Clearance Center.

Gene expression analysis

In Study II, gene expression data generated with the NanoString technology was used. Gene expression is represented by number of counts registered from the target-specific probe representing a corresponding sequence of the human transcriptome. Probes span the entire gene of interest from 5' end to 3' end (Figure 9). In contrary to global gene expression analysis, the NanoString technology is limited by the number of targets to be included in the probe set. In Study II, gene expression data were not normalized but a background correction using spiked-in references was performed. Gene expression data corresponding to 11 genes associated with NSCLC histology classification were extracted for further classification purposes.

Fusion gene detection

Fusion gene detection using the NanoString technology was performed in Studies I and II. Probe sets were designed with the intent to identify fusion genes frequently involved in lung cancer development and treatment and to simultaneously retrieve gene expression data on genes of significance in NSCLC. These genes correspond to protein markers that are used in routine histopathological diagnostic setting,

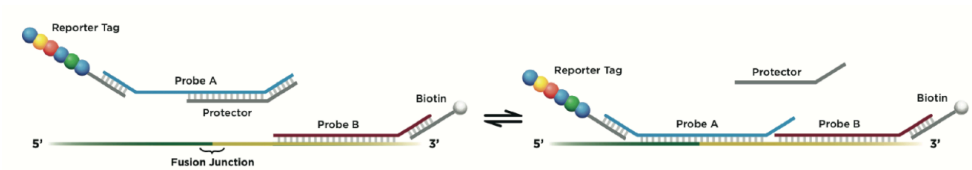


Figure 10. NanoString probe set. A target specific biotinylated capture probe and a target specific reporter probe linked to a fluorescently labeled barcode adress is referred to as probe set.

Reproduced with permission from[130], Springer Nature and Copyright Clearance Center.

markers for prognosis and immunomarkers. Using the NanoString technology for fusion gene detection involves: 1) target-specific sequences that spans the gene of interest from 3' end to the 5' end (Figure 9) and 2) fusion specific probes that spans the junction of known fusion genes. A so-called protector probe assures hybridization specificity (Figure 10). Fusion is called on the basis of the two above mentioned criteria. An imbalance in counts registered from the probes representing the 3' end of the affected gene (containing the kinase domain) compared to 5' end indicates fusion involving the investigated gene. The fusion specific probe reveals the fusion partner as elevated counts corresponding to elevated expression of the fusion transcript[126].

Immunohistochemistry

Immunohistochemistry (IHC) is a well-established and widely applied technique for documenting protein marker expression. Lung cancer subtypes are associated with differentiating protein marker expression (Figure 3). Morphology (derived from H&E stains) and scoring of immunomarkers is standard histological subtyping techniques used by pathologists worldwide. Guidelines for histology subtyping based on morphology and immunomarkers are administered through WHO and revised periodically[27]. All histopathological subtyping in this thesis work has been performed in line with the WHO guidelines. In addition, complimentary immunohistochemistry has been used to further discriminate subtypes or verify findings. In Studies III and IV, additional stainings to stratify LCC and LCNEC cases were performed including neuroendocrine markers (chromogranin A, synaptophysin, CD56), squamous cell markers (CK5, P40), adenocarcinoma markers (TTF-1, Napsin A) and RB1 (associated with SCLC)[131-133]. In Study II, additional stainings for mucin markers, in an attempt to histologically classify tumors clinically diagnosed as NSCLC-NOS, were performed including CDX2 and Periodic acid–Schiff–diastase (PAS-D).

Microarrays

Methodological concept

The overall concept of microarrays, regardless of the intention to study the transcriptome (RNA), genome (DNA) or epigenome (DNA methylation patterns), is to simultaneously derive information on thousands of genes of the human

transcriptome/genome/epigenome in a high throughput manner. Microarrays have shown to be a stable and trustworthy technique for investigating the transcriptional/genomic/epigenetic landscape since the first introduction of the technique in the late 1990s. Many commercial arrays are available such as the Illumina BeadArray technology[134]. Here, beads coated with probes corresponding to a sequence of the human transcriptome/genome/epigenome are scattered onto a silica surface. Pre-processed RNA/DNA is hybridized to the probe sequences, labeled (single or dual), washed and scanned. Emitted fluorescence is registered and calculated as overexpression/loss of expression (transcriptome), gain/loss (genome) or hypermethylation/hypomethylation (epigenome). Microarrays are used in Studies III and V to explore the transcriptional and epigenetic landscape of lung cancer (Figure 11).

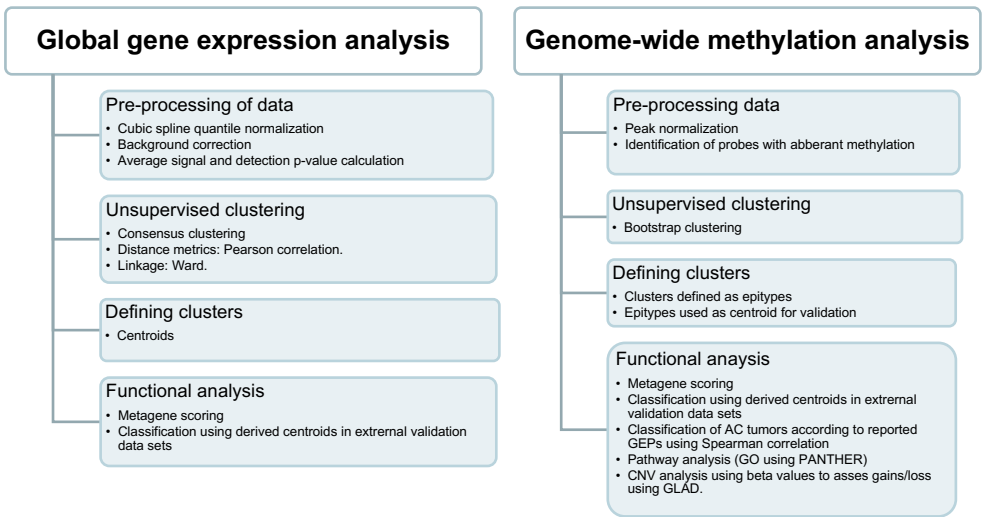


Figure 11. Microarray data analysis methods used in Study III and V. Microarrays were used to map the transcriptional and epigentic landscape of NSCLC. After an initial pre-processing of data, unsupervised clustering was performed. After defining clusters generated by clustering, gene expression centroids and epitypes were created and validated in external datasets. Functional analysis of differentially expressed genes, hypo- or hypermethylated genes was performed using pathway analysis.

Global gene expression analysis using microarrays

To explore the transcriptional landscape of lung cancer in Studies III and V, Illumina Human-HT-12v4.0 BeadChip arrays[134] were used for single-channel detection (Figure 12). The HT-12 gene expression microarrays allow for investigation of 47 231 probes (oligonucleotides) corresponding to about 25 000 annotated genes of the human genome. Genes are annotated based on the content

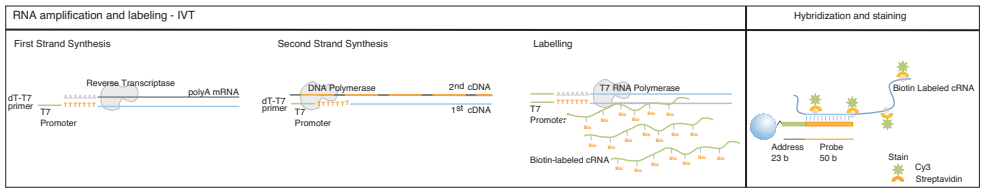


Figure 12. Illumina BeadArrays for gene expression profiling. Gene expression profiling using Illumina microarrays is initiated by linear amplification of total RNA resulting in biotin-labelled cDNA. The cDNA is hybridized to a BeadArray where oligo coated beads are randomly distributed onto a silica surface. The bead oligos contain information on which specific bead is placed on a particular spot on the array (address code) and a gene specific sequence (probe). Hybridized cDNA is stain with fluorescent Cy3 streptavidin and emitted fluorescence due to scanning is registered (probe), decoded (address) and further processed.

of National Center for Biotechnology Information (NCBI) RefSeq Release 38 and each probe-bead pair is replicated about 30 times per array. To retain gene expression data, the first steps include in vitro transcription (IVT) linear amplification of total RNA that results in biotin-labeled cRNA. The cRNA is hybridized to the BeadArray, stained with Cy3 streptavidin (a green fluorescent dye), washed and scanned. Emitted fluorescence is registered as a representative measurement of gene expression and subjected to further data pre-processing.

Methylation microarrays

Study V describes the epigenetic landscape of lung cancer based on DNA methylation patterns. To be able to discriminate between a methylated C (C_m) and an unmethylated C, a bisulfite conversion was initially performed. In this process, C is converted to U while C_m remains intact. In Study V, the 450K Infinium HD Methylation Assay (Illumina[134]) was used to investigate if subgroups are present in lung cancer based on methylation patterns. The platform covers information on about 480 000 CpGs distributed over the genome, and its accuracy and reproducibility has been proven to be stable[135, 136]. The Infinium Assay consists of two different probe categories (I and II) that deploy different chemistries in the extension and staining process. Infinium I probes were developed to be used in the first methylation BeadArray (27K) supplied by Illumina and the assay is based on a one-channel approach, using allele-specific primer extension of the pre-processed and hybridized DNA. Infinium II probes uses a single base extension dual fluorescence chemistry (Figure 13). The major difference between these two

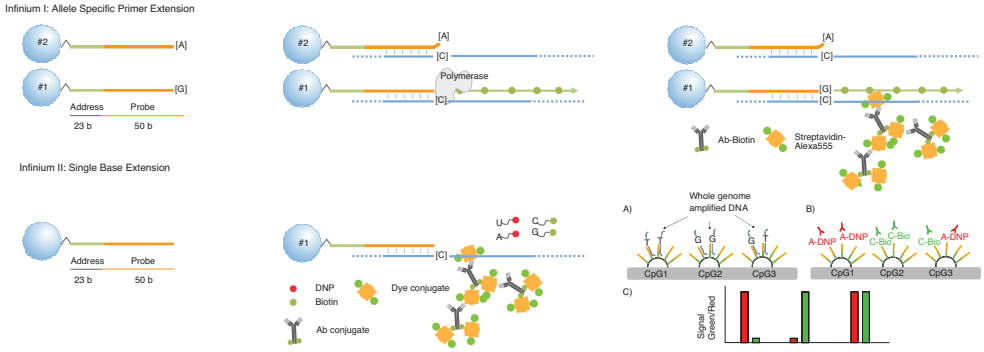


Figure 13. Infinium I and II probe chemistry. Illumina 450K Methylation BeadArray is based on the Infinium HD technology. The platform deploys two types of chemistry during the extension and staining process.

techniques is the fact that Infinium I probes indicates whether there is signal from bead 1 or bead 2, i.e. if the tumor CpG that these beads represents is methylated or unmethylated. With the Infinium II probes, information not only includes if the CpG is methylated or unmethylated but also if there is a no-change in methylation status (equal fluorescence from the two different dyes). Infinium I probes have been proven to produce more stable results, and when dealing with methylation data derived from this platform, the two different probe chemistries can be considered as two different data batches that must be corrected for[137]. Bisulfite conversion was performed in 96-well format (plates) balanced for biological and etiological sample properties. To exclude any potential batch effect, probes were adjusted between plates and principal component analysis (PCA) was performed to verify that no technical artifacts caused systematic bias in the final data[138]. The data generated were used to investigate multiple biological aspects including: 1) methylation patterns 2) copy number alterations and 3) gene expression profiles.

High-dimensional data processing

Pre-processing of data

Reasons for gene expression/methylation data variation can be biological or technical. Pre-processing of data is essential to minimize gene expression variations that are due to technical issues such as staining efficiency, batch effects but also sample related issues such as RNA/DNA quality or input. All samples need to be adjusted to a common baseline to be compared to other samples within or between experiments. This process is referred to as normalization. A variety of normalization methods exist, and new methods are constantly being developed. To generalize the methods, two main categories exist: 1) methods assuming expression doesn't change radically between samples and adjusts overall expression based on this assumption or 2) adjusts overall expression based on co-

hybridized references or spiked-in housekeeping genes. Normalization of gene expression/methylation data in studies included in this thesis was performed using an algorithm for cubic spline quantile normalization[139], proven to be robust for this microarray platform[140], and it belongs to above described category 1. Technical varieties were handled through background correction based on emitted fluorescence from negative control probes available in the BeadArray probe set by Illumina. This technical background is subtracted from the remaining beads in order to be able to assume that the signal emitted from these probes is due to biological changes and not technical issues. Outlier beads were removed, average bead signals and detection p-values were calculated. Annotated, pre-processed data was exported and further processed using the R statistical programming language[141]. Data was $\log_2$ transformed and selected based on their detection p-value. Detection p-value means the probability of a transcript being expressed above background. Probes with high confidence (detection p-value<0.01) were kept if present in more than 80% of the samples. Poorly annotated probes or signals due to cross-hybridizing were filtered out and the data were reannotated[142]. In order to deal with the biological fact, which can affect data interpretation and skewness, that some genes are highly expressed than others, data was scaled using standard deviations >0.3 across all samples. Further processing includes transformation of data using mean centering across all samples. After performing all above steps, the data set derived can be used to make biological interpretations. Methylation data is generated as beta-values for each CpG probe and range from 0 to 1, where 0 represent unmethylated events and 1 represent methylated events. To establish aberrant methylation patterns in tumors, the included normal tissue was used as reference. Based on beta-values hyper- or hypomethylation in tumors was compared with normal tissue, and 4136 CpGs were selected as the most varying CpGs if comparing tumor and normal tissue methylation patterns.

Clustering

Clustering is a broad definition of a large number of different algorithms and is a useful approach for unsupervised analysis (where no prior information is traditionally used to guide the analysis) of high-dimensional data such as gene expression or genome-wide methylation data. Briefly, clustering aggregates samples that are most similar to each other into clusters, forming a hierarchical tree. Similarity between samples may be defined in many different ways, using different distance metrics and methods to link clusters (linkage methods). In Study III, a consensus clustering[143] , with Pearson correlation was used as the distance metric and Ward linkage for linking clusters to each other. In this unsupervised analysis, gene expression patterns based on intrinsic features with no external information are used to investigate how many groups were present and with what confidence these groups were present. Samples are subgrouped and clustered in an iterative manner to investigate whether there is a consensus over iterations. The

consensus determines the number of clusters and assesses the stability of the observed clusters. Once clusters have been identified, differentially expressed genes between subgroups may be identified through supervised analysis. Supervised analysis, in contrast to unsupervised analysis, aims to identify differences between a priori defined groups of samples by using various statistical methods. Clusters produced in Study III were used as centroids for validation. In Study V, bootstrap clustering was used, a method similar to consensus clustering where resampling of the data is used to verify stable clusters. Bootstrap clustering identified five epitypes in Study V, which were used as centroids in the validation process (classification).

Copy number analysis retrieved from genome-wide methylation microarrays

Matched normal lung tissue specimens were available for 12 patients and were included in the tumor sample set in Study V. After normalization (performed as described for gene expression analysis), intensities from these 12 normal samples were used to create $\log_2$ copy number estimates from unmethylated and methylated CpG probe signals. Calculated mean signals for each sample probe were correlated with mean signal from the corresponding probe in the normal samples. Genomic profiles were generated using Gain and Loss of Analysis of DNA (GLAD)[144] and fixed thresholds were used to call copy number gain and loss.

Classification and validation of gene expression and methylation data

To reveal the underlying fact of why samples cluster together, and to assure the validity of the established clusters, gene expression clusters and epitypes were classified according to reported GEPs and scored according to metagene signatures. In Study III, 10 consensus clusters were identified and subgroups were used to classify an external dataset[39] by measuring the nearest distance of each sample in relation to the clusters (transformed into gene expression centroids). This was done to assure the biological validity in the derived clusters. In Study V, NSCLC tumors of all histological subtypes were included. Merely AC and SqCC tumors were classified according to reported GEPs available and described[47, 55] since no GEPs for the remaining histological subtypes (i.e. LCC/LCNEC/SCLC) were available. Gene Ontology (GO)[145] using PANTHER[146] was used to perform pathway analysis of differentially expressed genes in Study V. To perform integrative analysis of methylation and gene expression, correlation analysis was made for the largest histological subgroup (AC) using Spearman correlation.

Single Sample Prediction

Centroid classification is widely used and adapted but has reproducibility limitations. Centroids derived from stable clusters are highly platform, and sometimes even dataset, dependent. To overcome these limitations of

classification, rank-based classifiers have developed. A single-sample predictor (SSP) is platform-independent, robust to data normalization and yields transparent decision rules[147]. The concept of SSPs is to define certain decision rules based on annotated gene expression data (training set) and apply those rules to gene expression data lacking annotation. One of the most widely used SSP approaches is k-Top Scoring Pairs (kTSP)[148]. Here, the decision rule is entirely determined by the ordering of two features (i.e. the relative expression of two genes). Decision rules, or prediction classes, are based on ranking of the genes. In Study II, however, we aimed to create an SSP of NSCLC histology using three prediction classes. Since kTSP is designed to only handle two prediction classes, Absolute Intrinsic Molecular Subtyping (AIMS)[149] was chosen due to its capability to handle more than two classes of prediction. AIMS machine-learning method has performed well in other types of cancer (originally developed in breast cancer) and proven successful in prediction of multiple classes irrespective of normalization method and gene expression data generation platform.

Statistical methods

A variety of statistical methods have been used throughout this thesis work, depicted in Table 3.

Table 3. Statistical methods.
Various statistical methods used in respective study.

STATISTICAL METHOD	STUDY I	STUDY II	STUDY III	STUDY IV	STUDY V
Fishers' exact test	X				
SAM			X		
PCA					X
Kruskal-Wallis					X
ROC		X			
Kaplan-Meier			X	X	X

Statistical significance is based on rejecting or retaining the null hypothesis, the null hypothesis being no difference between two comparisons. In this hypothesis testing, a probability value (p-value) with a pre-determined significance level, typically the two-tailed 5% of sampling distribution, indicates statistical difference between two observations. Various methods for group comparisons have been used in this thesis including Fisher's exact test.

Significance Analysis of Microarrays (SAM) is a method that uses multiple t-tests to identify differentially expressed genes by statistical significance.

Principle Component Analysis (PCA) performs dimension reduction by finding components that maximizes the variation. PCA has been used to assure consistency in pooling of gene expression data using replicates included in the gene expression data generation[138].

The Kruskal-Wallis test is a non-parametric rank-based test used to investigate whether two samples originates from the same distribution. Since it is a non-parametric test, there is no assumption that the data set is normally distributed. Kruskal-Wallis was used to identify differentially expressed genes between epitypes in Study V.

Receiver operating characteristics (ROC) was calculated using the predicted classification and histological classification made by a pathologist in a confusion matrix. ROC investigates the performance by comparing observed positive and negative prediction conditions and taking false event into account. Predictions are plotted in a ROC curve. Accuracy is measured by the area under the ROC curve (AUC). An area of 1.0 means perfect discriminating power and 0.5 is the performance by pure chance [150]. Accuracy value is calculated as the sum of true positive predictions plus the sum of all true negative predictions divided by the sum of all predictions. A balanced accuracy value considers the false negative/positive predictions.

Survival analysis was performed using Kaplan-Meier estimates, a method that censors observations and calculates the risk for the remaining population. To test if the observed differences between groups are likely to happen by chance, a log-rank test is usually used. Kaplan-Meier curves were compared using the log rank test and Hazard ratios were calculated through univariate Cox regression.

Results and Discussion

This thesis is based on five studies designed and executed with the intent to molecularly classify and stratify NSCLC tumors with focus on potential or direct clinical implications. The order of the studies is not presented in a chronological order. Instead, the studies are presented according to patient category size and clinical impact through three main sections: 1) High-dimensional data generation and processing, 2) Molecular implications, and 3) Clinical implications. Study I focuses on treatment predictive testing applicable to the largest groups of lung cancer patients, patients with late-stage, advanced disease. The study describes a direct application of novel technologies in clinical diagnostics to aid clinical management. Study II describes a new approach to simultaneously predict tumor histology and fusion gene status, by using a multicomponent RNA assay, applicable to clinical tissue. A multicomponent tool is highly desirable, especially in cases with advanced disease, since diagnosis is based mainly on biopsy or cytology specimens that usually provide very small amounts of tissue. Therefore, the approach in Study II has potential clinical value and the multicomponent tool was developed in early disease and subsequently validated in both early and advanced disease. Moreover, Study II represents an example of how an assay could be used for different applications, and also as a platform for future, additional, applications to be integrated. In Study III, the newly adopted 2015 WHO guidelines on classification of a relatively small histological subgroup (LCC), representing an important differential diagnosis, was investigated to validate whether this update was also reflected in the transcriptional landscape. The results in Study III support the WHO guidelines, used in clinical management and therapy selection, to further stratify a specific NSCLC subgroup. This conclusion was further supported by Study IV, a study that was initiated before the 2015 WHO guideline revision but finalized in time for the revised guidelines to be adopted. Findings in Study IV support the revised guidelines through an extensive genetic characterization of LCC and LCNEC tumors. Since LCC is a differential diagnosis made on the basis of resected tumor material, Studies III and IV are based on low-stage tumors from early disease patients. In Study V, resected NSCLC tumors were epigenetically profiled resulting in DNA methylation patterns that stratified tumors into epitypes correlated with histology, copy number alterations, gene expression patterns, and patient survival. Study V demonstrates that epigenetic characterization adds another genomic layer that contribute to a

deeper understanding of NSCLC tumor development. This deepened understanding, in combination with other genomic data, e.g. gene expression, may eventually be translated into clinical management in the future through refined patient stratification. Figure 14 visualizes the structure and implications of this thesis work.

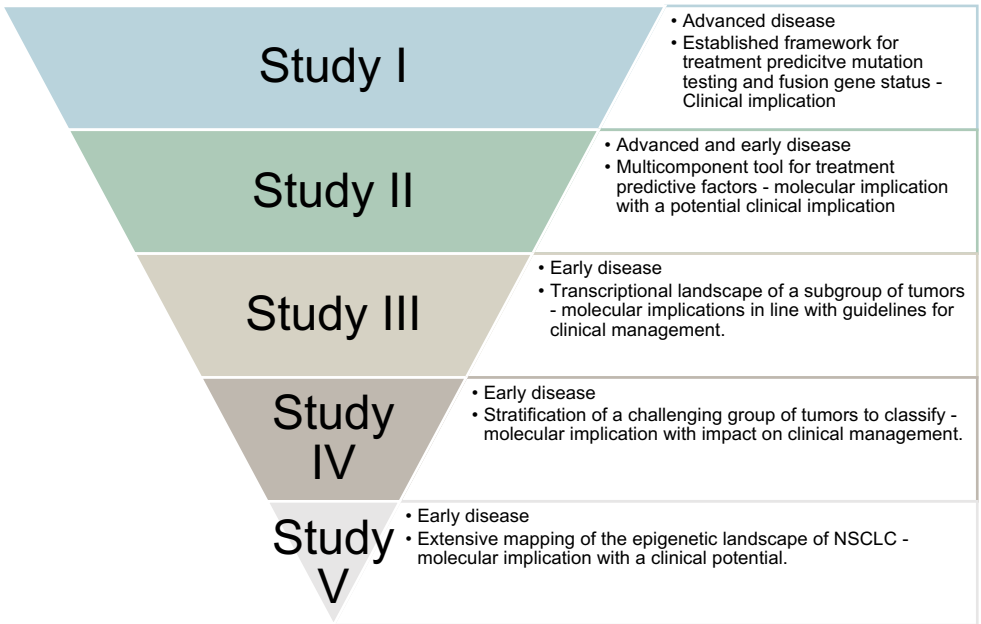


Figure 14. Visualization of thesis structure and implications of included studies.

High-dimensional data generation and processing

Technology development

Microarrays and NGS are two high-dimensional techniques used with a variety of intents in research. These technologies share many similarities but also have many differences. The two techniques are able to both investigate a high number of targets within a sample and processing multiple samples simultaneously, i.e. high-dimensional and high-throughput. With the introduction of microarrays to the field of cancer research in the late 1990s the possibility to study vast landscapes of copy number variation and transcriptional patterns opened up. This resulted in the

discovery of distinct phenotypes correlated with a variety of tumor properties, for instance, in breast cancer[151]. These phenotypes showed prognostic and therapeutic value and contributed to deeper biological understanding of the disease development. The early days of microarrays were technically challenging, though, and bioinformatic pipelines did not exist to handle the amount of data generated. The first generation of non-commercial (often in-house produced) microarrays required lots of hands-on time and a large amount of intact RNA/DNA. Due to these requirements, FFPE tumor tissue or scarce amounts of fresh frozen tumor tissue were almost never an option. Therefore, the lung cancer genomes investigated with early-days' microarrays were biased to: 1) large size tumors to fulfill the requirement of input DNA/RNA, and 2) surgically resected tumors, i.e. only low stage tumors. The resolution of the non-commercial microarrays was, for the time being, high and foremost genome-wide in coverage. With the introduction of commercially produced microarrays, the former instabilities associated with the production and handling of microarrays were reduced, and stability and reproducibility were greatly improved. The resolution increased as well as the throughput, and the field of applicability expanded to study, for instance, miRNA expression and epigenetic alterations such as DNA methylation and chromatin immunoprecipitation (chIP). To some extent, microarrays were also adapted to handle degraded DNA/RNA. Instead, the microarrays were limited with respect to resolution, often associated with technical difficulties, and required large amounts of input DNA/RNA. In this thesis, microarrays were used as the primary data generation platform in Studies III and V using fresh frozen tumor tissue from low-stage, surgically-treated patients, providing extracted nucleic acids of high quality suitable for microarrays (gene expression and DNA methylation). Today, the gene expression profiling in Studies III and V would likely have been performed through RNA sequencing, while the same type of DNA methylation analysis would have been performed with a greater number of features. In Studies I and II, due to the specific limitations of archival tissue, an alternative platform, NanoString, was used. While the NanoString method shares features with microarrays, e.g. hybridization and fluorescent detection, it represents a simpler platform applicable to the specific aims of Studies I and II. Currently, NanoString is a widely used platform for these types of analyses.

Early-days' of NGS also demanded high quality extracted nucleic acids, and protocols for library preparation were laboratory tedious, hands-on demanding, and very expensive. As technology developed, protocols for library preparation were simplified, sequencing instruments were adjusted to bench top versions, and costs dropped. Soon, even nucleic acids extracted from archival tissue could be used as template for library preparation. One of the original aims of Study IV was actually to evaluate a specific platform and protocol for mutation detection in archival NSCLC tumor tissue. At that time, an NGS-based platform for mutation

detection applicable to DNA extracted from archival, FFPE tissue had recently been developed by Illumina[134]. The Illumina TruSight Tumor 26 gene panel for targeted NGS was designed to overcome many of the issues when dealing with DNA extracted from FFPE. Particularly, the fixation process affects the DNA by crosslinking, which reduces PCR efficiency and fragments the DNA during the extraction. Also, fixation induces sequence alterations (most frequently C-T transversions) with as high rates as 1/500 bp[152]. To overcome these obstacles, design of probe sequences to be captured in the DNA of interest need to be adjusted to a length that is represented in FFPE DNA. A specific feature of the TruSight Tumor panel was that probe sequences corresponding to the target of interest were designed for both DNA strands, creating a bi-directional assay. DNA was separately prepared as two independent libraries for each patient, sequenced, and later pooled in the analysis. This allowed sequence alterations present in only both pools to be identified and considered true mutations, effectively removing artifacts caused by fixation. In Studies I and IV we found this feature to be a key component for the interpretation of the data, scientifically in Study IV, but more importantly in the clinical analyses described in Study I. Specifically, the bi-directional nature of the panel greatly reduced the interpretation time (filtering and classification) of identified variants, and also provided greater confidence in the results. While NGS panels for diagnostic use are continuously expanding in size and likely also changing to be become hybridization based, the bi-directional feature still represents a powerful way for amplicon-based panels to reduce artifacts and thus potential false positive calls through a straightforward analysis pipeline, despite the requirement of more input DNA.

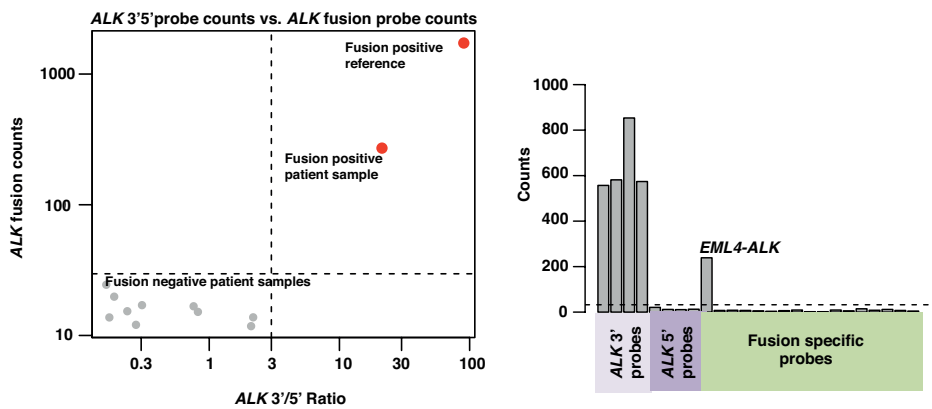


Figure 15. Fusion gene detection using the NanoString technology. Fusion gene detection is employed by calculating the ratio of expression of the 3' and 5' end of the *ALK* gene as originally described by Lira et al.[126]. Specific fusion probes identify the fusion gene partner of *ALK*.

In parallel to treatment predictive mutation testing performed in Study I, the NanoString technology to detect gene fusions was evaluated (Figure 15). Probe sets were designed to detect specific fusions but also novel fusions by calculation of 3'/5' gene expression ratio in genes frequently rearranged in NSCLC. The technology proved accurate in detecting gene fusions as all detected *ALK* and *ROS1* fusions were confirmed by the routine diagnostics (IHC and FISH). Using the ratio calculation, with overexpression of probes corresponding to the 3' end of the *RET* gene compared with low expression of the 5' end of the *RET* gene, a novel *TRIM24-RET* fusion (validated with targeted RNAseq, Archer Dx) was detected. Since the NanoString technology proved accurate and applicable to scarce amounts of archival tissue (the typical situation in analysis of advanced disease), the probe set was expanded in Study II. The updated set included probes corresponding to genes used as protein markers in routine diagnostics to distinguish histological subgroups as well as immunomarkers. In Study II we sought to investigate whether the NanoString technology could be used as a multicomponent assay for solving two clinically relevant features: gene fusion status and histological subtyping.

Molecular implications

The molecular implications of the studies in this thesis work ranges from mapping treatment predictive mutations and gene fusions in NSCLC advanced disease to comprehensive descriptions of the transcriptional and epigenetic landscape of lung cancer.

Mutations, gene fusions and implementing molecular tools

Classification of lung cancer is crucial for therapy guidance. Whether it is by histology or mutation/rearrangement of targetable genes, stratifying lung cancer is beneficial for choice of treatment and as a prognostic factor. In Study I, clinically relevant mutations were reported and used to guide treatment. An entire framework for NGS-based analysis of treatment predictive mutations was built to handle everything included in clinical mutation testing of NSCLC. At the time of Study I only *EGFR* mutations were considered treatment predictive, although all findings were reported back to the diagnostic pathologist. However, building the framework for clinical treatment predictive mutation testing offered a splendid opportunity to map the mutation spectrum of 26 classical oncogenes and tumor suppressors in advanced NSCLC in a south Swedish clinical testing population. As depicted in Figure 16, the mutation spectrum was summarized for the entire cohort

as an entity but also stratified by histology. In the 533 samples included, 889 variants were detected with predominantly 1-2 variants per sample reported reflecting the pan-cancer nature of the TruSight Tumor 26 panel.

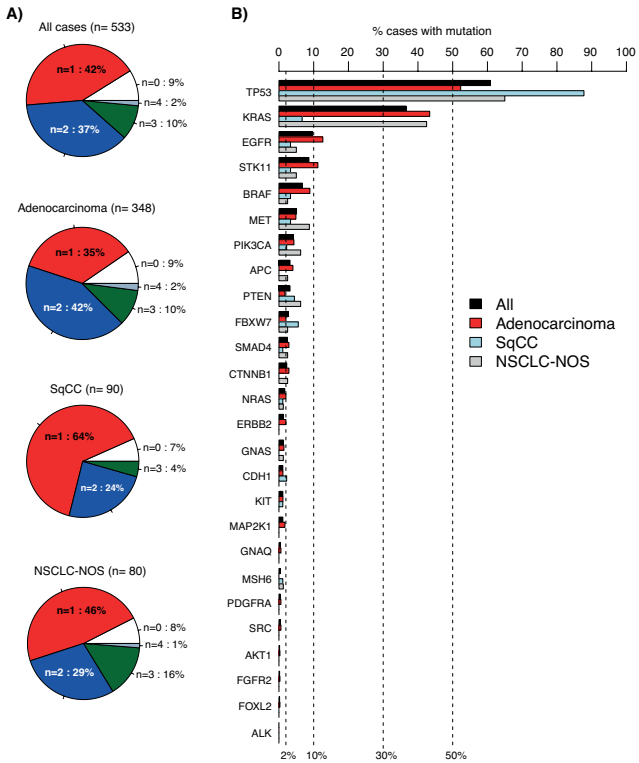


Figure 16. Detected variants in 533 consecutive lung cancers analyzed by the 26-gene Illumina TruSight Tumor panel. (A) Pie charts of number of called variants per sample for different sample groups. (B) Variant frequency for the analyzed 26 genes across different sample groups (bars). Genes are ordered according to decreasing frequency in the total cohort. In A and B, all detected non-synonymous variants by the vendor supplied analysis pipeline are included.

In parallel, we aimed to develop an assay to investigate gene fusion status that was quick, reliable, applicable to archival tissue, and possible to include in the NGS framework from Study I. In the framework established, both DNA for gene mutation analysis as well as RNA were extracted. In order to evaluate the feasibility of an RNA-based approach to deliver gene fusion status based on archival tissue, we selected 169 patients screened negative for treatment predictive mutations in *EGFR*, *KRAS* or *BRAF* during 2015. The NanoString technology fulfilled our requirements in terms of the ability to robustly detect gene fusions in RNA derived from archival tissue quickly while still being economically

affordable. The probe set design was based on a reported approach by Lira et al.[126] including the possibility to detect known fusions as well as a possibility to find novel fusions in *ALK*, *RET* and *ROS1*. Today, gene fusions involving the *ALK* or *ROS1* genes are targetable, and to a minor extent also *RET*. As a result of this, only *ALK* and *ROS1* fusions detected by NanoString could be verified through clinical routine analyses. For the non-adenocarcinoma cases that were ALK positive by IHC in the triple-negative cohort, NanoString analysis suggested overexpression of the entire gene by some other mechanism than gene rearrangement. This more detailed view of gene fusion events supports the usage of multiplexed methods like NanoString as a complementary method, or even replacement, for IHC/FISH when possible. Due to the flexibility and capacity of the NanoString technology, additional gene fusions as well as *MET* exon 14 skipping events and genes corresponding to IHC markers used in daily routine pathological assessment of histology were included in an updated probe set used in Study II.

In Study II we investigated the possibility of developing a multicomponent tool for simultaneous gene fusion detection and histological assessment based on RNA expression. The latter included developing a predictor of NSCLC histology from a mixed training cohort (n=68) of never-smokers and additional tumors from other histological subgroups analyzed by the updated NanoString probe set (SMIL in Figure 7). An SSP classifier was developed through machine learning in a NanoString-derived gene expression data set using 11 genes corresponding to protein markers used in routine diagnostics to assess NSCLC histology. Simultaneous gene fusion status was retrieved and the fact that five gene fusions and two *MET* exon 14 skipping events were detected can be ascribed to the composition of the SSP training cohort, as these rearrangements are associated with AC histology and a never-smoking history. After an initial feasibility test, a final SSP model was validated in three external cohorts. Two of the cohorts composed of NanoString derived gene expression data and the third composed of gene expression data generated by RNA sequencing (NGS). The SSP could successfully stratify tumors originally classified as LCC by WHO2004 guidelines into the revised WHO2015 guidelines as AC and SqCC. One of the validation cohorts was composed of tumors classified as NSCLC-NOS in Study I. These tumors were subjected to an in-depth pathological re-analysis including complementary IHC stains and morphology assessment.

Discrepancy between SSP prediction and pathology assessment for these NOS tumors included pathology classification performed on the basis of markers, such as mucin markers, not included in the probe set or metaplasia of cells of differing histology than the tumor (Figure 17).

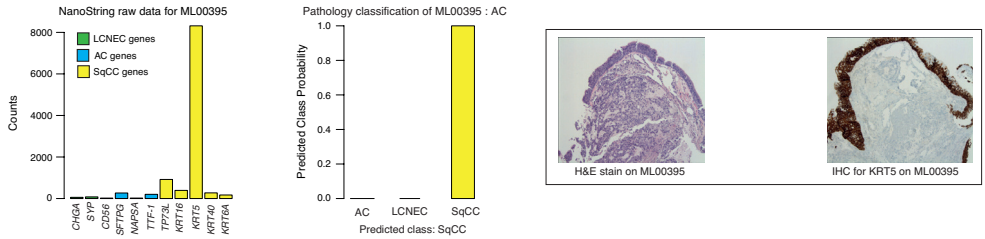


Figure 17. Discrepancy between SSP histology prediction and pathology assessment. Due to high *KRT5* gene expression, the SSP classifies this sample as SqCC with high probability. Evident from H&E stains and IHC fort KRT5, a SqCC metaplasia of this AC tumor causes high expression of the SqCC marker KRT5.

Despite the fact that the SSP was developed in NanoString gene expression data, the SSP could successfully predict NSCLC histology in a large (n=199) gene expression data set of early stage tumors generated using RNA sequencing, thus demonstrating platform independence. The requirements we set for the SSP are dramatically different from classification using centroids as described in Study III and V. Centroid classification is highly dependent on gene expression/methylation data to be pre-processed and centered across samples (i.e. relative expression instead of absolute). Study III and V instead provide comprehensive knowledge on global gene expression and methylation patterns based on unsupervised analysis, revealing underlying biological mechanisms in contrast to mapping specific mutations or gene fusions.

Molecular subtypes in resected, low-stage lung cancer

To comprehensively investigate the molecular characteristics of lung cancer (including all major histological groups), two genome-wide studies were performed in this thesis. Study III investigates the transcriptional landscape, while Study V describes the epigenetic landscape of lung cancer. In addition, Studies III and IV also analyze the genomic characteristics of a specific subset of lung cancer, LCC and LCNEC tumors, in more detail.

Since histology is an important factor for both clinical management and prognosis, Studies III and IV focused on a group of tumors, LCC, that up until the revision of the WHO guidelines in 2015 had been used as a somewhat “trash can” group of

NSCLC tumors. In Study IV we aimed to characterize LCC and LCNEC tumors using the same platform for mutation detection as in Study I and also to investigate whether gene fusions occurred in these tumors through RNA sequencing. Gene mutation and fusion analyses are part of the routine diagnostic framework and we aimed for screening this small, but important, histological subgroup using techniques that could be implemented in clinical diagnostics. At the time that Study IV was finalized, WHO revised their guidelines in terms of histology subgrouping, putting more emphasis on IHC analyses to define histology. Importantly, Studies III and IV were both performed in the context of the revised guidelines, and the main findings supported the revised guidelines on both the genetic and transcriptional level.

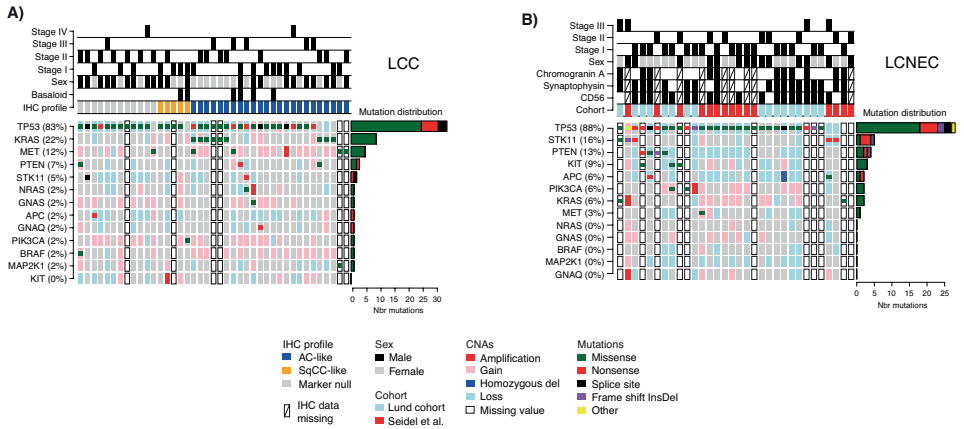


Figure 18: Detected mutations and copy number alterations in LCC and LCNEC. A) Detected gene variants and copy number alterations (CNAs) (rows) in 41 LCC cases (columns), ordered by immunomarker profile of adenocarcinoma-like (AC-like), squamous cell carcinoma like (SqCC-like), or marker null phenotype (TTF-1/Napsin A and CK5/P40 negative). Copy number status is shown as larger background rectangles and mutations as squares for each sample and gene. Right side bar plot summarizes the distribution of the different mutation types for each gene. B) Detected variants and copy number alterations in 32 LCNEC cases displayed as in A). Samples are ordered according to gene variant frequency.

Specifically, in Studies III and IV we observed: 1) a similarity in mutational (e.g. *TP53* and *RB1* inactivating mutations) and transcriptional patterns in LCNEC and SCLC tumors, 2) general absence of prototypical AC and SqCC oncogenes alterations in LCC tumors defined according to WHO2015, and 3) a transcriptional pattern of LCC tumors separating them from the other histological subgroups (see below). In Study IV, lung cancer cases (n=33) classified as LCC according to the WHO 2004 guidelines were reclassified on the basis of the WHO 2015 guidelines with 70% as variants of AC or SqCC. Specifically, 19 cases (58%) were reclassified as AC on the basis of positive expression of TTF1/Napsin

A, four (12%) were reclassified as SqCC on the basis of positive expression of CK5/ P40, and 10 (30%) did not express any of these IHC markers ("marker-null"). These results in Study III are supported by the mutational and copy number variation differences seen in Study IV, distinguishing LCC from LCNEC and stratifying the LCC group into AC-like, SqCC-like and marker null phenotypes (Figure 18). To investigate whether the WHO 2015 guidelines translated into a better transcriptional subgrouping of LCC in Study III, we performed unsupervised consensus clustering of a discovery cohort comprising 159 lung cancers of all histological subtypes. We first performed iterative consensus clustering without respect to sample annotations by using variable number of genes to assess the optimum cluster solution. Next, we performed an in-depth comparison of unsupervised transcriptional subgroups with sample molecular and clinicopathological data.

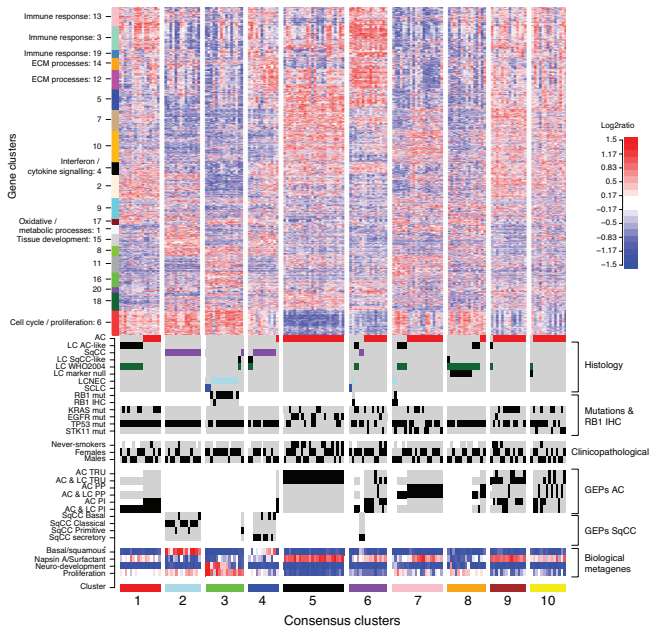


Figure 19. Unsupervised gene expression analysis stratifies large cell lung cancer (LCC) and large cell neuroendocrine carcinoma (LCNEC) into molecular subgroups. Gene expression heatmap of 2730 Illumina probes across 159 lung cancers stratified by 10 specified consensus clusters. The 2730 probes correspond to a $\log^2$ratio standard deviation cutoff of more than 0.5. Annotations for histological subtypes, clinicopathological variables, selected mutations, retinoblastoma 1 immunohistochemistry (RB1 IHC), classification according to reported gene expression phenotypes (GEPs) for adenocarcinoma (AC) and squamous cell carcinoma (SqCC), and expression of selected biological metagenes are provided. For annotations, black corresponds to a positive/presence call, gray to a negative call, and white to not applicable or not available. Gene cluster functional annotations are provided for some specific clusters in the heatmap. ECM, extracellular matrix; mut, mutation; RB1 mut, retinoblastoma 1 gene mutation; TP53, tumor protein p53 gene; STK11, serine/threonine kinase 11 gene; TRU, terminal respiratory unit; PP, proximal proliferative; and PI, proximal inflammatory.

Acknowledging that the histological subtypes of lung cancer strongly influence the transcriptional landscape[153] and that subgroups within the histological subtypes likely exist, we chose a 10-group consensus cluster solution to be able to also study characteristics for minor subgroups. Consistent with previous studies[48, 153], we observed a clear separation of AC, SqCC, and SCLC cases into subclusters driven by specific transcriptional programs (Figure 19). In agreement with recent studies and in agreement with mutational patterns in Study IV, LCNEC tumors clustered strongly (79% of cases) with SCLC tumors, forming a neuroendocrine subcluster (Figure 19). For LCC, 84% of the WHO 2004 cases reclassified as adenocarcinoma-like clustered in an adenocarcinoma-dominated subcluster, whereas 50% of the LCC SqCC-like cases clustered in an SqCC-dominated subcluster (Figure 19). Notably, 90% of marker-null cases (nine of 10) aggregated in a separate transcriptional cluster (see Figure 19 [cluster 8]), referred to as the marker-null–enriched subtype. To investigate the reproducibility of our findings in Study III, we created gene expression centroids for each consensus cluster and classified an independent validation cohort of 199 tumors comprising all histological subtypes by a nearest centroid approach. Three of five LCNEC and four of six LCC marker-null cases were classified into the LCNEC and LCC marker-null clusters, respectively. The two LCNEC cases not in the neuroendocrine cluster did not express high mRNA levels of neuroendocrine marker genes. These two discrepant LCNEC cases were further investigated in Study II where it became evident that these cases were of mixed histology, containing tumors cells of both AC and LCNEC features. RNA extraction had been performed solely on the AC component, explaining the lack of expression of neuroendocrine markers associated with LCNEC histology. For the two outlier LCC marker-null cases, one case was found in predicted cluster 10, whereas the second was found in predicted cluster 3 (the neuroendocrine cluster) despite not showing increased prototypical neuroendocrine gene expression.

On the basis of massive parallel sequencing studies, it is becoming evident that a subset of LCNEC tumors share mutational patterns with SCLC, whereas others carry mutations typically altered in non-neuroendocrine tumors[153-156]. Rekhtman et al.[154] hypothesized a genetic division of LCNECs into SCLC-like and NSCLC-like subgroups based on *TP53* and *RB1* alterations, in which the SCLC-like group was defined by concomitant *TP53* and *RB1* alterations, reflecting their ubiquitous inactivation in SCLC[64]. In contrast, the NSCLC-like subset was characterized by the lack of concomitant *TP53* and *RB1* alterations and occurrence of other NSCLC-type mutations (e.g., *KRAS* and *STK11* mutations). Interestingly, Study III indicates that this stratification may potentially be mimicked also on the transcriptional level, providing a speculative link between the mutational and transcriptional landscape of LCNEC. While our analyses in Studies III and IV provide some support for the hypothesis that LCNEC tumors may be refined into

NSCLC-like/SCLC-like tumors, larger studies are needed to confirm such subgroups, and importantly their clinical relevance.

Within both AC and SqCC different gene expression phenotypes have been proposed[47, 48, 55-58], although the consensus between phenotypes is not absolute in independent multicohort analysis[157]. In Study V, global methylation analysis of 124 lung cancers revealed four distinct AC epitypes and one neuroendocrine epitype, supporting further stratification of lung cancer and lung cancer histological subtypes. Using genome-wide methylation microarrays, we identified 4136 CpGs with aberrant methylation in >10% (n = 13) of tumors compared with normal lung tissue. Unsupervised bootstrap clustering identified five distinct clusters, referred to as epitypes (Figure 20).

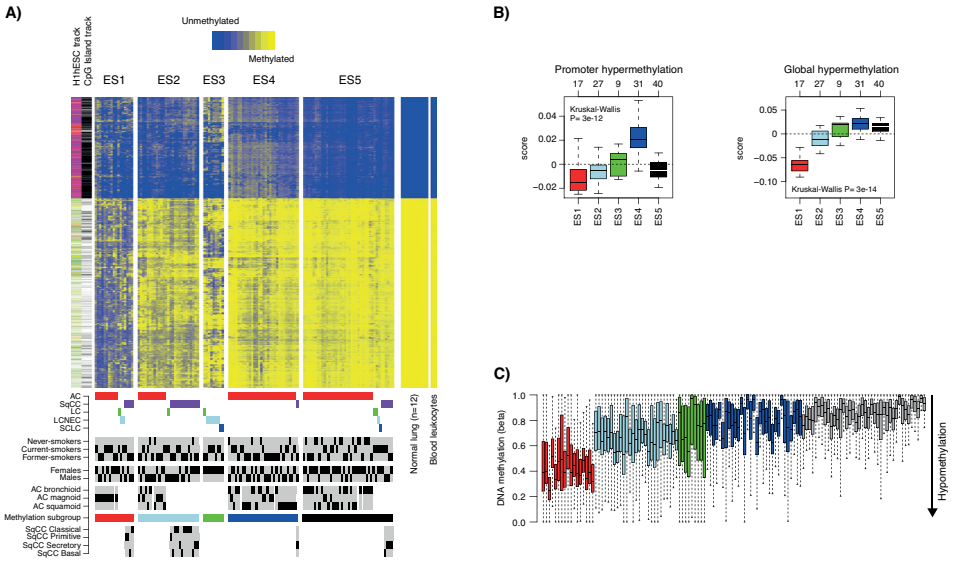


Figure 20. Identification of five DNA methylation subtypes. A) DNA methylation subtypes in 124 lung cancers based on bootstrap clustering of 4,136 variant CpGs. Heatmap displays beta values (rows) from unmethylated (blue) to methylated (yellow) for three sample groups (columns): 124 tumors divided into five subtypes by bootstrap clustering, 12 matched normal lung tissues, and blood leukocytes, with associated clinical characteristics and reported adenocarcinoma (AC) and SqCC gene expression phenotypes[47, 55]. Left hand CpG tracks, CpG island track; black, island; gray, shore/shelf; white, open sea, H1hESC track (ref.[158]; embryonic stem cell chromatin state): purple, poised promoter; red, active promoter; yellow, enhancer; green, transcribed; blue, insulator; white, heterochromatin. Sample annotations: black, yes; gray = no. B) global promoter hypermethylation (left) and global hypermethylation (right) score for methylation clusters (based on all filtered CpGs on the platform). C) box plots of DNA methylation for 629 CpGs matching repetitive elements from the set of 4,136 for each tumor in the discovery cohort across epitypes. Tumors are colored according to epitype as in A, with exception for ES5 (gray).

To validate the identified epitypes from the discovery cohort, we created DNA methylation centroids for each epitype based on the 4136 CpGs. Next, we classified two independent cohorts analyzed by the same methylation platform[42, 50] comprising 122 SqCC tumors and 695 adenocarcinomas. PCA performed in the

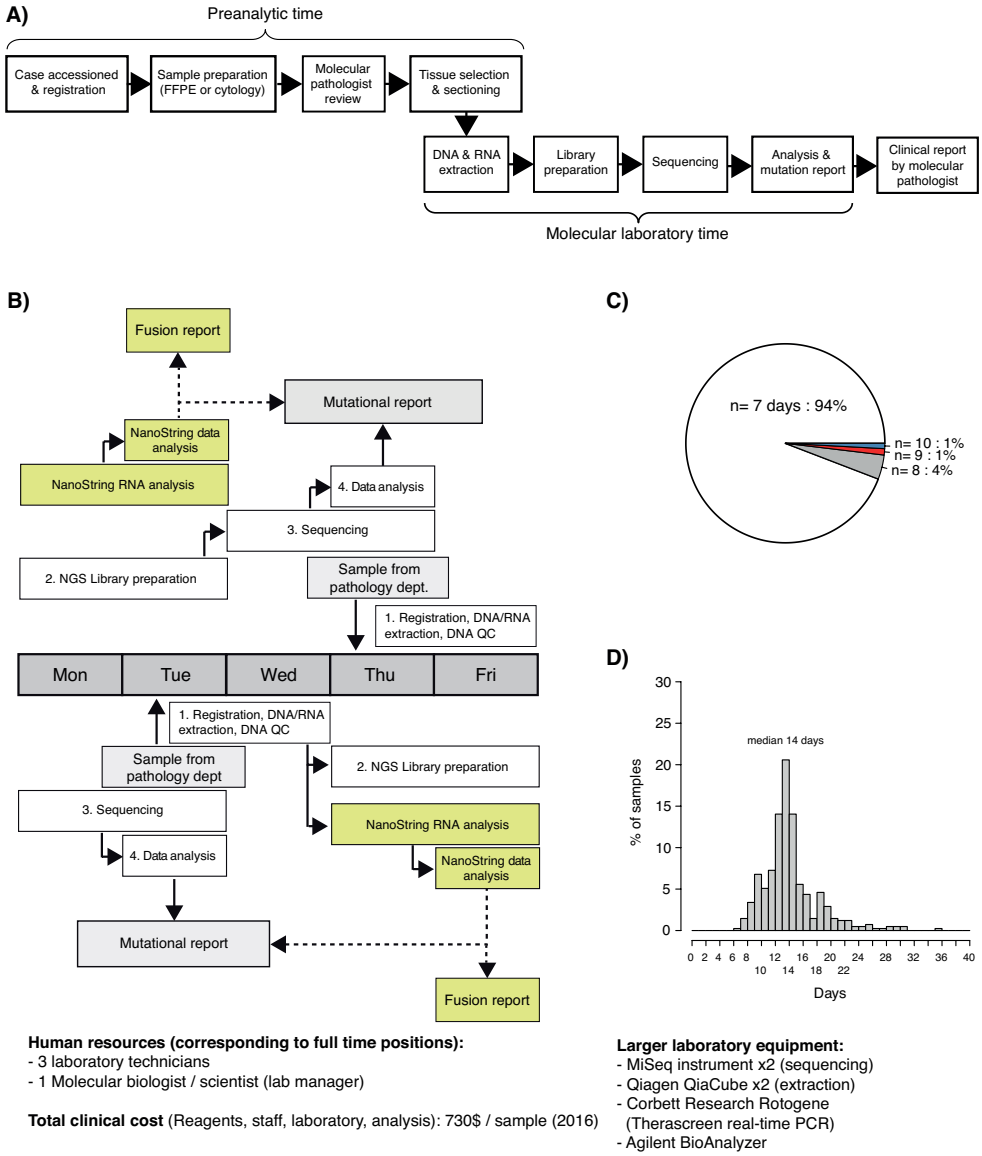
validation cohorts confirmed that the centroid classification explained most of the total variation in DNA methylation compared with available clinicopathological, technical (batch and beadchip data), and molecular factors, including clinical smoking history, sex, tumor stage, tumor size, histology (AC or SqCC), *EGFR*, *KRAS*, and *TP53* mutations. Notably, most of these factors (e.g. smoking status) contributed little to the total variation in DNA methylation. Moreover, the classification of the validation cohorts was robust across different sets of CpGs, and overlapped extensively with independently derived unsupervised bootstrap groups in these cohorts. In both validation cohorts, 1% of the cases were classified as the neuroendocrine epitype (ES3), supporting the theory that this epitype is highly distinct for lung cancers expressing neuroendocrine marker genes. The derived and validated epitypes were correlated with global hyper/hypomethylation, specific gene expression patterns, histopathological features, smoking status and, for AC tumors, correlated with survival. Although resected stage I NSCLC patients have the most favorable prognosis, the 5-year survival rate is 52% to 89%[159]. Thus, improved molecular subclassification and stratification of early-stage NSCLC remains highly relevant.

Clinical implications

The clinical implications of this thesis work extend from establishment of an NGS-based framework for clinical treatment predictive mutation testing to potential selection of patient groups for personalized medicine based on refined stratification of tumor biology.

Classification in advanced NSCLC disease

In Study I, the established NGS-based framework had a direct impact on treatment guidance of NSCLC patients diagnosed with advanced disease in the south Swedish health care region during January 2015 to June 2016. During this time approximately 1200 patients with suspected lung cancer or malignant melanoma were analyzed. Analysis included nucleic acids extraction, library preparation, sequencing, data processing and clinical reporting (Figure 21). With a few exceptions, turnaround time (TAT) were seven calendar days (i.e. five working days). Of all originally referred suspected lung cancer cases, 4.7% could not be analyzed during the first pass through the centralized NGS laboratory due to insufficient DNA quality in the qPCR quality control step. The latter was caused by either degraded DNA, or more often by insufficient amounts of extracted DNA from the FFPE sections sent for analysis (mainly small biopsy specimens).



Human resources (corresponding to full time positions):
- 3 laboratory technicians
- 1 Molecular biologist / scientist (lab manager)

Total clinical cost (Reagents, staff, laboratory, analysis): 730$ / sample (2016)

Larger laboratory equipment:
- MiSeq instrument x2 (sequencing)
- Qiagen QiaCube x2 (extraction)
- Corbett Research Rotogene (Therascreen real-time PCR)
- Agilent BioAnalyzer

Figure 21. Clinical implementation of an NGS-based diagnostic framework. (A) Workflow for the NGS- based mutational screening of treatment predictive alterations in lung cancer. Key turnaround time (TAT) metrics are displayed. (B) GANNT scheme of the workflow for the NGS mutational assay, with two parallel analyses running each week. Inclusion of a NanoString RNA based gene fusion assay is added to the NGS GANNT scheme, together with information about human resources, cost and larger laboratory equipment needed for the implementation. Reports from the NanoString assay can be issued at two different time points depending on whether a combined or separate reports are desired. (C) TAT for the molecular NGS part in calendar days for lung cancers analyzed during 2015 in the NGS central laboratory. (D) Total TAT including both the preanalytic and molecular laboratory time in calendar days for lung cancers analyzed during 2015.

However, with a few exceptions, all of these cases were analyzed either by resampling followed by new NGS-analysis, or by a real-time PCR method. During the period of time of Study I, mutation status of *EGFR* and *KRAS* (as a negative indicator of *ALK* fusions) were initially the only genes of interest for therapy guidance (in lung cancer) in the healthcare region (*ALK* alterations were analyzed by other techniques). Today, multiple new targets for therapy are emerging in NSCLC treatment, e.g. *BRAF* V600 alterations, demonstrating the importance of multigene analysis in clinical diagnostics today and even more so tomorrow. To investigate the potential additional clinical benefit of a combined NGS-based mutation analysis and multiplexed gene fusion assay (NanoString) compared to the current targeted therapy options in the health care region (*EGFR* and *ALK* inhibitor treatment), we analyzed actionable alterations defined from the literature in the 533 patients screened during 2015. Firstly, we defined a set of both acknowledged and proposed actionable oncogene mutations in specific oncogenes (*KRAS, EGFR, BRAF, PIK3CA, NRAS, ERBB2, MAP2K1*, and *AKT1*) in addition to *ALK, RET*, and *ROS1* gene fusions, using information from public sources[160] and reported studies[161, 162]. Next, we stratified the 533 samples based on existence of these actionable alterations in individual cases (Figure 22A). Of these actionable variants, alterations in *KRAS* dominated in all subgroups, followed by *EGFR, BRAF* and *PIK3CA* (Figure 22B). Gene fusions accounted for 7% of actionable alterations in adenocarcinomas. While the majority of detected actionable variants appeared mutually exclusive across samples (Figure 22C), a number of cases showed multiple actionable variants, e.g., concurrent *KRAS* and *PIK3CA* mutations, concurrent *KRAS* and *BRAF* mutations, and concurrent *KRAS/EGFR/BRAF* mutations and *ALK* fusions. While the two former observations may be explained by tumor subclonality, the high proportion of the latter observation is intriguing given the reported near to mutual exclusiveness of these alterations [162, 163]. Possible explanations may be tumor subclonality, however technical/interpretation issues in the *ALK* diagnostic scheme cannot be excluded. Second, we sought to determine the subset of patients with different histological subtypes that could be eligible for potential emerging treatments based on the defined actionable alterations (Figure 22D). In adenocarcinoma, this analysis suggested that 10.6% (50.3% if including *KRAS*) of cases could be eligible for emerging targeted treatments, beyond the 15.3% of cases eligible for standard *EGFR* or *ALK* targeted therapy (Figure 22D). For SqCC, similar proportions were lower, 4.4% of cases (11.1% if including *KRAS*) could be eligible for emerging treatments, in addition to the 2.2% eligible for *EGFR* targeted therapy. Finally, for NSCLC- NOS 2.5% of cases (45% with *KRAS*) could be eligible for novel/emerging targeted treatments, in addition to the 11.2% of cases eligible for standard *EGFR* or *ALK* targeted therapy.

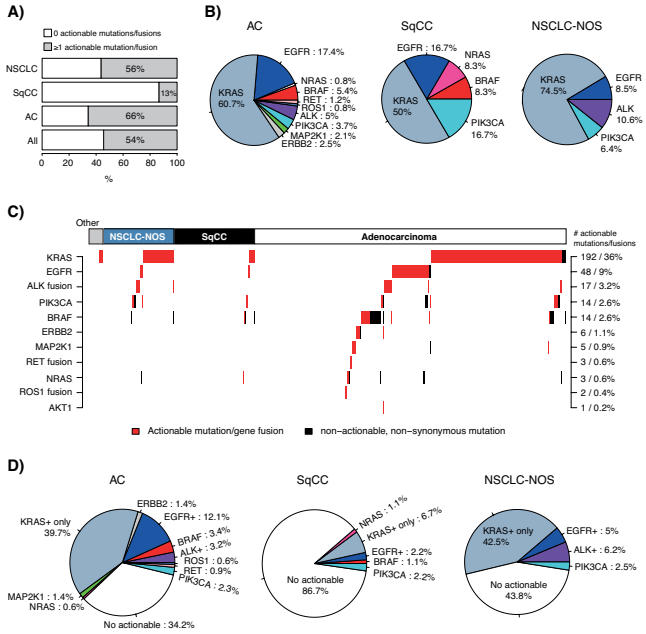


Figure 22. Integration of actionable mutations and gene fusions in the consecutive 533-sample cohort. (A) Proportion of cases with ≥1 actionable alteration in the total 533-sample cohort, adenocarcinomas (AC), SqCCs, and NSCLC-NOS. (B) Distribution of detected actionable variants according to the gene in which they fall for adenocarcinomas (AC, n=242 detected variants), SqCCs (n=12 variants) and NSCLC-NOS (n=47 variants). (C) Heatmap describing defined actionable and non-actionable non-synonymous variants and gene fusions in investigated genes identified in each case. Each column represents a sample; each row represents a gene. Numbers and proportions displayed on the right axis correspond to the total cohort (533 samples). (D) Proportion of cases with actionable mutations in adenocarcinoma (AC, n=348 samples), SqCC (n=90 samples), and NSCLC-NOS (n=80 samples). In each pie chart, EGFR+ corresponds to the proportion of cases with an actionable EGFR mutation irrespective of other alterations, ALK+ corresponds to cases with an ALK gene fusion irrespective of other alterations, and KRAS+_only corresponds to cases with only an actionable KRAS mutation. Consequently, some BRAF mutated cases may for instance harbor also an actionable KRAS variant. In all panels, not all cases were analyzed for gene fusions by the NanoString assay; consequently these estimates (mainly ROS1 and RET) should be interpreted as low frequency proportions.

As demonstrated in Figure 22, the gene mutation spectrum is strongly correlated with histological subtype. AC tumors appear more oncogene addictive with a wider spectrum and higher rates of actionable variants (n=242) compared with SqCC (n=12) and NOS (n=47). The wider spectrum of specific, targetable mutations in AC translates to an increase in potential therapy options compared with, e.g. SqCC and NOS. The two latter groups are characterized by a higher frequency of "no actionable" mutations or gene mutations (for instance *KRAS*) that are more difficult to target. However, it should be noted that the gene selection in the Illumina TruSight Tumor 26 panel (classical oncogenes and tumor suppressors) may represent a source of bias in these interpretations, as it has been shown, that for instance, that SqCC tumors harbor a different spectrum of mutations than AC tumors[41, 42]. Of importance, the high incidence of gene fusion in *EGFR/KRAS/BRAF* mutation-negative cases stresses the importance of

screening these "triple oncogene negative" cases more in depth. In Study I, we could show that NanoString based *ALK* fusion gene detection had a success rate equivalent to the then clinically used in situ methods (approximately 80%). *ROS1* gene fusions have been shown to be treatment predictive for *ALK*-inhibitor drugs[164]. In our investigated cohort of triple oncogene negative NSCLCs, the number of cases with either *ROS1* or *RET* fusions were similar to that of *ALK*-positive cases, supporting the need of multiplexed gene fusion diagnostics in NSCLC and AC specifically. Moreover, recent studies highlight the importance of knowing the specific type of fusion, as this might influence therapy response and resistance development[165, 166, 167], information that is not provided by routine IHC or FISH analysis. Clearly, to acquire information on multiple gene mutations and gene fusions in ideally a single assay is of great clinical value. The power and importance of a multigene assay was further demonstrated in Study II through the creation of a multicomponent NanoString assay targeting two relevant diagnostic questions in lung cancer. In Study II, we evolved the assay from Study I to not only include more targetable fusion genes, but also genes that could be used for other purposes, e.g. gene expression-based prediction of tumor histology, using a single extract of RNA. For histological prediction, we used a machine-learning algorithm, AIMS[149], to develop an SSP based on a set of transparent gene rules for prediction of three histological subtypes, AC, SqCC, and LCNEC. The basis for the predictor was the RNA expression of 11 genes whose protein expression had been associated with specific subtypes. The developed SSP proved remarkably accurate for NSCLC histology prediction, irrespective of analysis platform, when validated in external data sets, suggesting that it can be provided to the clinic as a complement to existing techniques. In contrast to previously published predictors[36-38, 168], the SSP derived in Study II is based on and applied to tumors histologically classified according to the WHO 2015 guidelines, representing the golden standard in clinical routine today. A combined tool for fusion gene detection and histology prediction would likely be highly useful in lung cancer diagnostics, especially in advanced disease, as tissue amounts are scarce due to small biopsies and cytologies. An advantage of the NanoString technology in this context is the flexibility of the probe design, allowing extension of the assay over time based on new knowledge. For instance, the NOS cases included in one of the validation cohorts in Study II had been identified as NOS in Study I, i.e. in a clinical setting. NOS tumors represent a likely obvious group of tumors where both an RNA-based histological predictor (Study II) and comprehensive mutation testing (like in Study I) may be of complementary value to routine diagnostics. Patients with NOS tumors are associated with poorer outcome, possibly due to the lack of therapy options that histological subgrouping provides coupled with a generally more aggressive undifferentiated disease. The challenging nature of NSCLC-NOS tumors is evident in the initially observed poor concordance rate between the SSP and the pathological re-review for AC cases in one of the

validation cohorts. However, discordant cases could be explained by either insufficient RNA quality (a challenge in archival tissue) or biological reasons such as SqCC dysplasia (Figure 17), or diagnosis based on markers (mucins) not included in the NanoString design. Importantly, these shortcomings can be addressed by: 1) an assay quality control step as in Study I, 2) appropriate micro/macro dissection considering the non-*in situ* type of analysis, and 3) update of the NanoString probe content, respectively. As depicted in Figure 22 many NOS cases presented with gene mutations often associated with certain NSCLC histology subgroups, e.g. *EGFR* mutations more frequently associated with AC and *PIK3CA* mutations associated with SqCC in Study I. Importantly, the NOS category in advanced disease included tumors that are "marker null" i.e. lack expression of markers associated with other NSCLC subtypes such as AC or SqCC and would hypothetically be classified as LCC if resected tissue had been available. Combining mutation status with, for instance, gene fusion status and RNA-based histology prediction in one assay would mean creating layers of information for a potentially better final classification that could impact patient treatment.

Stratification and classification in resected tumors

In advanced disease there are multiple genetic alterations crucial for first-line therapy choice – hence the emphasis is often on analysis of a small set of targetable genes (as in Studies I and II). While the latter fulfills the current clinical needs and provide some limited insight into the driver landscape of these tumors (Figure 22) the question of whether there exist other molecular subgroups of clinical value remains less clear. Two goals of comprehensive genomic analyses in lung cancer, e.g. whole genome transcriptional analyses, have been to try to refine the taxonomy of the disease through definition of molecular subgroups of tumors with clinical importance or to identify specific prognostic or treatment predictive signatures better than the current methods. However, comprehensive molecular studies in advanced disease are rare, likely due to tissue limitations (amount and type). Therefore, the aforementioned genomic studies have mainly used resected tumor material as this provides the necessary amounts and quality of nucleic acids. Naturally, this means that such studies are biased in terms of low stage disease and, to some extent, large tumors that must be considered when translating findings from molecular studies into clinical implications, especially concerning treatment. In this context, adjuvant therapy in early stage lung cancer does not generally include targeted therapy or immunotherapy. Thus, these types of studies are limited towards prognosis (risk of relapse) or treatment prediction of mainly adjuvant chemotherapy. It has been noted that "normal" tissue adjacent to a tumor infested area of the lung contains aberrant molecular/epigenetic changes associated

with higher risk of recurrence and secondary tumors. Therefore, greater surgical margins and further molecular analysis of surrounding tissue can be of both prognostic value as in identifying patients that would benefit from adjuvant chemotherapy[34].

This thesis contains two studies, III and V, aimed at refining the molecular landscape of early lung cancer to potentially unravel novel subgroups of clinical relevance. Study V was chronologically first in this thesis work and provided at the time comprehensive knowledge on the epigenetic landscape in lung cancer using a then state-of-the-art DNA methylation assay. In Study V, we derived five epitypes that were correlated with histology, gene expression subtypes and for adenocarcinomas also survival. Epigenetic subtypes associated with outcome in adenocarcinoma have been reported also in other studies[50, 89], supporting that this type of analysis can identify early stage tumors with different prognosis.

Although the findings in Studies III and V require extensive validation and are far from clinical implementation today, they together with several other reports serve as an example to illustrate that a likely subgrouping within histological subgroups associated with prognosis in at least early stage AC and SqCC appear highly possible[36, 37, 41, 47, 48, 53, 55, 57, 88]. If properly and successfully validated, such subtypes may ultimately become clinically relevant for patient management similar to the development and usage of molecular subtypes in breast cancer. Thus, molecular subtypes defined from extensive genomic characterization may contribute with information to personalize treatment for specific lung cancer patients. Besides prognosis, a better stratification of histological subgroups could also identify patients without targetable mutations that have similar epigenetic or transcriptional patterns as patient groups possessing a targetable mutation, potentially indicating a similar dependency of a specific signaling pathway (which may be targetable). One such speculative example is seen in Studies III and V, with a subset of adenocarcinomas without detectable *EGFR* mutations sharing similar transcriptional and epigenetic patterns as tumors harboring actionable *EGFR* mutations. Whether such tumors respond similarly or at least partially to EGFR inhibitors remains to be proven. There is also the possibility to identify a patient subset that would probably benefit from immunotherapy representing an immune hot subtype through such analyses. In summary, the type of stratification that genome-wide epigenetic and transcriptional profiling provide in early stage disease is valuable not only in the context of unraveling lung cancer biology, but hopefully also for those surgically treated patients that relapse.

Today a plethora of studies exist in the literature about molecular subtypes and different prognostic/predictive gene expression, copy number, or epigenetic signatures in early stage lung cancer using high throughput techniques applied to fresh frozen tissue (including Studies III and V). However, fewer studies exist on

how to move such tools closer to actual clinical use, e.g. involving analysis of fixated tissue and dealing with single sample prediction. Here, assays such as the multicomponent tool in Study II, including the concept of SSP predictors, may represent one possible route forward to bring novel signatures closer to future clinical use.

Therapy surveillance and tumorigenic adaption

Studies included in this thesis report on molecular findings that add layers to tumor biology and stratification of patient groups as well as tools to perform/aid in daily molecular routine analysis as part of clinical management. Molecular characterization and classification of tumors is important to gain understanding of biological processes that promote tumor formation but also to understand the mechanisms of therapy resistance which may be intrinsic or acquired. Interestingly, many patients undergoing treatment relapse or progress. Therefore, longitudinal monitoring during treatment with chosen therapy would be preferable in order to: 1) assess treatment response and 2) monitor tumorigenic adaption the treatment (e.g. resistance development during TKI treatment). Such monitoring would enhance our understanding of changes in biological processes in the tumor when subjected to therapeutic agents (therapeutic pressure), as well as to serve as an early indicator of when these changes will cause progression due to, e.g. emergence of subclonal resistance mutations. Due to the fact that tumors leak fragments of DNA into the blood stream, great efforts to retrieve this cell-free, circulating tumor DNA (cfDNA/ctDNA) by extraction from plasma is ongoing (although not a focus of this thesis). Therapy surveillance and observing tumorigenic adaption to treatment using ctDNA or analysis of exhaled breath condensates[169, 170] is an appealing approach as they represent non-invasive procedures compared to a conventional tissue biopsy, which may for various reasons be difficult to achieve for metastatic lesions. However, using ctDNA as a sample analyte requires highly sensitive detection techniques. Digital droplet PCR (ddPCR)[171] performs massive sample partitioning and individual PCR on each partition (droplet) using TaqMan assays. The technique is highly sensitive and suitable for ctDNA (as ctDNA is fragmented to a great extent) but is a single-gene assay. In response to this, a variety of commercially available library preparation kits for NGS are available, providing the possibility of multi-gene screening also in this context. Demonstrated in Study I, the introduction of NGS in the clinical setting, represents a major leap forward but current commercial amplicon-based panels (e.g. the TruSight Tumor and Ion Torrent AmpliSeq Colon and Lung panels) are biased towards analyzing hotspot alterations in a limited set of oncogenes often selected through a pan-cancer approach. To some extent, these

panels offer the possibility to detect intrinsic or acquired resistance mechanisms to targeted treatment, mainly T790M and C797S mutations in *EGFR* and specific gatekeeper mutations in *ALK* (like L1196M and G1269A), in patients re-biopsied after treatment failure. However, most panels (including TruSight Tumor and frequently used Ion Torrent panels) are less well suited to detect *EGFR/ALK* resistance mechanisms caused by alterations in other genes. Here, panel design and size constraints, but also problems in calling copy number alterations (like *MET* amplification as a mechanism of resistance to *EGFR* inhibitors) reliably in tumors with considerable non-malignant infiltration represent limiting factors. Therefore, diagnostic platforms based on, e.g., hybrid capture methods of either DNA alone[172, 173] or DNA and RNA combinations (e.g. the Illumina TruSight Tumor 170 panel and the AmpliSeq based Thermo Fisher Oncomine™ Focus/Comprehensive panels) that allow considerably more sequence to be analyzed could be the next preferable step also outside large comprehensive cancer centers. These assays could allow simultaneous detection of mutations, gene fusions, and copy number alterations (like drug targetable *MET* and *FGFR1* amplifications) in a large number of genes. Non-invasive techniques for therapy surveillance and tumorigenic adaption observation are compelling as it spares patients from often challenging tissue biopsies. However, in case of contradictory results or low amounts of ctDNA (causing e.g. inclusive results), a biopsy of tumor tissue is still required. Whether the technique to retrieve tumor DNA is non-invasive or more invasive to the patient, longitudinal observations of therapy response and tumorigenic adaption will likely be key factors in tomorrow's clinical management and molecular understanding of lung cancer.

Conclusions

Study I

A framework for treatment predictive mutation testing using NGS is feasible and useful in routine diagnostics. In parallel, gene fusion detection can be performed using the NanoString technology. Both technologies used are applicable to archival tissue and provides with information on multiple targets in one assay.

Study II

A multicomponent tool for gene fusion detection and histology prediction provides with two clinically important aspects in one assay applicable to archival tissue.

Study III

Investigating the transcriptional landscape of lung cancer supports the revised WHO guidelines with specific focus on LCC and LCNEC tumors.

Study IV

Specific mutations are associated with LCC and LCNEC tumors. LCC tumors can be further stratified based on markers associated with other NSCLC subgroups.

Study V

Global methylation patterns reveal distinct epitypes correlated with histopathological features, gene expression patterns and survival.

Future perspectives

As this thesis illustrates, lung cancer is a molecularly diverse and lethal disease. The options in lung cancer therapy are constantly expanding through molecular discoveries. Although much focus lie on development of novel molecular tools and therapies, one cannot disregard from the fact that smoking is the most prominent cause of lung cancer. With the overwhelming body of evidence of the detriment of tobacco to human health, many control policies have been implemented as health promotion actions. Such prevention methods include taxation of smoking, health warnings on tobacco products, marketing restrictions, and banning smoking in public places. Even so, smoking prevalence is still not expected to decrease globally. Therefore, screening of selected groups of the population with an increased risk of lung cancer (i.e. current or former smokers) would be the second-best option for decreasing mortality rates in lung cancer. Screening methods such as spirometry and chest x-ray could potentially alter the proportions of early versus advanced disease. Detection of low stage tumors increase survival rates through surgery. While smoking is a major health risk and undisputedly the major cause of lung cancer, about 20-25% of lung cancer patients are never-smokers. Irrespective of smoking status, proper clinical stratification of lung cancer patients is of great importance and would, in this context, be the third-best tactics in improving survival rates in lung cancer. Tumor characteristics such as specific gene mutation/fusion status and histology impacts therapy choice. Still, a vast majority of patients diagnosed with advanced disease lack obvious therapy options, historically leaving the remaining options to be of higher toxicity and/or less efficacy. With the addition of immunotherapy to the list of therapy options for those not eligible for targeted therapy, a remarkably prolonged survival of a large patient group has been observed. This particular patient group could perhaps, with time, be considered diagnosed with chronic disease. In the future, genomics will likely play an even larger role in the clinical management of lung cancer. Further molecular characterization and stratification of lung cancer is needed to understand resistance mechanisms and to identify patients and patient groups that would benefit from specific therapies or be spared of ineffective and/or harsh therapies. This will likely take place through continuous monitoring and analysis of a patient's tumor during the course of the disease to tailor treatment. Deepened understanding of lung cancer biology is essential for development of novel

therapeutics that, hopefully, will translate to better survival for patients diagnosed with lung cancer.

References

1. Tortora G, Grabowski S. *Principles of anatomy and physiology*. New York: Wiley; 2003.
2. Cotes JE, Chinn DJ, Miller MR. Lung Function: Physiology, Measurement and Application in Medicine. Blackwell Publishing Ltd; 2009.
3. Gray H. Anatomy of the Human Body. 1918.
4. Harding R, Pinkerton KE, Plopper CG. *The Lung. Development, Aging and the Environment*. Elsevier; 2004.
5. Rock JR, Hogan BL. Epithelial progenitor cells in lung development, maintenance, repair, and disease. *Annu Rev Cell Dev Biol* 2011;27:493-512.
6. Whitsett JA, Wert SE, Weaver TE. Alveolar surfactant homeostasis and the pathogenesis of pulmonary disease. *Annu Rev Med* 2010;61:105-119.
7. Adamson IY, Bowden DH. Derivation of type 1 epithelium from type 2 cells in the developing rat lung. *Lab Invest* 1975;32:736-745.
8. Williams MC. Alveolar type I cells: molecular phenotype and development. *Annu Rev Physiol* 2003;65:669-695.
9. Rock JR, Randell SH, Hogan BL. Airway basal stem cells: a perspective on their roles in epithelial homeostasis and remodeling. *Dis Model Mech* 2010;3:545-556.
10. Evans MJ, Van Winkle LS, Fanucchi MV, et al. Cellular and molecular characteristics of basal cells in airway epithelium. *Exp Lung Res* 2001;27:401-415.
11. Hong KU, Reynolds SD, Watkins S, et al. In vivo differentiation potential of tracheal basal cells: evidence for multipotent and unipotent subpopulations. *Am J Physiol Lung Cell Mol Physiol* 2004;286:L643-649.
12. Rock JR, Onaitis MW, Rawlins EL, et al. Basal cells as stem cells of the mouse trachea and human airway epithelium. *Proc Natl Acad Sci U S A* 2009;106:12771-12775.
13. Torre LA, Siegel RL, Jemal A. Lung Cancer Statistics. *Adv Exp Med Biol* 2016;893:1-19.
14. Health. USNIo. National Cancer Institute.; Available at https://seer.cancer.gov/statfacts/html/lungb.html.
15. Cheng TY, Cramb SM, Baade PD, et al. The International Epidemiology of Lung Cancer: Latest Trends, Disparities, and Tumor Characteristics. *J Thorac Oncol* 2016;11:1653-1671.
16. Islami F, Torre LA, Jemal A. Global trends of lung cancer mortality and smoking prevalence. *Transl Lung Cancer Res* 2015;4:327-338.
17. Bain C, Feskanich D, Speizer FE, et al. Lung cancer rates in men and women with comparable histories of smoking. *J Natl Cancer Inst* 2004;96:826-834.
18. Govindan R, Ding L, Griffith M, et al. Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell* 2012;150:1121-1134.

19. Sun S, Schiller JH, Gazdar AF. Lung cancer in never smokers--a different disease. *Nat Rev Cancer* 2007;7:778-790.
20. Samet JM, Avila-Tang E, Boffetta P, et al. Lung cancer in never smokers: clinical epidemiology and environmental risk factors. *Clin Cancer Res* 2009;15:5626-5645.
21. Hirsch FR, Harper P. *Lung Cancer.* Remedica; 2010.
22. AJCC. *AJCC Cancer Staging Manual.* New York.: Springer.; 2016.
23. UICC. *TNM Classification of Malignant Tumours.* Hoboken NJ.: Wiley-Blackwell; 2016.
24. Goldstraw P, Chansky K, Crowley J, et al. The IASLC Lung Cancer Staging Project: Proposals for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer. *J Thorac Oncol* 2016;11:39-51.
25. Detterbeck FC, Boffa DJ, Kim AW, et al. The Eighth Edition Lung Cancer Stage Classification. *Chest* 2017;151:193-203.
26. Swanton C, Govindan R. Clinical Implications of Genomic Discoveries in Lung Cancer. *N Engl J Med* 2016;374:1864-1873.
27. Travis WD, Brambilla E, Nicholson AG, et al. The 2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances Since the 2004 Classification. *J Thorac Oncol* 2015;10:1243-1260.
28. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell* 2000;100:57-70.
29. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell* 2011;144:646-674.
30. Alavanja MC. Biologic damage resulting from exposure to tobacco smoke and from radon: implication for preventive interventions. *Oncogene* 2002;21:7365-7375.
31. Van Lommel A. Pulmonary neuroendocrine cells (PNEC) and neuroepithelial bodies (NEB): chemoreceptors and regulators of lung development. *Paediatr Respir Rev* 2001;2:171-176.
32. Phillips DH. Smoking-related DNA and protein adducts in human tissues. *Carcinogenesis* 2002;23:1979-2004.
33. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. *Nature* 2013;500:415-421.
34. Herbst RS, Heymach JV, Lippman SM. Lung cancer. *N Engl J Med* 2008;359:1367-1380.
35. Wistuba, II, Mao L, Gazdar AF. Smoking molecular damage in bronchial epithelium. *Oncogene* 2002;21:7298-7306.
36. Charkiewicz R, Niklinski J, Claesen J, et al. Gene Expression Signature Differentiates Histology But Not Progression Status of Early-Stage NSCLC. *Transl Oncol* 2017;10:450-458.
37. Zhang A, Wang C, Wang S, et al. Visualization-aided classification ensembles discriminate lung adenocarcinoma and squamous cell carcinoma samples using their gene expression profiles. *PLoS One* 2014;9:e110052.

38. Girard L, Rodriguez-Canales J, Behrens C, et al. An Expression Signature as an Aid to the Histologic Classification of Non-Small Cell Lung Cancer. *Clin Cancer Res* 2016;22:4880-4889.
39. Djureinovic D, Hallstrom BM, Horie M, et al. Profiling cancer testis antigens in non-small-cell lung cancer. *JCI Insight* 2016;1:e86837.
40. Karlsson A, Jonsson M, Lauss M, et al. Genome-wide DNA methylation analysis of lung carcinoma reveals one neuroendocrine and four adenocarcinoma epitypes associated with patient outcome. *Clin Cancer Res* 2014;20:6127-6140.
41. Cancer Genome Atlas Research N. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 2012;489:519-525.
42. Cancer Genome Atlas Research N. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* 2014;511:543-550.
43. Imielinski M, Berger AH, Hammerman PS, et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* 2012;150:1107-1120.
44. Suda K, Tomizawa K, Yatabe Y, et al. Lung cancers unrelated to smoking: characterized by single oncogene addiction? *Int J Clin Oncol* 2011;16:294-305.
45. Karlsson A, Ringner M, Lauss M, et al. Genomic and transcriptional alterations in lung adenocarcinoma in relation to smoking history. *Clin Cancer Res* 2014;20:4912-4924.
46. Staaf J, Jonsson G, Jonsson M, et al. Relation between smoking history and gene expression profiles in lung adenocarcinomas. *BMC Med Genomics* 2012;5:22.
47. Wilkerson MD, Yin X, Walter V, et al. Differential pathogenesis of lung adenocarcinoma subtypes involving sequence mutations, copy number, chromosomal instability, and methylation. *PLoS One* 2012;7:e36530.
48. Bhattacharjee A, Richards WG, Staunton J, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci U S A* 2001;98:13790-13795.
49. Director's Challenge Consortium for the Molecular Classification of Lung A, Shedden K, Taylor JM, et al. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat Med* 2008;14:822-827.
50. Sandoval J, Mendez-Gonzalez J, Nadal E, et al. A prognostic DNA methylation signature for stage I non-small-cell lung cancer. *J Clin Oncol* 2013;31:4140-4147.
51. Selamat SA, Chung BS, Girard L, et al. Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression. *Genome Res* 2012;22:1197-1211.
52. Walter K, Holcomb T, Januario T, et al. DNA methylation profiling defines clinically relevant biological subsets of non-small cell lung cancer. *Clin Cancer Res* 2012;18:2360-2373.
53. Hayes DN, Monti S, Parmigiani G, et al. Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in multiple independent patient cohorts. *J Clin Oncol* 2006;24:5079-5090.
54. Ettinger DS, Akerley W, Bepler G, et al. Non-small cell lung cancer. *J Natl Compr Canc Netw* 2010;8:740-801.

55. Wilkerson MD, Yin X, Hoadley KA, et al. Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types. *Clin Cancer Res* 2010;16:4864-4875.
56. Takeuchi T, Tomida S, Yatabe Y, et al. Expression profile-defined classification of lung adenocarcinoma shows close relationship with underlying major genetic changes and clinicopathologic behaviors. *J Clin Oncol* 2006;24:1679-1688.
57. Garber ME, Troyanskaya OG, Schluens K, et al. Diversity of gene expression in adenocarcinoma of the lung. *Proc Natl Acad Sci U S A* 2001;98:13784-13789.
58. Raponi M, Zhang Y, Yu J, et al. Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. *Cancer Res* 2006;66:7466-7472.
59. Bhora FY, Chen DJ, Detterbeck FC, et al. The ITMIG/IASLC Thymic Epithelial Tumors Staging Project: A Proposed Lymph Node Map for Thymic Epithelial Tumors in the Forthcoming 8th Edition of the TNM Classification of Malignant Tumors. *J Thorac Oncol* 2014;9:S88-96.
60. Dogan S, Shen R, Ang DC, et al. Molecular epidemiology of EGFR and KRAS mutations in 3,026 lung adenocarcinomas: higher susceptibility of women to smoking-related KRAS-mutant cancers. *Clin Cancer Res* 2012;18:6169-6177.
61. Luo SY, Lam DC. Oncogenic driver mutations in lung cancer. *Transl Respir Med* 2013;1:6.
62. Paez JG, Janne PA, Lee JC, et al. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* 2004;304:1497-1500.
63. Lynch TJ, Bell DW, Sordella R, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N Engl J Med* 2004;350:2129-2139.
64. George J, Lim JS, Jang SJ, et al. Comprehensive genomic profiles of small cell lung cancer. *Nature* 2015;524:47-53.
65. Soda M, Choi YL, Enomoto M, et al. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature* 2007;448:561-566.
66. Horn L, Pao W. EML4-ALK: honing in on a new target in non-small-cell lung cancer. *J Clin Oncol* 2009;27:4232-4235.
67. Kohno T, Nakaoku T, Tsuta K, et al. Beyond ALK-RET, ROS1 and other oncogene fusions in lung cancer. *Transl Lung Cancer Res* 2015;4:156-164.
68. Dholaria B, Hammond W, Shreders A, et al. Emerging therapeutic agents for lung cancer. *J Hematol Oncol* 2016;9:138.
69. Tong JH, Yeung SF, Chan AW, et al. MET Amplification and Exon 14 Splice Site Mutation Define Unique Molecular Subgroups of Non-Small Cell Lung Carcinoma with Poor Prognosis. *Clin Cancer Res* 2016;22:3048-3056.
70. Scotti MM, Swanson MS. RNA mis-splicing in disease. *Nat Rev Genet* 2016;17:19-32.
71. Paik PK, Drilon A, Fan PD, et al. Response to MET inhibitors in patients with stage IV lung adenocarcinomas harboring MET mutations causing exon 14 skipping. *Cancer Discov* 2015;5:842-849.
72. Available at https://clinicaltrials.gov/ct2/show/NCT00585195.

73. Drilon A, Cappuzzo F, Ou SI, et al. Targeting MET in Lung Cancer: Will Expectations Finally Be MET? *J Thorac Oncol* 2017;12:15-26.
74. Drilon A. MET Exon 14 Alterations in Lung Cancer: Exon Skipping Extends Half-Life. *Clin Cancer Res* 2016;22:2832-2834.
75. Jones PA, Baylin SB. The epigenomics of cancer. *Cell* 2007;128:683-692.
76. Berdasco M, Esteller M. Aberrant epigenetic landscape in cancer: how cellular identity goes awry. *Dev Cell* 2010;19:698-711.
77. Saxonov S, Berg P, Brutlag DL. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A* 2006;103:1412-1417.
78. Deaton AM, Bird A. CpG islands and the regulation of transcription. *Genes Dev* 2011;25:1010-1022.
79. Vogelstein B, Papadopoulos N, Velculescu VE, et al. Cancer genome landscapes. *Science* 2013;339:1546-1558.
80. Guo M, House MG, Hooker C, et al. Promoter hypermethylation of resected bronchial margins: a field defect of changes? *Clin Cancer Res* 2004;10:5131-5136.
81. Bhutani M, Pathak AK, Fan YH, et al. Oral epithelium as a surrogate tissue for assessing smoking-induced molecular alterations in the lungs. *Cancer Prev Res (Phila)* 2008;1:39-44.
82. Licchesi JD, Westra WH, Hooker CM, et al. Promoter hypermethylation of hallmark cancer genes in atypical adenomatous hyperplasia of the lung. *Clin Cancer Res* 2008;14:2570-2578.
83. Belinsky SA, Liechty KC, Gentry FD, et al. Promoter hypermethylation of multiple genes in sputum precedes lung cancer incidence in a high-risk cohort. *Cancer Res* 2006;66:3338-3344.
84. Machida EO, Brock MV, Hooker CM, et al. Hypermethylation of ASC/TMS1 is a sputum marker for late-stage lung cancer. *Cancer Res* 2006;66:6210-6218.
85. Kim JS, Kim JW, Han J, et al. Cohypermethylation of p16 and FHIT promoters as a prognostic factor of recurrence in surgically resected stage I non-small cell lung cancer. *Cancer Res* 2006;66:4049-4054.
86. Brock MV, Hooker CM, Ota-Machida E, et al. DNA methylation markers and early recurrence in stage I lung cancer. *N Engl J Med* 2008;358:1118-1128.
87. Sato T, Arai E, Kohno T, et al. Epigenetic clustering of lung adenocarcinomas based on DNA methylation profiles in adjacent lung tissue: Its correlation with smoking history and chronic obstructive pulmonary disease. *Int J Cancer* 2014;135:319-334.
88. The Cancer Genome Atlas. Available at https://cancergenome.nih.gov.
89. Shinjo K, Okamoto Y, An B, et al. Integrated analysis of genetic and epigenetic alterations reveals CpG island methylator phenotype associated with distinct clinical characters of lung adenocarcinoma. *Carcinogenesis* 2012;33:1277-1285.
90. Chen F, Zhang Y, Parra E, et al. Multiplatform-based molecular subtypes of non-small-cell lung cancer. *Oncogene* 2017;36:1384-1393.
91. Cancercentrum R. Nationellt vårdprogram Lungcancer. 2015 Available at www.cancercentrum.se. Accessed 2015-03-10.

92. Available at https://www.nbt.nhs.uk/sites/default/files/filedepot/incoming/WHO_Performance_Status.doc.
93. Bria E, Gralla RJ, Raftopoulos H, et al. Magnitude of benefit of adjuvant chemotherapy for non-small cell lung cancer: meta-analysis of randomized clinical trials. *Lung Cancer* 2009;63:50-57.
94. Group NM-aC, Arriagada R, Auperin A, et al. Adjuvant chemotherapy, with or without postoperative radiotherapy, in operable non-small-cell lung cancer: two meta-analyses of individual patient data. *Lancet* 2010;375:1267-1277.
95. Ardizzoni A, Boni L, Tiseo M, et al. Cisplatin- versus carboplatin-based chemotherapy in first-line treatment of advanced non-small-cell lung cancer: an individual patient data meta-analysis. *J Natl Cancer Inst* 2007;99:847-857.
96. Hotta K, Matsuo K, Ueoka H, et al. Meta-analysis of randomized clinical trials comparing Cisplatin to Carboplatin in patients with advanced non-small-cell lung cancer. *J Clin Oncol* 2004;22:3852-3859.
97. Scagliotti GV, Parikh P, von Pawel J, et al. Phase III study comparing cisplatin plus gemcitabine with cisplatin plus pemetrexed in chemotherapy-naive patients with advanced-stage non-small-cell lung cancer. *J Clin Oncol* 2008;26:3543-3551.
98. Rossi G, Cavazza A, Marchioni A, et al. Role of chemotherapy and the receptor tyrosine kinases KIT, PDGFRalpha, PDGFRbeta, and Met in large-cell neuroendocrine carcinoma of the lung. *J Clin Oncol* 2005;23:8774-8785.
99. Onishi H, Araki T, Shirato H, et al. Stereotactic hypofractionated high-dose irradiation for stage I nonsmall cell lung carcinoma: clinical outcomes in 245 subjects in a Japanese multiinstitutional study. *Cancer* 2004;101:1623-1631.
100. Timmerman R, Papiez L, McGarry R, et al. Extracranial stereotactic radioablation: results of a phase I study in medically inoperable stage I non-small cell lung cancer. *Chest* 2003;124:1946-1955.
101. Nyman J, Johansson KA, Hulten U. Stereotactic hypofractionated radiotherapy for stage I non-small cell lung cancer--mature results for medically inoperable patients. *Lung Cancer* 2006;51:97-103.
102. Baumann P, Nyman J, Lax I, et al. Factors important for efficacy of stereotactic body radiotherapy of medically inoperable stage I lung cancer. A retrospective analysis of patients treated in the Nordic countries. *Acta Oncol* 2006;45:787-795.
103. Onishi H, Shirato H, Nagata Y, et al. Hypofractionated stereotactic radiotherapy (HypoFXSRT) for stage I non-small cell lung cancer: updated results of 257 patients in a Japanese multi-institutional study. *J Thorac Oncol* 2007;2:S94-100.
104. Rosell R, Moran T, Queralt C, et al. Screening for epidermal growth factor receptor mutations in lung cancer. *N Engl J Med* 2009;361:958-967.
105. Rosell R, Carcereny E, Gervais R, et al. Erlotinib versus standard chemotherapy as first-line treatment for European patients with advanced EGFR mutation-positive non-small-cell lung cancer (EURTAC): a multicentre, open-label, randomised phase 3 trial. *Lancet Oncol* 2012;13:239-246.
106. Morgillo F, Della Corte CM, Fasano M, et al. Mechanisms of resistance to EGFR-targeted drugs: lung cancer. *ESMO Open* 2016;1:e000060.

107. Awad MM, Shaw AT. ALK inhibitors in non-small cell lung cancer: crizotinib and beyond. *Clin Adv Hematol Oncol* 2014;12:429-439.
108. Sandler A, Gray R, Perry MC, et al. Paclitaxel-carboplatin alone or with bevacizumab for non-small-cell lung cancer. *N Engl J Med* 2006;355:2542-2550.
109. Ramalingam SS, Dahlberg SE, Langer CJ, et al. Outcomes for elderly, advanced-stage non small-cell lung cancer patients treated with bevacizumab in combination with carboplatin and paclitaxel: analysis of Eastern Cooperative Oncology Group Trial 4599. *J Clin Oncol* 2008;26:60-65.
110. Reck M, von Pawel J, Zatloukal P, et al. Phase III trial of cisplatin plus gemcitabine with either placebo or bevacizumab as first-line therapy for nonsquamous non-small-cell lung cancer: AVAil. *J Clin Oncol* 2009;27:1227-1234.
111. Lima AB, Macedo LT, Sasse AD. Addition of bevacizumab to chemotherapy in advanced non-small cell lung cancer: a systematic review and meta-analysis. *PLoS One* 2011;6:e22681.
112. Labarga P, Medrano J, Seclen E, et al. Safety and efficacy of tenofovir/emtricitabine plus nevirapine in HIV-infected patients. *AIDS* 2010;24:777-779.
113. Jubb AM, Harris AL. Biomarkers to predict the clinical efficacy of bevacizumab in cancer. *Lancet Oncol* 2010;11:1172-1183.
114. Sgambato A, Casaluce F, Sacco PC, et al. Anti PD-1 and PDL-1 Immunotherapy in the Treatment of Advanced Non- Small Cell Lung Cancer (NSCLC): A Review on Toxicity Profile and its Management. *Curr Drug Saf* 2016;11:62-68.
115. Brahmer J, Reckamp KL, Baas P, et al. Nivolumab versus Docetaxel in Advanced Squamous-Cell Non-Small-Cell Lung Cancer. *N Engl J Med* 2015;373:123-135.
116. Seiler C, Sharpe A, Barrett JC, et al. Nucleic acid extraction from formalin-fixed paraffin-embedded cancer cell line samples: a trade off between quantity and quality? *BMC Clin Pathol* 2016;16:17.
117. Haile S, Pandoh P, McDonald H, et al. Automated high throughput nucleic acid purification from formalin-fixed paraffin-embedded tissue samples for next generation sequence analysis. *PLoS One* 2017;12:e0178706.
118. Kotorashvili A, Ramnauth A, Liu C, et al. Effective DNA/RNA co-extraction for analysis of microRNAs, mRNAs, and genomic DNA from formalin-fixed paraffin-embedded specimens. *PLoS One* 2012;7:e34683.
119. Patel PG, Selvarajah S, Guerard KP, et al. Reliability and performance of commercial RNA and DNA extraction kits for FFPE tissue cores. *PLoS One* 2017;12:e0179732.
120. Saal LH, Vallon-Christersson J, Hakkinen J, et al. The Sweden Cancerome Analysis Network - Breast (SCAN-B) Initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine. *Genome Med* 2015;7:20.
121. Agilent Technologies. Available at https://www.agilent.com/.
122. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 1977;74:5463-5467.
123. Archer DX. Available at http://archerdx.com/.

124. Heuckmann JM, Holzel M, Sos ML, et al. ALK mutations conferring differential resistance to structurally diverse ALK inhibitors. *Clin Cancer Res* 2011;17:7394-7401.
125. Technologies N. Available at https://www.nanostring.com/.
126. Lira ME, Choi YL, Lim SM, et al. A single-tube multiplexed assay for detecting ALK, ROS1, and RET fusions in lung cancer. *J Mol Diagn* 2014;16:229-243.
127. Geiss GK, Bumgarner RE, Birditt B, et al. Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nat Biotechnol* 2008;26:317-325.
128. Malkov VA, Serikawa KA, Balantac N, et al. Multiplexed measurements of gene signatures in different analytes using the Nanostring nCounter Assay System. *BMC Res Notes* 2009;2:80.
129. Reis PP, Waldron L, Goswami RS, et al. mRNA transcript quantification in archival samples using multiplexed, color-coded probes. *BMC Biotechnol* 2011;11:46.
130. Zhang DY, Chen SX, Yin P. Optimizing the specificity of nucleic acid hybridization. *Nat Chem* 2012;4:208-214.
131. Karlsson A, Brunnstrom H, Micke P, et al. Gene Expression Profiling of Large Cell Lung Cancer Links Transcriptional Phenotypes to the New Histological WHO 2015 Classification. *J Thorac Oncol* 2017;12:1257-1267.
132. Karlsson A, Brunnstrom H, Lindquist KE, et al. Mutational and gene fusion analyses of primary large cell and large cell neuroendocrine lung cancer. *Oncotarget* 2015;6:22028-22037.
133. Brunnstrom H, Johansson L, Jirstrom K, et al. Immunohistochemistry in the differential diagnostics of primary lung cancer: an investigation within the Southern Swedish Lung Cancer Study. *Am J Clin Pathol* 2013;140:37-46.
134. Illumina. Available at https://emea.illumina.com/.
135. Sandoval J, Heyn H, Moran S, et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* 2011;6:692-702.
136. Bibikova M, Barnes B, Tsan C, et al. High density DNA methylation array with single CpG site resolution. *Genomics* 2011;98:288-295.
137. Dedeurwaerder S, Defrance M, Calonne E, et al. Evaluation of the Infinium Methylation 450K technology. *Epigenomics* 2011;3:771-784.
138. Lauss M, Visne I, Kriegner A, et al. Monitoring of technical variation in quantitative high-throughput datasets. *Cancer Inform* 2013;12:193-201.
139. Workman C, Jensen LJ, Jarmer H, et al. A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol* 2002;3:research0048.
140. Ritchie ME, Dunning MJ, Smith ML, et al. BeadArray expression analysis using bioconductor. *PLoS Comput Biol* 2011;7:e1002276.
141. The R Foundation for Statistical Computing. Available at http://www.R-project.org.
142. Barbosa-Morais NL, Dunning MJ, Samarajiwa SA, et al. A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. *Nucleic Acids Res* 2010;38:e17.

143. Monti S. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning* 2003:91-118.
144. Hupe P, Stransky N, Thiery JP, et al. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics* 2004;20:3413-3422.
145. Consortium GO. Available at http://www.geneontology.org/.
146. Mi H, Poudel S, Muruganujan A, et al. PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res* 2016;44:D336-342.
147. Eddy JA, Sung J, Geman D, et al. Relative expression analysis for molecular cancer diagnosis and prognosis. *Technol Cancer Res Treat* 2010;9:149-159.
148. Tan AC, Naiman DQ, Xu L, et al. Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics* 2005;21:3896-3904.
149. Paquet ER, Hallett MT. Absolute assignment of breast cancer intrinsic molecular subtype. *J Natl Cancer Inst* 2015;107:357.
150. Altman DG, Bland JM. Diagnostic tests 3: receiver operating characteristic plots. *BMJ* 1994;309:188.
151. Perou CM, Sorlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature* 2000;406:747-752.
152. Williams C, Ponten F, Moberg C, et al. A high frequency of sequence alterations is due to formalin fixation of archival specimens. *Am J Pathol* 1999;155:1467-1471.
153. Clinical Lung Cancer Genome P, Network Genomic M. A genomics-based classification of human lung tumors. *Sci Transl Med* 2013;5:209ra153.
154. Rekhtman N, Pietanza MC, Hellmann MD, et al. Next-Generation Sequencing of Pulmonary Large Cell Neuroendocrine Carcinoma Reveals Small Cell Carcinoma-like and Non-Small Cell Carcinoma-like Subsets. *Clin Cancer Res* 2016;22:3618-3629.
155. Miyoshi T, Umemura S, Matsumura Y, et al. Genomic Profiling of Large-Cell Neuroendocrine Carcinoma of the Lung. *Clin Cancer Res* 2017;23:757-765.
156. Simbolo M, Mafficini A, Sikora KO, et al. Lung neuroendocrine tumours: deep sequencing of the four World Health Organization histotypes reveals chromatin-remodelling genes as major players and a prognostic role for TERT, RB1, MEN1 and KMT2D. *J Pathol* 2017;241:488-500.
157. Ringner M, Staaf J. Consensus of gene expression phenotypes and prognostic risk predictors in primary lung adenocarcinoma. *Oncotarget* 2016;7:52957-52973.
158. Ernst J, Kheradpour P, Mikkelsen TS, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 2011;473:43-49.
159. Crino L, Weder W, van Meerbeeck J, et al. Early stage and locally advanced (non-metastatic) non-small-cell lung cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol* 2010;21 Suppl 5:v103-115.
160. My Cancer Genome. Available at https://www.mycancergenome.org.
161. Hagemann IS, Devarakonda S, Lockwood CM, et al. Clinical next-generation sequencing in patients with non-small cell lung cancer. *Cancer* 2015;121:631-639.

162. Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. *Nature* 2012;490:61-70.
163. Takeuchi K, Soda M, Togashi Y, et al. RET, ROS1 and ALK fusions in lung cancer. *Nat Med* 2012;18:378-381.
164. Shaw AT, Ou SH, Bang YJ, et al. Crizotinib in ROS1-rearranged non-small-cell lung cancer. *N Engl J Med* 2014;371:1963-1971.
165. Woo CG, Seo S, Kim SW, et al. Differential protein stability and clinical responses of EML4-ALK fusion variants to various ALK inhibitors in advanced ALK-rearranged non-small cell lung cancer. *Ann Oncol* 2017;28:791-797.
166. Lin JJ, Zhu VW, Yoda S, et al. Impact of EML4-ALK Variant on Resistance Mechanisms and Clinical Outcomes in ALK-Positive Lung Cancer. *J Clin Oncol* 2018;36:1199-1206.
167. Christopoulos P, Endris V, Bozorgmehr F, et al. EML4-ALK fusion variant V3 is a high-risk feature conferring accelerated metastatic spread, early treatment failure and worse overall survival in ALK(+) non-small cell lung cancer. *Int J Cancer* 2018;142:2589-2598.
168. Wilkerson MD, Schallheim JM, Hayes DN, et al. Prediction of lung cancer histological types by RT-qPCR gene expression in FFPE specimens. *J Mol Diagn* 2013;15:485-497.
169. Youssef O, Knuuttila A, Piirila P, et al. Presence of cancer-associated mutations in exhaled breath condensates of healthy individuals by next generation sequencing. *Oncotarget* 2017;8:18166-18176.
170. Youssef O, Sarhadi VK, Armengol G, et al. Exhaled breath condensate as a source of biomarkers for lung carcinomas. A focus on genetic and epigenetic markers-A mini-review. *Genes Chromosomes Cancer* 2016;55:905-914.
171. Bio-Rad. Available at www.bio-rad.com.
172. Frampton GM, Fichtenholtz A, Otto GA, et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat Biotechnol* 2013;31:1023-1031.
173. Cheng DT, Mitchell TN, Zehir A, et al. Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. *J Mol Diagn* 2015;17:251-264.

Acknowledgements

Since I've managed to remain at the department of Oncology and Pathology for an impressive 13 years this means two things: 1) It's the best workplace ever and 2) the acknowledgement section of a thesis is extensive. Several excellent co-workers and persons close to me have contributed to the finalization of this thesis. Those not mentioned here are not less important, merely not stated!

First of all: thank you *Johan Staaf* for accepting me as a PhD student in the lung cancer research group. I am truly honored. I appreciate handing me a splendid opportunity to retrieve some of your vast knowledge in so many areas. It has been a pleasure being part of the expanding lung cancer research group and I'm privileged to have accompanied you on your first lap on the route to success in cancer research.

My co-supervisor *Maria Planck*: as the founder of the lung cancer research group you have created an atmosphere of high research standards, great profession and a deep devotion for clinical management in relation to research. Your everlasting enthusiasm, great professionalism and sense of humor is a fantastic combination. Thank you for including me and reminding us all of the inevitable string between research and "the real world": the clinic.

Thanks to my co-supervisor *Markus Ringnér* for giving me the opportunity to learn from the most skillful bioinformatician I've met. You have a perceptive outlook on research and I can always trust you'll share with me your point of view. Thanks for all friendly and humoristic talks over the years, scientific or non-scientific!

Professor *Åke Borg*: thank you for finding a position for me as a technician more than a decade ago in your ever-expanding group. You have created so many opportunities for research to thrive at the department and it has been a privilege to be a part of that infrastructure. And thank you for sharing my passion for novel techniques and cool machines!

To past and present heads of department: *Bo Baldetorp, Mef Nilbert, Signe Borgquist, Ingrid Wilson, Lars Ekblad* and *Karin Jirström* for running the department with excellence, providing an environment for fruitful doctoral studies.

Ingrid Wilson, Head of Department at Medicon Village. I am so proud of all the years I spent working with you and all fantastic colleagues at the SCIBLU microarray facility. Thanks for constantly inviting us to your home where we have spent so many joyful days discussing everything but work!

Thanks to *Björn Frostner* and *Susanne André* for flawless administration!

To my dear colleagues at the former *SCIBLU Microarray facility*, whom produced so many trustworthy research-results due to skillful laboratory personnel and bioinformaticians. And to *SCAN-B,* for handing me the opportunity to engage in this important and impressive initiative. *Åke Borg*, *Lao Saal, Ingrid Wilson, Johan Vallon-Christersson, Jari Häkkinen, Nicklas Nordborg, Cecilia Hegardt, Göran Jönsson, Olle Månsson, Inger Remse, Minerva Li, Karin Annersten, Cecilia Wahlström, Jeanette Valcich* and *Ralph Schulz*.

Special thanks to *Göran Jönsson* for including me in the melanoma research group and for inviting me to melanoma retreats worldwide!

To my fellow PhD students in the lung cancer research group, *Annette Salomonsson, Sofi Isaksson* and *Bassam Hazem* for all your support and fruitful discussions during (and anywhere in between) our group meetings. A special thanks to *Annette* for designing the cover of this thesis!

To my best partners in lab-crime: *Mats Jönsson*, *Christel Reuterswärd* and *Frida Rosengren*. There is no lab-mystery unsolved with you involved. Thanks for all laughs and true friendship over the years and for sharing with me your vast knowledge in a variety of methods.

Frida Rosengren, Karolina Holm, Karin Annersten, Christel Reuterswärd, Mats Jönsson, Olle Månsson, Inger Remse, Göran Jönsson and *Johan Staaf* aka "the MeLuDi crew" for all joint hard work building the treatment predictive mutation testing framework. Thanks for all friendly and humoristic weekly meetings and for contributing to Study I in this thesis.

Eva Rambech, *Karin Haraldsson*, *Cecilia Forsberg*, *Maria Christiansson*, *Helena Malmberg*, *Lina Tellhed*, *Camilla Olsson*, *Steina Aradottir* and *Therese Törngren* at BRCA-lab for all good times spent in and outside of the lab.

Hans Brunnström and *Kajsa Ericson Lindquist,* department of Pathology in Lund for all contributions to studies included in this thesis and for kindly sharing with me your expertise and vast knowledge in the very confusing (to a non-Pathologist) but intriguing field of pathology.

Katja Harbst for being a true friend and for all good talks over tea in the lunchroom.

To all present and former roommates over the years: *Adriana Sanna, Rita Cabrita, Nicolai Arildsen, Christian Brueffer, Sergii Gladchuk* and *Robert Rigo*. For fruitful discussions, mostly about non-work-related stuff! Special thanks to *Jeanette Valcich* and *Helena Cirenajwis* for your true friendship and all time spent outside of the lab!

To all co-authors for valuable input, insights and contributions!

To all funders of this work: *the Swedish Cancer Society, the Mrs Berta Kamprad Foundation, the Gunnar Nilsson Cancer Foundation, the Crafoord Foundation, BioCARE a Strategic Research Program at Lund University, the Gustav V:s Jubilee Foundation, Skåne University Hospital Foundation, and The National Health Services (Region Skåne/ALF)*.

To all patients, respectfully, whom so generously contributed to all biobanks.

To *mormor* and *morfar* for enriching my life!

In honor of two persons that I know are with me in spirit: *farmor Ingegerd* and my father-in-law *Mats*. Rest peacefully knowing your support carried me this far.

To my mother-in-law, *Helena* and her husband *Holger*, for always taking care of my family when I'm not able to.

To my brothers-in-law, *Andreas* and *Nicklas*, for constantly having a strong opinion on basically everything. You should try research! Thanks, *Nastassia* for being my own private proofreader and an excellent chef. And to *Julia* (Louise's BFF) for sharing my passion for research and cats.

I am surrounded by strong women, forces of nature that stops for nothing. You are my inspiration and my support when life strikes. You have my back and I have yours. To my beloved friends: *Lina*, *Frida*, *Anna*, *Marie*, *Mimmi*, *Jenny, Rebecca* and *Pia*.

To my brother *Martin* for being a solid rock always! And for putting things in perspective. Thanks for enriching my life by bringing your wife *Agnes* and your three lovely kids into my world!

To my parents, *Lars-Olof* and *Ann-Kristin*, for unconditional love, never-ending support and for being the role models I aspire to be for my own children.

To my partner for life (despite the fact that a ring is missing even after 18 years), *Daniel*, for supporting me always, for believing in me when I'm not and for sharing with me the joy of being a family. Life is brighter when spending it with you.

Louise and *Amanda*, I will love you always. You are my world and my reason for existing ♥

Study I

Clinical framework for next generation sequencing based analysis of treatment predictive mutations and multiplexed gene fusion detection in non-small cell lung cancer

Kajsa Ericson Lindquist[1,*], Anna Karlsson[2,*], Per Levéen[1], Hans Brunnström[1,3], Christel Reuterswärd[2], Karolina Holm[2], Mats Jönsson[2], Karin Annersten[2], Frida Rosengren[2], Karin Jirström[1,3], Jaroslaw Kosieradzki[4], Lars Ek[4], Åke Borg[2,5], Maria Planck[2,6], Göran Jönsson[2,5], Johan Staaf[2,5]

[1]Department of Pathology, Regional Laboratories Region Skåne, Lund SE 22185, Sweden

[2]Division of Oncology and Pathology, Department of Clinical Sciences Lund, Lund University, Medicon Village, Lund SE 22381, Sweden

[3]Division of Oncology and Pathology, Department of Clinical Sciences Lund, Lund University, Lund SE 22185, Sweden

[4]Department of Respiratory Medicine and Allergology, Skane University Hospital, Lund SE22185, Sweden

[5]CREATE Health Strategic Center for Translational Cancer Research, Lund University, Medicon Village, Lund SE 22381, Sweden

[6]Department of Oncology, Skåne University Hospital, Lund SE 22381, Sweden

[*]These authors have contributed equally to this work

Correspondence to: *Johan Staaf,* ***email:*** *johan.staaf@med.lu.se*

Keywords: *lung cancer, NGS, gene fusion, mutation, precision medicine*

Received: *November 01, 2016* ***Accepted:*** *March 01, 2017* ***Published:*** *March 16, 2017*

ABSTRACT

Precision medicine requires accurate multi-gene clinical diagnostics. We describe the implementation of an Illumina TruSight Tumor (TST) clinical NGS diagnostic framework and parallel validation of a NanoString RNA-based *ALK*, *RET*, and *ROS1* gene fusion assay for combined analysis of treatment predictive alterations in non-small cell lung cancer (NSCLC) in a regional healthcare region of Sweden (Scandinavia). The TST panel was clinically validated in 81 tumors (99% hotspot mutation concordance), after which 533 consecutive NSCLCs were collected during one-year of routine clinical analysis in the healthcare region (~90% advanced stage patients). The NanoString assay was evaluated in 169 of 533 cases. In the 533-sample cohort 79% had 1-2 variants, 12% >2 variants and 9% no detected variants. Ten gene fusions (five *ALK*, three *RET*, two *ROS1*) were detected in 135 successfully analyzed cases (80% analysis success rate). No *ALK* or *ROS1* FISH fusion positive case was missed by the NanoString assay. Stratification of the 533-sample cohort based on actionable alterations in 11 oncogenes revealed that 66% of adenocarcinomas, 13% of squamous carcinoma (SqCC) and 56% of NSCLC not otherwise specified harbored ≥1 alteration. In adenocarcinoma, 10.6% of patients (50.3% if including *KRAS*) could potentially be eligible for emerging therapeutics, in addition to the 15.3% of patients eligible for standard EGFR or ALK inhibitors. For squamous carcinoma corresponding proportions were 4.4% (11.1% with *KRAS*) vs 2.2%. In conclusion, multiplexed NGS and gene fusion analyses are feasible in NSCLC for clinical diagnostics, identifying notable proportions of patients potentially eligible for emerging molecular therapeutics.

INTRODUCTION

Discoveries of frequent and therapeutically targetable mutations and gene fusions in non-small cell lung cancer (NSCLC) have changed not only the clinical management of the disease, but also the procedures and techniques used in the diagnosis of the disease. In addition to the current cornerstones of targeted therapy in NSCLC, *EGFR* mutations and *ALK* gene fusions, a growing number of alterations, like *ROS1* gene fusions, are emerging as treatment predictive in lung cancer broadening the cohort of patients eligible for targeted treatment [1].

Until recently, clinical analyses of treatment predictive alterations in *EGFR* and *ALK* have predominantly been performed by different single gene assays, e.g., real-time PCR or pyrosequencing, and immunohistochemistry (IHC) or fluorescence *in situ* hybridization (FISH), respectively. Given the continuous discovery of new, potentially treatment predictive alterations in lung cancer (see e.g. [1]) and a growing understanding of treatment resistance mechanisms, iterative single gene diagnostics is becoming problematic. Specifically, multiple analyses per sample increase the cost, require more input material and a longer time to generate results, in addition to the cumbersome nature of some methods (e.g. FISH). With the introduction of next generation sequencing (NGS) to the field of molecular genetics, and more recently also to the field of clinical diagnostics by allowing formalin-fixed paraffin embedded (FFPE) tissues to be screened, new possibilities exist for cost-, time- and sample efficient analysis of many different treatment predictive alterations in one analysis. Today, NGS-based diagnostics of treatment predictive mutations are running in large scale in large cancer centers worldwide and numerous reports of different implementations and techniques exist (see e.g., [2–8]). However, the technology is also increasingly introduced in smaller, often decentralized, healthcare regions at regional/local pathology departments with limitations in sample flow, budget, trained personnel, NGS equipment and bioinformatics structures, but still obliged to deliver accurate and timely results to guide patient therapy decisions.

The aim of the present study was to: i) implement a centralized NGS-based framework in the southern health care region of Sweden, Scandinavia, corresponding to one of the larger decentralized healthcare regions in Sweden, for clinical analysis of treatment predictive mutations in NSCLC, ii) determine the potential diagnostic yield of the NGS testing based on a complete year of clinical analysis, and iii) to investigate the clinical potential of multiplexed gene fusion analysis of *ALK*, *RET*, and *ROS1* based on RNA expression (Figure 1).

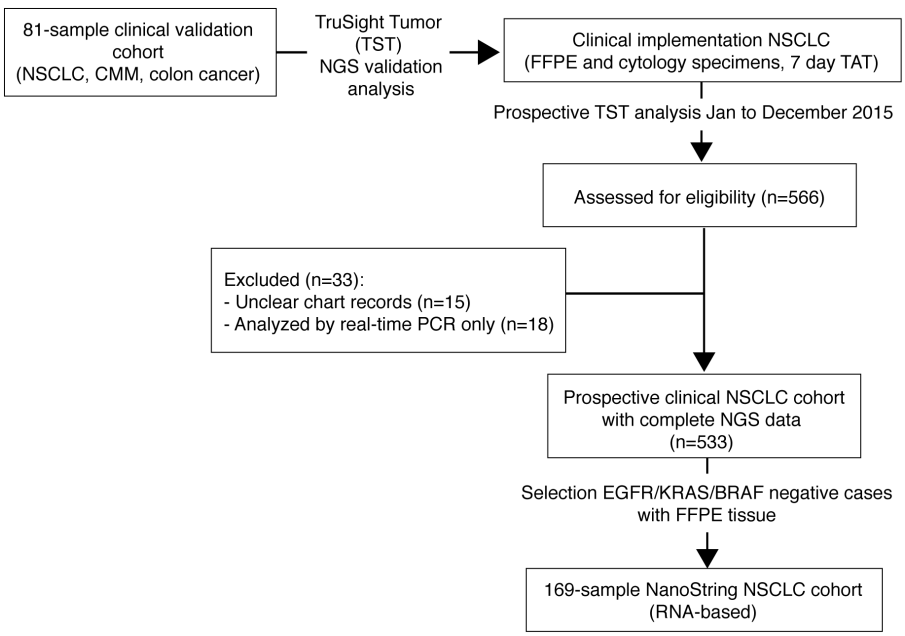


Figure 1: Study scheme outlining analyses and cohorts. FFPE: formalin-fixed paraffin embedded tissue, TAT: turnaround time.

RESULTS

Validation of an NGS-based assay compared to routine single gene diagnostics

To validate the Illumina TruSight Tumor (TST) NGS panel for clinical usage we analyzed 81 lung cancers, cutaneous malignant melanomas (CMMs) and colon cancers with existing clinical mutation data for hotspot mutations in *EGFR*, *KRAS*, *NRAS*, and *BRAF* (Table 1) (in addition to our previous validation of TST in a research setting, [9]). In total, the 81 cases harbored 29 known hotspot mutation calls and 63 calls of no mutation present for the investigated genes and loci. Of the total 92 mutation calls, concordance between previous single gene clinical testing methods and the TST assay was observed for 88 calls (96%) (Additional file 1). Three of the four discordant calls were due to a variant detected by TST but not analyzed by the corresponding single gene assay. Excluding these variants implied a concordance of 99% between TST and prior clinical methods. In the remaining single discordant sample (a colon cancer), a *NRAS* c.182A>G variant (38% TST variant allele frequency, VAF) was detected by all methods, with an additional c.35G>C *KRAS* variant called by the prior clinical real-time PCR method. A reanalysis was performed using sections from the same tissue block with the prior clinical real-time PCR method, TST, pyrosequencing, and complementary real-time PCR (Qiagen Therascreen). Reanalysis with the clinical real-time PCR method again identified the *KRAS* c.35G>C variant, while pyrosequencing identified a different *KRAS* variant (c.35G>T, 5% VAF). In contrast, TST analysis and Therascreen real-time PCR analysis agreed that no variants in *KRAS* were observed. The observation of both an activating *KRAS* and *NRAS* mutation in the same tumor is unlikely, suggesting that the discrepant *KRAS* variant might represent a false positive call (supported by the low VAF from pyrosequencing, and the different variants reported by pyrosequencing and the prior clinical real-time PCR method).

Using the Qiagen Therascreen *EGFR*, *KRAS*, and *BRAF* RGQ kits as reference methods in this study, we were able to validate detected hotspot variants down to 4% VAF from the NGS analysis in both the validation cohort and the subsequent prospective cohort in all tested cases, thus representing the effectively used limit of detection in later clinical samples (notably, a strict 10% tumor percentage cut-off was used for decision to perform clinical mutation testing at all).

Clinical implementation of an NGS-based diagnostic framework

The clinical implementation, including personnel and budget, of the NGS-based framework is described in Additional file 2. Following the clinical implementation (January 7, 2015), NGS analysis was the primary assay for routine clinical analysis of treatment predictive mutations in NSCLC and results were used to guide patient treatment. All identified mutations were reported to the diagnostic pathologist through a NGS report. During the prospective time period, only *EGFR* and *KRAS* mutation status were included in the pathological report returned to the treating clinician to guide treatment. For *ALK* fusions, the main method during the investigated time period was IHC and/or FISH. NanoString evaluation of RNA-based fusion detection was performed in parallel, but was not used to guide treatment. During the investigated time period (January 7 to December 31 2015), on average 12 suspected lung cancers were analyzed per week, of which 74.5% were FFPE sections and 25.5% cytology material. The turnaround time (TAT) for the molecular testing (DNA extraction, NGS analysis, and mutation report) was seven calendar days (i.e. five work days) in 94% of all cases analyzed during 2015, eight calendar days in 4%, and 9-10 days in 2% of cases (Additional file 2). The median TAT for the entire molecular process (from clinical referral, pathological evaluations, molecular analysis, to the final clinical report) was 14 calendar days (mean=15±6 calendar days) (Additional file 2). Of all originally referred suspected lung cancer cases, 4.7% could not be analyzed during the first pass through the centralized NGS laboratory due to insufficient DNA quality in the qPCR quality control step. The latter was caused by either degraded DNA, or more often by insufficient amounts of extracted DNA from the FFPE sections sent for analysis (mainly small biopsy specimens). With a few exceptions, all of these cases were however analyzed either by resampling followed by new NGS-analysis, or by a real-time PCR method (case then excluded from the prospective cohort analyzed in this study, see Figure 1). Based on this diagnostic framework, we collected 533 consecutively tested lung cancers by the TST NGS panel during 2015 to determine the diagnostic yield (Table 1). Notably, the proportions of the histological subtypes in the consecutive clinical testing cohort differ slightly from what might be expected from a Swedish population-based cohort (especially a lower proportion of squamous cell carcinomas, SqCC, 17% versus 21% based on data from the Swedish lung cancer registry). This suggests a potential selection bias between histological subtypes in the offered reflex-testing scheme. The bias could originate from the decentralized clinical management of patients in different regional hospitals within the healthcare region, coupled with a previous history of testing only adenocarcinomas.

NGS-based clinical analysis of a consecutive lung cancer cohort

Among the 533 cases, 889 variants were called by the standard vendor supplied data analysis pipeline.

Table 1: Clinicopathological characteristics of the validation and prospective cohorts

Validation cohort			
	Lung cancer	CMM[A]	Colon cancer
Number of patients	40	22	19
FFPE / Cytology (%)	70/30	100/0	100/0
Number of hot spot mutation calls (mut / no mutation)	8/32	9/13	12/18
KRAS	-	-	10/8
NRAS	-	0/1	1/5
BRAF	-	9/12	1/5
EGFR	8/32	-	-

Prospective cohort of lung cancer cases (n=533)				
	AC[B]	SqCC[B]	NSCLC-NOS[B]	Other[B]
Number of patients (%)	348 (65%)	90 (17%)	80 (15%)	15 (3%)
FFPE / Cytology (%)	79/21	84/16	47/53	60/40
Median age (years±sd)	69±9	72±8.3	70±8.5	68±7.6
Sex (female/male %)	53/47	38/62	51/49	60/40
Early stage cancer (%)	12.5%	7.8%	0%	20%
Clinical ALK analysis (n_{tot}=491)	324	86	68	13
ALK FISH-positive (n)	10	0	4	0
ALK IHC-positive (n)	2	2	1	1
NanoString analyzed cases[C] (%)	68 (50%)	54 (40%)	8 (6%)	5 (4%)

A: CMM: cutaneous malignant melanoma.
B: AC: adenocarcinoma, SqCC: squamous cell carcinoma, NSCLC-NOS: non-small cell lung cancer not otherwise specified, Other: including adenosquamous, large cell carcinoma, large cell neuroendocrine carcinoma, sarcomatoids.
C: Only cases that were *KRAS*/*EGFR*/*BRAF* mutation negative, from FFPE tissue, and with successful NanoString hybridizations are listed.

In general, analyzed cases showed few alterations in the investigated genes across different sample groups (total cohort, adenocarcinoma, SqCC, and NSCLC not otherwise specified, NSCLC-NOS), with ~80% of cases having 1-2 called variants and 7-9% no detected variants (Figure 2A). In all major sample groups (adenocarcinoma, SqCC, and NSCLC-NOS) *TP53* was the most frequently mutated gene, while the mutational pattern for *KRAS* and *EGFR* (second and third most frequently mutated genes in total) differed between sample/histological groups (Figure 2B). For 14 genes in the 26-gene panel, mutation frequencies in the total cohort were ~2% or less, suggesting that genetic alterations in these genes represent more rare driver events in NSCLC. Associations between mutation status for individual genes and clinicopathological variables (age, gender, and tumor histology) were scarce, with exception for *BRAF* (adenocarcinoma histology), *EGFR* (adenocarcinoma histology), *KRAS* (younger age, gender, adenocarcinoma histology), *CTNNB1* (gender), *PTEN* (adenocarcinoma histology), *STK11* (adenocarcinoma histology), and *TP53* (adenocarcinoma histology) (Additional file 3). In contrast with the literature there was no association between presence of *EGFR* mutation and gender in the total prospective clinical testing cohort (with 53/47% females/males), or in adenocarcinoma specifically (p=0.66 and 0.87, respectively, Fisher's exact test). In lack of complete patient smoking status (not consistently available in pathological referrals) this insignificant association is difficult to assess.

KRAS and *EGFR* (first and second most mutated oncogenes) showed a striking enrichment of specific, well-established, activating variants. In *KRAS*, variants in codons 12 and 13 constituted ~91% of all detected variants, while exon 19 deletions and p.L858R point mutations constituted

~75% of all detected variants in *EGFR* (see Additional file 4 for complete protein localization of variants in *EGFR*, *KRAS*, *BRAF* and *TP53*). For *EGFR*, mutations beside exon 19 deletions and p.L858R (c.2573T>G) variants, such as p.G719X (c.2155G>A, c.2156G>C, and c.2156G>T), p.L861Q (c.2582T>A, n=3), and exon 20 insertions (n=2), each represented ≤5% of the total *EGFR* mutation spectrum.

EGFR p.T790M resistance mutations (n=4) were only observed in cases re-biopsied after progression on targeted *EGFR* treatment. In these cases, the originally detected activating *EGFR* mutation (e.g. p.L858R) was always present, while the p.T790M alterations always showed lower VAFs in each case suggesting tumor heterogeneity (5.4-12.2% VAF versus 8.1-88% VAF for activating mutations).

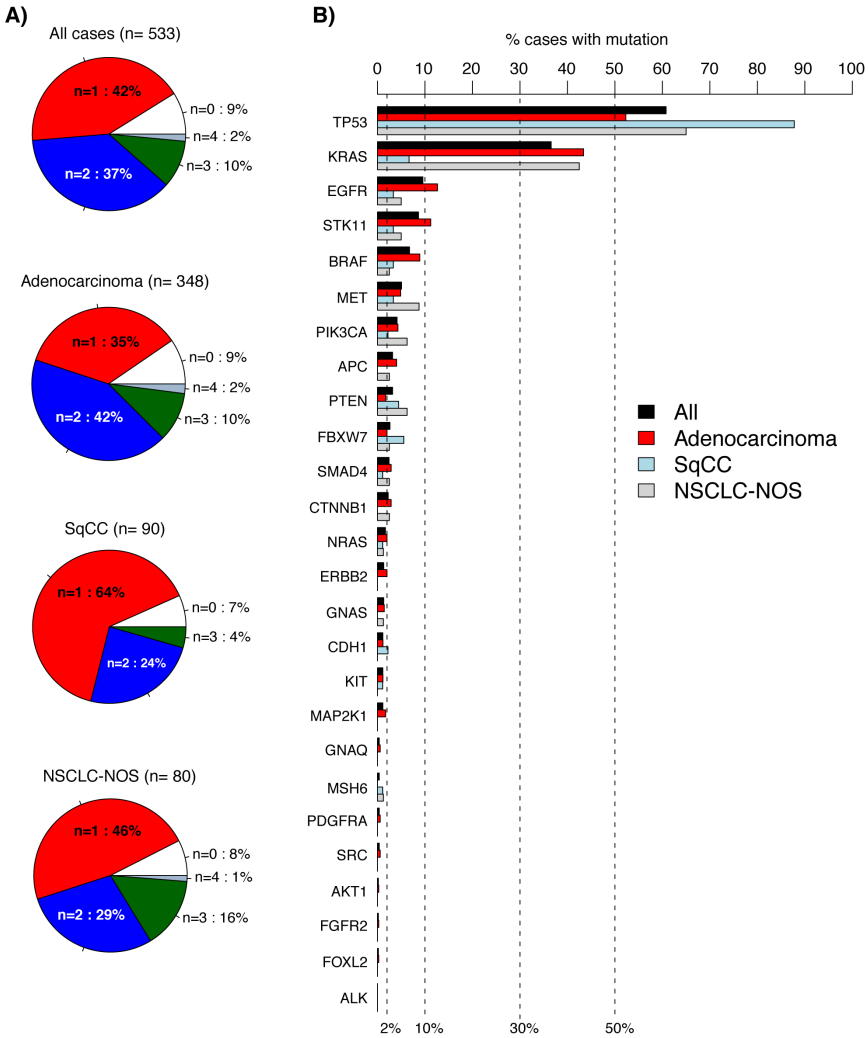


Figure 2: Detected variants in 533 consecutive lung cancers analyzed by the 26-gene TST panel. (A) Pie charts of number of called variants per sample for different sample groups. **(B)** Variant frequency for the analyzed 26 genes across different sample groups (bars). Genes are ordered according to decreasing frequency in the total cohort. In A and B, all detected non-synonymous variants by the vendor supplied analysis pipeline are included.

For *BRAF* (the third most mutated oncogene), a slightly higher variability in detected variants was observed compared to *EGFR* and *KRAS*. This included both codon 600 (38% of *BRAF* variants, 3.7% of adenocarcinomas, and 1.1% of SqCCs specifically) and codon 601 (5.4% of *BRAF* variants) variants known or suggested to be treatment predictive in malignant melanoma, but also variants in codons 466 (11%), 469 (5.4%), and 594 (22% of *BRAF* variants) for which the treatment predictive value to BRAF inhibitors are not fully elucidated (Additional file 4). *BRAF* variants were more often found in older patients (>60 years), with only 13.5% of all detected variants in patients younger than 60 years.

NanoString ALK, RET, and ROS1 gene fusion analysis

The ability of the NanoString assay to identify gene fusions in *ALK*, *RET* and *ROS1* was first successfully validated in four cell lines with known fusion gene rearrangements, HCC78 (*ROS1-SLC34A2*), KARPAS-299 (*ALK-NPM1*), LC-2/ad (*CCDC6-RET*), and H2228 (*EML4-ALK*). Next, gene fusion analysis was performed on RNA from FFPE tissue from 169 lung cancers in the prospective 533-sample cohort, which did not harbor mutations in *EGFR*, *KRAS* or *BRAF* (referred to as triple-negative cases hereon) (Figure 1). NanoString analysis was restricted to cases with FFPE tissue, as the clinical handling of cytology specimens at local pathology departments did not include combined RNA and DNA extraction. Of the 169 analyzed cases (representing 87% of all triple-negative FFPE cases in the 533-sample cohort), 34 hybridizations (20%) were deemed as failures based on too low signals from included housekeeping genes. The failure of these FFPE cases is likely due to extensive RNA degradation in the tissue blocks caused by the fixation process. During 2015, there was no standardization of the time for formalin fixation between different pathology departments in the healthcare region.

Interestingly, the proportion of inclusive NanoString cases was equivalent to the 17% of cases with an inconclusive ALK status by IHC and/or FISH in the total clinical cohort. However, there was no significant association between an inconclusive ALK IHC/FISH call with an inconclusive NanoString call (p=0.78, Fisher's exact test), suggesting that: i) different degradation processes and/or technical issues are in action, and ii) that the methods may complement each other in detecting gene fusion events.

Among the 135 triple-negative cases successfully analyzed by the NanoString assay, gene fusions were detected in ten cases (7.4%); five (3.7%) *ALK* gene fusions (four EML4-ALK_E13:A20 and one EML4-ALK_E6ab:A20 fusion), three (2.2%) *RET* fusions (two *CCDC6-RET_C1:R12* and one novel fusion), and two (1.5%) *ROS1* fusions (*SLC34A2-ROS1_S4:R32* and *SDC4-ROS1_S2:R32*) (Figures 3A-3C). All cases harboring gene fusions were adenocarcinomas, corresponding to 15% of the 67 analyzed triple-negative adenocarcinomas, consistent with the literature [10]. All NanoString called *ALK* and *ROS1* fusions were confirmed by clinical IHC and/or FISH data, and no ALK fusion positive case identified by FISH was missed by the NanoString assay. Three non-adenocarcinoma cases (two SqCC and one large cell neuroendocrine tumor) with positive ALK IHC staining, but inconclusive FISH calls, did not show gene fusions in the NanoString analysis (Figure 3A, pink labeled cases, lower left quadrant, all with tumor cell content >70%). Notably, all three cases showed high expression of both 3' and 5' probes of the ALK tyrosine kinase domain in the NanoString data (indicating lack of gene rearrangement, Figure 3D), suggesting that these ALK IHC stainings could represent false positive gene fusion calls (although treatment data is required to fully confirm such a hypothesis). Interestingly, despite their ALK IHC positive staining none of these three patients have so far received anti ALK therapy in the clinic.

In one sample we detected a probable *RET* fusion through the 3'/5' NanoString ratio not targeted by a fusion specific NanoString probe (Figure 3C, ML00682). Complementary experimental RNA-based NGS analysis (ArcherDX, Boulder, CO, US), performed as previously described [9], identified the suspected fusion to be a *TRIM24-RET* fusion, confirming the NanoString assay's ability to detect also novel fusions.

Co-occurrence of actionable mutations and gene fusions in a consecutive cohort of lung cancers referred to mutation and gene fusion screening

Integration of TST mutation data, ALK IHC/FISH, and NanoString *ALK*, *RET*, and *ROS1* gene fusion analysis for the complete 533-sample prospective cohort is shown in Additional file 5.

To investigate the potential additional clinical yield of a combined TST and multiplexed gene fusion assay (NanoString) compared to the current targeted therapy options in the health care region (EGFR and ALK inhibitor treatment), we analyzed actionable alterations defined from the literature in the 533-sample cohort. First, we defined a set of both acknowledged and proposed actionable oncogene mutations in specific oncogenes (*KRAS*, *EGFR*, *BRAF*, *PIK3CA*, *NRAS*, *ERBB2*, *MAP2K1*, and *AKT1*) in addition to *ALK*, *RET*, and *ROS1* gene fusions, using information from public sources [11] and reported studies [2, 12]. Next, we stratified the 533-sample cohort based on existence of these actionable alterations in individual cases, finding that 54% of all cases, 66% of adenocarcinomas, 13% of SqCCs and 56% of NSCLC-NOS harbored ≥1 alteration (Figure 4A). Of these actionable variants, alterations in *KRAS* dominated in all subgroups, followed by *EGFR*, *BRAF* and *PIK3CA*

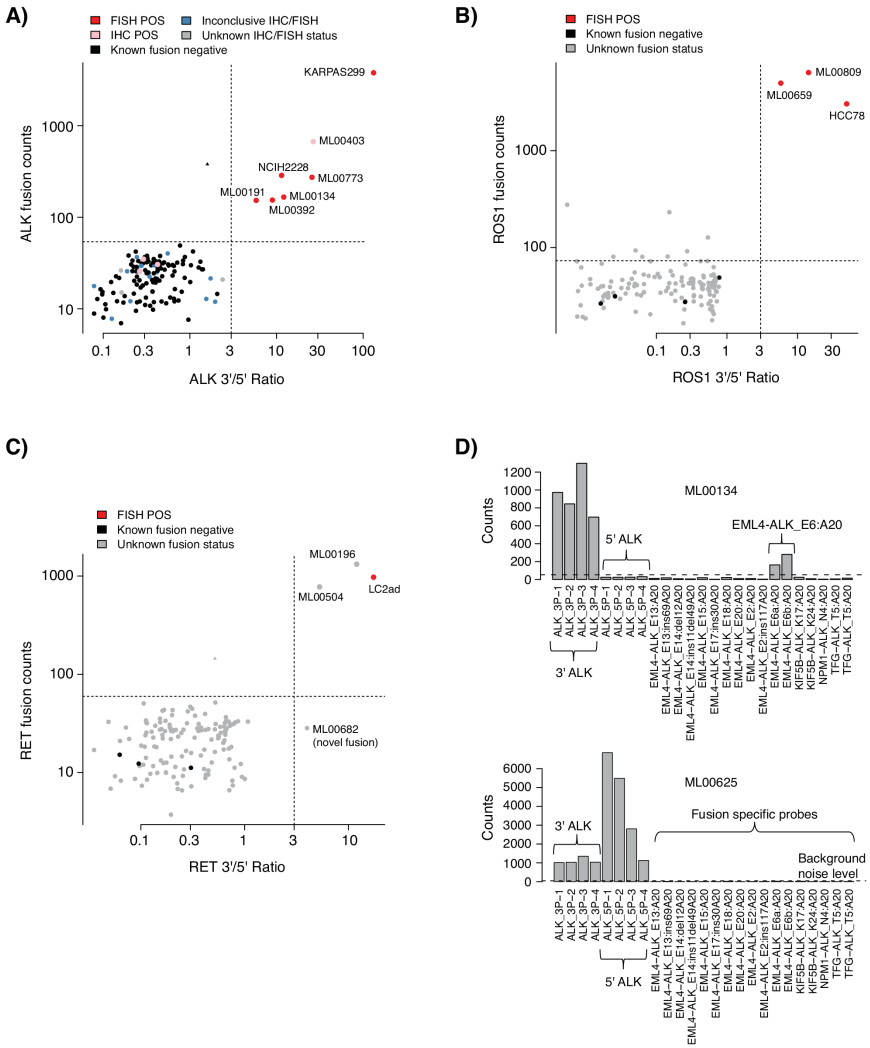


Figure 3: NanoString gene fusion analysis of 135 *EGFR/KRAS/BRAF* mutation negative tumors from a consecutive 533 NSCLC cohort. (A) NanoString *ALK* gene fusion analysis of 135 tumors from the consecutive prospective cohort, and four cancer cell line controls included for reference. **(B)** NanoString *ROS1* gene fusion analysis of the 135 tumors and four cell lines. **(C)** NanoString *RET* gene fusion analysis of the 135 tumors and four cell lines. One sample, ML00682 (lower right quadrant), displays a high 3'/5' ratio but no elevated fusion specific signal, suggesting a fusion not included in the fusion specific probe set. In A to C, analyses were performed as described in Lira et al. [15], using the same thresholds (dotted horizontal and vertical lines). Briefly, based on the dual plotting of a gene fusion specific signal and the 3'/5' expression ratio of probes located around the tyrosine kinase exon, a gene fusion positive case with a known fusion should be located in the upper right quadrant, while a gene fusion positive case without an included fusion specific probe should be located in the lower right quadrant. Negative cases should be located in the lower left quadrant. **(D)** Top panel shows an example of an *ALK* gene fusion positive adenocarcinoma identified by both FISH and NanoString analysis. Bottom panel shows a LCNEC case with a positive ALK IHC call, but inconclusive FISH, that do not display any fusion event based on NanoString analysis. The latter is based on the simultaneously high 3' and 5' expression of probes around the tyrosine kinase exon in the *ALK* gene (case corresponds to a pink labeled sample in panel A).

(Figure 4B). Gene fusions accounted for 7% of actionable alterations in adenocarcinomas. While the majority of detected actionable variants appeared mutually exclusive across samples (Figure 4C), a number of cases showed multiple actionable variants, e.g., concurrent *KRAS* and *PIK3CA* mutations, concurrent *KRAS* and *BRAF* mutations, and concurrent *KRAS*/*EGFR*/*BRAF* mutations and *ALK* fusions. While the two former observations may be explained by tumor subclonality, the high proportion of the latter observation is intriguing given the reported near to mutual exclusiveness of these alterations [10, 13]. Possible explanations may be tumor subclonality, however technical/interpretation issues in the ALK diagnostic scheme cannot be excluded.

Second, we sought to determine the subset of patients with different histological subtypes that could be eligible for potential emerging treatments based on the defined actionable alterations (Figure 4D). In

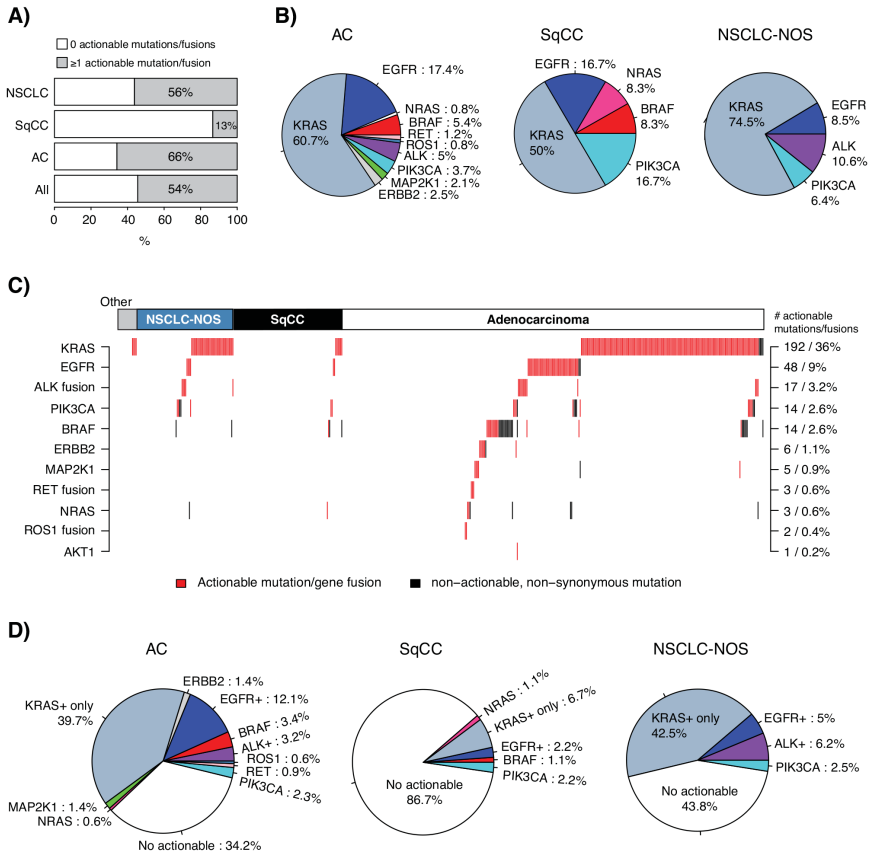


Figure 4: Integration of actionable mutations and gene fusions in the consecutive 533-sample cohort. (A) Proportion of cases with ≥1 actionable alteration in the total 533-sample cohort, adenocarcinomas (AC), SqCCs, and NSCLC-NOS. **(B)** Distribution of detected actionable variants according to the gene in which they fall for adenocarcinomas (AC, n=242 detected variants), SqCCs (n=12 variants) and NSCLC-NOS (n=47 variants). **(C)** Heatmap describing defined actionable and non-actionable non-synonymous variants and gene fusions in investigated genes identified in each case. Each column represents a sample; each row represents a gene. Numbers and proportions displayed on the right axis correspond to the total cohort (533 samples). **(D)** Proportion of cases with actionable mutations in adenocarcinoma (AC, n=348 samples), SqCC (n=90 samples), and NSCLC-NOS (n=80 samples). In each pie chart, EGFR+ corresponds to the proportion of cases with an actionable *EGFR* mutation irrespective of other alterations, ALK+ corresponds to cases with an *ALK* gene fusion irrespective of other alterations, and KRAS+_only corresponds to cases with only an actionable *KRAS* mutation. Consequently, some *BRAF* mutated cases may for instance harbor also an actionable *KRAS* variant. In all panels, not all cases were analyzed for gene fusions by the NanoString assay; consequently these estimates (mainly *ROS1* and *RET*) should be interpreted as low frequency proportions.

adenocarcinoma, this analysis suggested that 10.6% (50.3% if including *KRAS*) of cases could be eligible for emerging targeted treatments, beyond the 15.3% of cases eligible for standard EGFR or ALK targeted therapy (Figure 4D). For SqCC, similar proportions were lower, 4.4% of cases (11.1% if including *KRAS*) could be eligible for emerging treatments, in addition to the 2.2% eligible for EGFR targeted therapy. Finally, for NSCLC-NOS 2.5% of cases (45% with *KRAS*) could be eligible for novel/emerging targeted treatments, in addition to the 11.2% of cases eligible for standard EGFR or ALK targeted therapy.

DISCUSSION

In the era of personalized medicine accurate multi-gene diagnostics is crucial. In the present study, we describe the clinical implementation of an NGS-based diagnostic framework and a parallel validation of a RNA based gene fusion assay for analysis of treatment predictive alterations in a prospective and consecutive clinical testing cohort of mainly advanced NSCLC patients analyzed during a single year within a regional health care region in a Nordic country (Sweden).

Together with several recent reports [2-8, 14-16], our clinical validation and implementation of a commercial DNA amplicon-based NGS assay support the usage of this technique in routine clinical diagnostics of NSCLC compared to previous single gene diagnostics also in smaller regional healthcare regions, based on concordance between techniques (99% in this study), turnaround time, sample success rate over time, accuracy, limit of detection, and cost. In agreement with both Fisher et al. [5] and Hagemann et al. [2], the success rate and clinical feasibility of our NGS framework is highly dependent on central pathological review by experienced diagnostic pathologists together with standardized and quality controlled tissue handling, to ensure sufficiently high proportions of malignant cells in specimens with adequate nucleic acid quality. A challenge for regional/county hospitals may be the bioinformatics aspect of NGS. Using the TST vendor-supplied bioinformatics pipeline we were able to detect and validate by orthogonal methods known activating driver mutations in *EGFR*, *KRAS* and *BRAF* below 5% VAF in cases with ≥10% tumor cell content by routine pathological assessment. This sensitivity was especially important for analysis of *EGFR* p.T790M mutations in patients undergone re-biopsy after progression on first generation EGFR inhibitors. For all such patients, we observed tumor subclonality (inferred based on differences in VAF) between p.T790M mutations and original activating mutations (p.T790M always with a lower VAF). In agreement with Fisher et al. [5], we did observe cases where the vendor-supplied bioinformatics pipeline failed to adequately annotate complicated insertions and deletions in, e.g., *EGFR* (exon 19 deletions and exon 20 insertions), calling for continuous development of these pipelines and/or usage or orthogonal data analysis protocols. Detection of larger insertions and deletions is challenging using amplicon-based techniques, especially in cases with low tumor cell content or tumor subclonality. In our consecutive 533-sample cohort we identified *EGFR* exon 19 deletions down to 8% VAF, and participation in the ESP Lung Quality Assessment Scheme [17] and analysis of samples referred to testing after December 2015 also confirmed detection of *EGFR* exon 20 insertions down to 7% VAF. We believe these findings show that while additional work is needed for challenging indels, vendor supplied analysis pipelines to us appear adequately robust and sensitive for routine clinical use in regional healthcare units lacking strong bioinformatical infrastructure.

The low number of detected variants per sample in this study is consistent with similar targeted NGS-based reports [5, 16] and the gene driver selection process and pan-cancer approach of TST and similar gene panels (like the Ion Torrent AmpliSeq Colon and Lung panel). Ethnicity plays a role in the prevalence of certain genetic markers in NSCLC [18]. For several of the investigated genes (e.g. *TP53*, *PTEN*, *EGFR*, *KRAS*, *ERBB2*) the observed mutation patterns and frequencies in our Swedish cohort agree with previous reports on clinical patient cohorts (predominantly comprising of advanced cancers) of similar ethnicity and/or geographic origin (Scandinavia) [19–25], but also with cohorts consisting of selected non-consecutive patients with operable disease [26, 27]. Alteration frequencies in the NSCLC-NOS subgroup are difficult to interpret and compare, as this subgroup comprises of a mixture of different histological subtypes (a majority is expected to be adenocarcinoma) due to mainly insufficient tissue material (>50% of NSCLC-NOS cases were cytology specimens) that precluded comprehensive histological subtyping by IHC. A few notable discrepancies in our cohort are apparent. For *PIK3CA*, we observe a considerably lower mutation rate (only 2%) in SqCC cases compared with literature reports of 7-16% [16, 19, 26]. The cause of this difference is difficult to determine without extensive comparison of the tested clinical SqCC cohort versus a more population-based cohort from our region. For *BRAF*, we observe a high general mutation frequency in adenocarcinomas (9%), with a 3.7% V600 mutation rate. While the overall mutation frequency is clearly higher compared to some recent studies [19, 28, 29], it is in line with others using e.g. the Ion Torrent AmpliSeq Colon and Lung panel [16]. Consistently, the proportion of specific V600 alterations was slightly higher in our clinical testing cohort than previous literature reports [19, 28, 29] (3.7% versus ~2%).

The main purpose of NGS (and multiplexed gene fusion assays in general) in the clinic is to increase the list of actionable variants for a patient, without increasing the cost, time and tissue requirement compared to serial

single gene testing. In addition, occurrence and/or co-occurrence of mutations in tumor suppressor genes like *TP53* and *STK11* with typical oncogene driver mutations in lung cancer have been suggested to have implications for prognosis and treatment response [30–32], which may be of complimentary clinical value. Our analysis of 533 consecutive NSCLCs screened during a single year showed considerable differences between histological subgroups in the proportion of cases harboring a known or suggested actionable variant, with adenocarcinomas having the greatest potential benefit from this type of analyses (Figure 4). We acknowledge that inclusion of *KRAS* as an actionable gene in this type of analysis is not unproblematic (and hence we report frequencies with and without *KRAS*). For SqCC, results must be interpreted with great care given the small number of cases and individual mutations. Irrespectively, despite the individually low frequency of many potentially actionable variants defined in the current study (e.g. *ERBB2*, *BRAF*, *RET*, *ROS1*, *PIK3CA*), the high incidence of lung cancer implies that a large population worldwide is affected, supporting clinical trials or routine molecular screening programs in the disease [19].

ROS1 gene fusions have been shown to be treatment predictive for ALK-inhibitor drugs [33]. In our investigated cohort of triple-negative NSCLCs (*EGFR*, *KRAS*, and *BRAF* mutation negative) the number of cases with either *ROS1* or *RET* fusions were similar to that of ALK-positive cases, supporting the need of multiplexed gene fusion diagnostics in NSCLC and adenocarcinoma specifically. For the non-adenocarcinoma cases that were ALK positive by IHC in the triple-negative cohort, NanoString analysis suggested overexpression of the entire gene by some other mechanism than gene rearrangement (see, e.g., Figure 3D). This more detailed view of gene fusion events supports the usage of multiplexed methods like NanoString as a complementary orthogonal method, or even replacement, for IHC/FISH when possible. Moreover, due to the flexibility and capacity of the NanoString technology, additional gene fusions as well as *MET* exon 14 skipping events can easily be added in a design update (see e.g. [34]). Finally, the experimental TAT for the NanoString assay may be very short, potentially down to three working days including nucleic acid extraction (Additional file 2).

While the introduction of NGS in the clinical setting represents a major leap forward; current commercial amplicon-based panels (e.g. the TST and Ion Torrent Ampliseq Colon and Lung panels) are biased towards analyzing hotspot alterations in a limited set of oncogenes often selected through a pan-cancer approach. To some extent, these panels offer the possibility to detect intrinsic or acquired resistance mechanisms to targeted treatment, mainly p.T790M (as shown in this study) and p.C797S mutations in *EGFR* (first to third generation inhibitors) and specific gatekeeper mutations in *ALK* (like p.L1196M and p.G1269A), in patients re-biopsied after treatment failure. However, most panels (including TST and frequently used Ion Torrent panels) are less well suited to detect EGFR/ALK resistance mechanisms caused by alterations in other genes. Here, panel design and size constraints, but also problems in calling copy number alterations (like *MET* amplification as a mechanism of resistance to EGFR inhibitors) reliably in tumors with considerable non-malignant infiltration represent limiting factors. Therefore, diagnostic platforms based on, e.g., hybrid capture methods of either DNA alone (see e.g. [35, 36]) or DNA and RNA combinations (e.g. the Illumina TruSight Tumor 170 panel and the AmpliSeq based Thermo Fisher Oncomine™ Focus/Comprehensive panels) that allow considerably more sequence to be analyzed could be the next preferable step also outside large comprehensive cancer centers. These assays could allow simultaneous detection of mutations, gene fusions, and copy number alterations (like drug targetable *MET* and *FGFR1* amplifications) in a large number of genes. However, considering the observed failure rate of 20% for the NanoString RNA gene fusion assay in our prospective clinical samples, it remains to be shown that RNAseq approaches can do better in daily clinical practice. Finally, while tissue-based diagnostics is the cornerstone of diagnostic tumor pathology today, less invasive sampling methods like blood-based assays targeting e.g. circulating tumor DNA, or analysis of exhaled breath condensates [37, 38] are increasingly gaining interest as they could facilitate a more active and less invasive treatment monitoring. However, these applications may require more sensitive sequencing techniques, different logistics, and optimized sample preparations than presently used in most local diagnostic pathology departments.

MATERIALS AND METHODS

Ethics statement

The study was approved by the Regional Ethical Review Board in Lund, Sweden (Registration no. 2014/748 and 2015/575). By decision of the Ethical Review Board, specific written informed consent from patients were not required. No personal data was used for this study. In accordance with the decision of the Ethical Review Board, patients were informed about the study through local advertisement in news media in the region.

Tumor validation cohort

A tumor cohort comprising of 40 NSCLCs, 22 CMMs, and 19 colon cancers with available *BRAF*, *KRAS*, *NRAS*, and/or *EGFR* mutational data from routine clinical analysis within the southern health care region of Sweden using single gene assays (see below) were collected (Table 1). Tissue types from included patient tumors

included routine FFPE sections from resected material or needle biopsies (typically 6x5um sections), sections from cytology cellblocks (typically 10x5um or 10x10um sections), or DNA extracted from cytology slides (see below).

Consecutive prospective tumor cohort

In the southern Swedish health care region (comprising close to 1.8 million inhabitants), ~800 new lung cancer cases (of any stage and histology) are identified annually. A consecutive prospective clinical testing cohort of 533 lung cancers, representing 526 unique patients, subjected to routine NGS-based mutational analysis within the southern health care region of Sweden, including two university-affiliated and four regional pathology departments, with additional samples from a third university-affiliated pathology department outside the healthcare region were collected between January 7 to December 31 2015 (Table 1, Figure 1). During this period a reflex testing procedure was allowed in the health care region, meaning that all lung cancers that were not SCLC or carcinoids could be sent for clinical mutation testing, including also some early stage tumors. All cases were analyzed at an established central NGS laboratory within the Division of Oncology and Pathology, Department of Clinical Sciences Lund, Lund University through the Center for Molecular Diagnostics (www.skane.se/cmd). All identified mutations were reported back to the diagnostic pathologists in a molecular report by the central NGS laboratory. During the time period, only actionable mutations in *EGFR* and hotspot mutations in *KRAS* were included in the final clinical report (signed by a diagnostic pathologist) returned to the treating clinician. Data on lung cancer histology and tumor stage was obtained from patient charts and in accordance with the classification scheme used at the time of diagnosis. Tissue sources included primary lung tumors, lymph node metastases, or extranodal distant metastases. Sample types from included patient tumors were either tissue blocks or cytology specimens.

Tissue selection for routine clinical mutation analysis

Tumor morphology was determined by the clinical pathologist. In cases with apparent keratinizing SqCC, IHC was normally not performed. In case of morphologically apparent adenocarcinoma, the standard immunohistochemical panel included at least TTF-1, while in case of NSCLC without clear morphology the panel included at least TTF-1 and either CK5 or p40. If these markers were negative, further stains including CK7 or a broad cytokeratin were performed. Also, the morphological appearance, patient history and clinical and radiological findings guided the initial selection of stainings. Neuroendocrine markers were added in cases with neuroendocrine morphology. If a diagnosis of primary lung cancer was uncertain, or if the DNA content and/or quality was to low for NGS-analysis (requiring real-time PCR analysis), the case was excluded from this study (Figure 1). During the study period, encountered cases of pulmonary adenocarcinoma, SqCC, adenosquamous carcinoma, sarcomatoid carcinoma, NSCLC-NOS, large cell carcinoma and LCNEC based on biopsy or cytology were unselectively tested for predictive molecular alterations.

The suitability of a material for mutational analysis was assessed by a pathologist based on hematoxylin and eosin (H&E) stains of archived FFPE tissue blocks and / or cytology specimens. A representative area with high frequency of malignant cells was identified, from which sections for mutational analysis was taken followed by new H&E sections to ensure that a representative material had been taken. An estimate of tumor cell content was made by a diagnostic pathologist, with a requirement of ≥10% for the mutational analysis. In addition to FFPE tissue blocks, tissue material for mutation analysis could also originate from cytology slides, or sections from centrifuged and paraffin embedded cytology material (cell blocks). Sections were stored at -20°C until nucleic acid extraction, due to logistical batching with frozen DNA aliquots from cytology specimens.

In case of preparation of cell lysate from cytology slides, a representative tumor cell rich area of a cytology slide was identified, the slide was scanned (to enable future clinical review), and the glass cover slip was removed using xylene followed by a rehydration step in ethanol. Thereafter, the cells were lysed using 180ul ATL Buffer from Qiagen (Qiagen, Hilden, Germany). DNA was extracted from the lysate within 24h and stored at -20°C.

DNA and RNA extraction

DNA and RNA for NGS-based mutation analysis and NanoString (Seattle, WA, US) gene fusion analysis were extracted using the Qiagen AllPrep kit for FFPE tissue and automated on the QIAcube instrument (Qiagen). The protocol was modified with an extended proteinase K digestion (overnight) for the DNA extraction to obtain higher DNA yields. DNA from cytology slides was extracted using the QiaAmp DNA Micro kit (Qiagen). RNA was not extracted from cytology specimens, as these extractions were not performed at the central NGS laboratory in contrast to FFPE extractions.

Mutational validation techniques

Mutational status for hotspot mutations in *NRAS*, *KRAS*, *BRAF*, and *EGFR* were obtained for the validation cohort using pyrosequencing for *EGFR* or real-time PCR for *KRAS* and *NRAS* (Entrogen, Woodland Hills, CA,

US) and *BRAF* (Qiagen RGQ Therascreen®) performed and validated for routine clinical diagnostics within the health care region (Region Skåne, Sweden). Independent validation analysis of NGS results for: i) samples with very low variant allele frequencies (VAFs) in *EGFR*, *KRAS*, or *BRAF* (VAF <5%) or ii) randomly selected *EGFR* and *KRAS* mutation negative cases on a regular basis was performed using Qiagen Therascreen® RGQ PCR Kits for *EGFR*, *KRAS*, and *BRAF* according to the manufacturer's protocol.

NGS-based mutational analysis

NGS-based mutation analysis was performed using the Illumina TST panel on a MiSeq instrument according to manufacturer's instructions (Illumina, San Diego, CA, US). The TST panel is an exon-focused panel, allowing theoretical identification of all variants in screened exons, opposed to a specific hotspot mutation panel. Analyzed regions included a selected set of complete exons in 26 genes: *AKT1* (exon 2), *ALK* (exon 23), *APC* (exon 15), *BRAF* (exons 11, 15), *CDH1* (exons 8, 9, 12), *CTNNB1* (exon 2), *EGFR* (exons 18, 19, 20, 21), *ERBB2* (exon 20), *FBXW7* (exons 7, 8, 9, 10, 11), *FGFR2* (exon 6), *FOXL2* (exon 1), *GNAQ* (exons 4, 5, 6), *GNAS* (exons 6, 8), *KIT* (exons 9, 11, 13, 17, 18), *KRAS* (exons 1, 2, 3, 4), *MAP2K1* (exon 2), *MET* (exons 1, 4, 13, 15, 16, 17, 18, 20), *MSH6* (exons 5), *NRAS* (exons 1, 2, 3, 4), *PDGFRA* (exons 11, 13, 17), *PIK3CA* (exons 1, 2, 7, 9, 20), *PTEN* (exons 1, 2, 3, 4, 5, 6, 7, 9), *SMAD4* (exons 8, 11), *SRC* (exon 10), *STK11* (exons 1, 4, 6, 8), and *TP53* (exons 2, 3, 4, 5, 6, 7, 8, 9, 10, 11). Prior to library preparation a quality control qPCR assay was performed as described in the TST instructions. In the TST assay, sample DNA amount is not fixed. Instead, the quality control assay determines a sample volume used as assay input (maximum of 2x10ul) based on a calculated delta Ct value (higher values implies poorer DNA). Thus, actual DNA input in the NGS assay may vary dramatically between samples of high quality (e.g. 16x dilution) to samples with low quality (no dilution). Routinely, samples with a quality score of 7-8 could be analyzed by NGS (recommended Illumina upper threshold was 6). Samples with higher delta Ct scores were directly assayed by the Qiagen EGFR Therascreen assay to reduce the number of inconclusive NGS runs. Four to six samples were multiplexed using the Illumina V2 sequencing chemistry, while 7-12 samples were multiplexed if using the V3 sequencing chemistry. Alignment, quality filtering, variant calling, and variant annotation were performed as described [9], using the vendor supplied data analysis pipeline. Base coverage >1000X were used as a sequencing quality control threshold for variant calling. Limit of detection for variants were not fixed in percentage, as the main variant filtering step in the variant calling was the requirement of occurrence of a variant in both TST library pools for a sample (see www.illumina.com for explanation of the bidirectional design of the TST assay). Effectively, a limit of detection of 4% were set in the clinical context, as all such hotspot variants in *EGFR*, *BRAF*, and *KRAS* could be validated by a real-time PCR assay. Detected *TP53* variants were screened against the IARC database [39], with no variants being annotated as a known polymorphism, and 95% of annotated variants being considered as deleterious by both the AVGVD and SIFT prediction tools.

Actionable mutations (defined in [2, 11, 12]) in *KRAS* (codon 12, 13, and 61 variants), *EGFR* (exon 19 deletion, exon 20 insertion, T790M, codon 719 (exon 18), 858, and 861 variants), *BRAF* (codon 600 variants), *PIK3CA* (codon 542, 545, 1047, and 1047 variants), *NRAS* (codon 12, 13, and 61), *ERBB2* (exon 20 insertions), *MAP2K1* (codon 56 and 57 variants), and *AKT1* (L52R variant) and gene fusions in *ALK*, *RET*, and *ROS1* were summarized for each sample.

NanoString gene fusion analysis

Analysis of *ALK*, *RET*, and *ROS1* gene fusions in FFPE tissue were performed using a RNA-based NanoString nCounter Elements assay. For each gene, a probe set was designed using the approach reported in Lira et al. [15] using two sequence-specific probe cocktails consisting of a mixture of 5' capture and 3' reporter probes with a target specific sequence. In addition to the 3' 5' approach, fusion specific target probes spanning the known exon-exon junction of fusion transcripts in the *ALK*, *RET* and *ROS1* genes were added based on the toehold approach established by NanoString and reported by Zhang et al. [40] (see this study for exact listing of specific fusions analyzed). This dual design allows gene fusions to be detected by the 3'/5' ratio difference alone (if the specific gene fusions is not included among the toehold designed probes), or by both the 3' 5' probes and expression of a specific toehold probe (see Lira et al. [15] for details). All probes where synthesized by Integrated DNA Technology (IDT Inc., Coralville, USA). A RNA pool of the HCC78 (*ROS1-SLC34A2*), KARPAS-299 (*ALK-NPM1*), LC-2/ad (*CCDC6-RET*), and H2228 (*EML4-ALK*) cell lines were used as controls on all NanoString Elements cartridges, and prepared as described [9]. 100-250 ng of total RNA was hybridized for each sample for 24h at 67°C. Data analysis, including background correction, scaling based on positive controls, calculation of 3'/5' fusion ratios, and thresholds for calling gene fusions were performed/used as described by Lira et al. [15] using the R statistical language [41]. An analysis was called as failure if its H/H_i ratio as described by Lira et al. [42] was >8. In a series of dilution experiments using clinical tumors with different gene fusions and pathologically estimated tumor percentages, we estimated the limit of detection to be at least 5% for the NanoString

assay, i.e., the assay may detect a fusion in a sample with ≥5% tumor cells mainly due to the order of expression of a gene fusion on the RNA level.

ALK and ROS1 IHC and/or FISH analyses

ALK IHC and/or *ALK* FISH data was available for 98.2% of all NanoString analyzed cases as part of the routine clinical diagnostic scheme in lung cancer within the health care region. ALK IHC was performed using the D5F3 antibody (Ventana Medical Systems, Tucson, AZ, US), and *ALK* FISH analysis using the Vysis ALK break apart FISH probe (Vysis, Abbot molecular, Des Plaines, IL, US) according to manufacturers' instructions. *ROS1* gene fusions were validated using the Vysis ROS1 break apart FISH probe according to manufacturers' instructions. *RET* fusions were not validated by FISH, as no validated assay was available in collaborating pathology departments.

CONCLUSIONS

The present study describes a clinical implementation of NGS-based diagnostics for analysis of treatment predictive mutations in NSCLC, demonstrating that such methods can be incorporated into daily clinical practice in regional healthcare regions with constraints in budget, personnel and infrastructure. Although mutation profiles in our prospective Swedish cohort, comprising mainly of advanced stage patients, does not differ considerably from other Western patients some differences exist. Importantly, multiplexed gene diagnostics provide information for both current and emerging treatments, as well as insights into mechanisms of treatment resistance to targeted therapy. In order to allow a more personalized cancer care for lung cancer patients, innovative clinical trials and programs should take advantage of improvements in clinical diagnostics through these multigene assays to determine their actual clinical benefit.

Abbreviations

CMM: cutaneous malignant melanoma; FFPE: formalin-fixed paraffin embedded; FISH: fluorescence *in situ* hybridization; LCNEC: large cell neuroendocrine carcinoma; NGS: next generation sequencing; NSCLC: non-small cell lung cancer; NSCLC-NOS: non-small cell lung cancer not otherwise specified; SCLC: small cell lung cancer; SqCC: squamous cell carcinoma; TAT: turnaround time; TST: TruSight Tumor; VAF: variant allele frequency

CONFLICTS OF INTEREST

Authors declare that they have no competing interests.

FUNDING

Financial support for this study was provided by the Swedish Cancer Society, the Mrs Berta Kamprad Foundation, the Gunnar Nilsson Cancer Foundation, the Crafoord Foundation, BioCARE a Strategic Research Program at Lund University, governmental funding (ALF), and the Gustav V:s Jubilee Foundation.

Authors' contributions

JS, AK and MP conceived of the study. AK, JS, KH, KA, MJ, FR, and CR performed the mutational analyzes and gene fusion experiments. KJ, KEL, PL, HB provided validation samples, pathological evaluations and clinical data. AK and JS performed statistical analyses. MP, JK, and LE included patients. JS and AK drafted and wrote the manuscript with support from KEL, MP, and GJ. The study was methodologically and scientifically supported by ÅB. All authors approved the final manuscript.

REFERENCES

1. Popper HH, Ryska A, Timar J, Olszewski W. Molecular testing in lung cancer in the era of precision medicine. Translational lung cancer research. 2014; 3:291-300.
2. Hagemann IS, Devarakonda S, Lockwood CM, Spencer DH, Guebert K, Bredemeyer AJ, Al-Kateb H, Nguyen TT, Duncavage EJ, Cottrell CE, Kulkarni S, Nagarajan R, Seibert K, et al. Clinical next-generation sequencing in patients with non-small cell lung cancer. Cancer. 2015; 121:631-639.
3. Scarpa A, Sikora K, Fassan M, Rachiglio AM, Cappellesso R, Antonello D, Amato E, Mafficini A, Lambiase M, Esposito C, Bria E, Simonato F, Scardoni M, et al. Molecular typing of lung adenocarcinoma on cytological samples using a multigene next generation sequencing panel. PLoS ONE. 2013; 8:e80478.
4. Boland GM, Piha-Paul SA, Subbiah V, Routbort M, Herbrich SM, Baggerly K, Patel KP, Brusco L, Horombe C, Naing A, Fu S, Hong DS, Janku F, et al. Clinical next generation sequencing to identify actionable aberrations in a phase I program. Oncotarget. 2015; 6:20099-20110.
5. Fisher KE, Zhang L, Wang J, Smith GH, Newman S, Schneider TM, Pillai RN, Kudchadkar RR, Owonikoko TK, Ramalingam SS, Lawson DH, Delman KA, El-Rayes BF, et al. Clinical Validation and Implementation of a Targeted Next-Generation Sequencing Assay to Detect Somatic Variants in Non-Small Cell Lung, Melanoma, and Gastrointestinal Malignancies. J Mol Diagn. 2016; 18:299-315.
6. Sie D, Snijders PJ, Meijer GA, Doeleman MW, van Moorsel MI, van Essen HF, Eijk PP, Grunberg K, van Grieken NC, Thunnissen E, Verheul HM, Smit EF, Ylstra B, et al. Performance of amplicon-based next generation

DNA sequencing for diagnostic gene mutation profiling in oncopathology. Cell Oncol (Dordr). 2014; 37:353-361.

7. Endris V, Penzel R, Warth A, Muckenhuber A, Schirmacher P, Stenzinger A, Weichert W. Molecular diagnostic profiling of lung cancer specimens with a semiconductor-based massive parallel sequencing approach: feasibility, costs, and performance compared with conventional sequencing. J Mol Diagn. 2013; 15:765-775.
8. Cottrell CE, Al-Kateb H, Bredemeyer AJ, Duncavage EJ, Spencer DH, Abel HJ, Lockwood CM, Hagemann IS, O'Guin SM, Burcea LC, Sawyer CS, Oschwald DM, Stratman JL, et al. Validation of a next-generation sequencing assay for clinical molecular oncology. J Mol Diagn. 2014; 16:89-105.
9. Karlsson A, Brunnstrom H, Lindquist KE, Jirstrom K, Jonsson M, Rosengren F, Reutersward C, Cirenajwis H, Borg A, Jonsson P, Planck M, Jonsson G, Staaf J. Mutational and gene fusion analyses of primary large cell and large cell neuroendocrine lung cancer. Oncotarget. 2015; 6:22028-22037.
10. Takeuchi K, Soda M, Togashi Y, Suzuki R, Sakata S, Hatano S, Asaka R, Hamanaka W, Ninomiya H, Uehara H, Lim Choi Y, Satoh Y, Okumura S, et al. RET, ROS1 and ALK fusions in lung cancer. Nature medicine. 2012; 18:378-381.
11. My Cancer Genome. https://www.mycancergenome.org. Accessed 15 April, 2016.
12. Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. Nature. 2012; 490:61-70.
13. Gainor JF, Varghese AM, Ou SH, Kabraji S, Awad MM, Katayama R, Pawlak A, Mino-Kenudson M, Yeap BY, Riely GJ, Iafrate AJ, Arcila ME, Ladanyi M, et al. ALK rearrangements are mutually exclusive with mutations in EGFR or KRAS: an analysis of 1,683 patients with non-small cell lung cancer. Clin Cancer Res. 2013; 19:4273-4281.
14. Takeda M, Sakai K, Terashima M, Kaneda H, Hayashi H, Tanaka K, Okamoto K, Takahama T, Yoshida T, Iwasa T, Shimizu T, Nonagase Y, Kudo K, et al. Clinical application of amplicon-based next-generation sequencing to therapeutic decision making in lung cancer. Ann Oncol. 2015; 26:2477-2482.
15. Lira ME, Choi YL, Lim SM, Deng S, Huang D, Ozeck M, Han J, Jeong JY, Shim HS, Cho BC, Kim J, Ahn MJ, Mao M. A single-tube multiplexed assay for detecting ALK, ROS1, and RET fusions in lung cancer. J Mol Diagn. 2014; 16:229-243.
16. Maki-Nevala S, Sarhadi VK, Ronty M, Kettunen E, Husgafvel-Pursiainen K, Wolff H, Knuuttila A, Knuutila S. Hot spot mutations in Finnish non-small cell lung cancers. Lung Cancer. 2016; 99:102-110.
17. ESP Lung External Quality Assessment Scheme. http://lung.eqascheme.org/. Accessed 20 March, 2016.
18. Midha A, Dearden S, McCormack R. EGFR mutation incidence in non-small-cell lung cancer of adenocarcinoma histology: a systematic review and global map by ethnicity (mutMapII). Am J Cancer Res. 2015; 5:2892-2911.
19. Barlesi F, Mazieres J, Merlio JP, Debieuvre D, Mosser J, Lena H, Ouafik L, Besse B, Rouquette I, Westeel V, Escande F, Monnet I, Lemoine A, et al. Routine molecular profiling of patients with advanced non-small-cell lung cancer: results of a 1-year nationwide programme of the French Cooperative Thoracic Intergroup (IFCT). Lancet. 2016; 387:1415-1426.
20. Sandelin M, Berglund A, Sundstrom M, Micke P, Ekman S, Bergqvist M, Bergstrom S, Koyi H, Branden E, Janson C, Botling J. Patients with Non-small Cell Lung Cancer Analyzed for EGFR: Adherence to Guidelines, Prevalence and Outcome. Anticancer Res. 2015; 35:3979-3985.
21. Gahr S, Stoehr R, Geissinger E, Ficker JH, Brueckl WM, Gschwendtner A, Gattenloehner S, Fuchs FS, Schulz C, Rieker RJ, Hartmann A, Ruemmele P, Dietmaier W. EGFR mutational status in a large series of Caucasian European NSCLC patients: data from daily practice. British journal of cancer. 2013; 109:1821-1828.
22. Maki-Nevala S, Ronty M, Morel M, Gomez M, Dawson Z, Sarhadi VK, Telaranta-Keerie A, Knuuttila A, Knuutila S. Epidermal growth factor receptor mutations in 510 Finnish non--small-cell lung cancer patients. J Thorac Oncol. 2014; 9:886-891.
23. Weber B, Hager H, Sorensen BS, McCulloch T, Mellemgaard A, Khalil AA, Nexo E, Meldgaard P. EGFR mutation frequency and effectiveness of erlotinib: a prospective observational study in Danish patients with non-small cell lung cancer. Lung Cancer. 2014; 83:224-230.
24. Berg J, Fjellbirkeland L, Suhrke P, Jebsen P, Lund-Iversen M, Kleinberg L, Helgeland L, Brustugun OT, Helland A. EGFR mutation testing of lung cancer patients - Experiences from Vestfold Hospital Trust. Acta oncologica (Stockholm, Sweden). 2016; 55:149-155.
25. Smits AJ, Kummer JA, Hinrichs JW, Herder GJ, Scheidel-Jacobse KC, Jiwa NM, Ruijter TE, Nooijen PT, Looijen-Salamon MG, Ligtenberg MJ, Thunnissen FB, Heideman DA, de Weger RA, et al. EGFR and KRAS mutations in lung carcinomas in the Dutch population: increased EGFR mutation frequency in malignant pleural effusion of lung adenocarcinoma. Cell Oncol (Dordr). 2012; 35:189-196.
26. Hammerman PS, Lawrence MS, Voet D, Jing R, Cibulskis K, Sivachenko A, Stojanov P, McKenna A, Lander ES, Gabriel S, Getz G, Sougnez C, Imielinski M, et al. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012; 489:519-525.
27. The Cancer Genome Atlas Network A. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014; 511:543-550.
28. Brustugun OT, Khattak AM, Tromborg AK, Beigi M, Beiske K, Lund-Iversen M, Helland A. BRAF-mutations in non-small cell lung cancer. Lung Cancer. 2014; 84:36-38.

29. Tissot C, Couraud S, Tanguy R, Bringuier PP, Girard N, Souquet PJ. Clinical characteristics and outcome of patients with lung cancer harboring BRAF mutations. Lung Cancer. 2016; 91:23-28.
30. Pecuchet N, Laurent-Puig P, Mansuet-Lupo A, Legras A, Alifano M, Pallier K, Didelot A, Gibault L, Danel C, Just PA, Riquet M, Le Pimpec-Barthes F, Damotte D, et al. Different prognostic impact of STK11 mutations in non-squamous non-small-cell lung cancer. Oncotarget. 2015. doi: 10.18632/oncotarget.6379.
31. Huang S, Benavente S, Armstrong EA, Li C, Wheeler DL, Harari PM. p53 modulates acquired resistance to EGFR inhibitors and radiation. Cancer research. 2011; 71:7071-7079.
32. Ma X, Le Teuff G, Lacas B, Tsao MS, Graziano S, Pignon JP, Douillard JY, Le Chevalier T, Seymour L, Filipits M, Pirker R, Janne PA, Shepherd FA, et al. Prognostic and Predictive Effect of TP53 Mutations in Patients with Non-Small Cell Lung Cancer from Adjuvant Cisplatin-Based Therapy Randomized Trials: A LACE-Bio Pooled Analysis. J Thorac Oncol. 2016; 11:850-861.
33. Shaw AT, Ou SH, Bang YJ, Camidge DR, Solomon BJ, Salgia R, Riely GJ, Varella-Garcia M, Shapiro GI, Costa DB, Doebele RC, Le LP, Zheng Z, et al. Crizotinib in ROS1-rearranged non-small-cell lung cancer. The New England journal of medicine. 2014; 371:1963-1971.
34. Sunami K, Furuta K, Tsuta K, Sasada S, Izumo T, Nakaoku T, Shimada Y, Saito M, Nokihara H, Watanabe S, Ohe Y, Kohno T. Multiplex Diagnosis of Oncogenic Fusion and MET Exon Skipping by Molecular Counting Using Formalin-Fixed Paraffin Embedded Lung Adenocarcinoma Tissues. J Thorac Oncol. 2016; 11:203-212.
35. Frampton GM, Fichtenholtz A, Otto GA, Wang K, Downing SR, He J, Schnall-Levin M, White J, Sanford EM, An P, Sun J, Juhn F, Brennan K, et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nature biotechnology. 2013; 31:1023-1031.
36. Cheng DT, Mitchell TN, Zehir A, Shah RH, Benayed R, Syed A, Chandramohan R, Liu ZY, Won HH, Scott SN, Brannon AR, O'Reilly C, Sadowska J, et al. Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. J Mol Diagn. 2015; 17:251-264.
37. Youssef O, Knuuttila A, Piirila P, Bohling T, Sarhadi V, Knuutila S. Presence of cancer-associated mutations in exhaled breath condensates of healthy individuals by next generation sequencing. Oncotarget. 2017. doi: 10.18632/oncotarget.15233.
38. Youssef O, Sarhadi VK, Armengol G, Piirila P, Knuuttila A, Knuutila S. Exhaled breath condensate as a source of biomarkers for lung carcinomas. A focus on genetic and epigenetic markers-A mini-review. Genes Chromosomes Cancer. 2016; 55:905-914.
39. IARC TP53 Database. http://www-p53.iarc.fr. Accessed April 1, 2016.
40. Zhang DY, Chen SX, Yin P. Optimizing the specificity of nucleic acid hybridization. Nat Chem. 2012; 4:208-214.
41. The R Project for Statistical Computing. http://www.r-project.org. Accessed 2016.
42. Lira ME, Kim TM, Huang D, Deng S, Koh Y, Jang B, Go H, Lee SH, Chung DH, Kim WH, Schoenmakers EF, Choi YL, Park K, et al. Multiplexed gene expression and fusion transcript analysis to detect ALK fusions in lung cancer. J Mol Diagn. 2013; 15:51-61.

Study II

Study III

ORIGINAL ARTICLE

Gene Expression Profiling of Large Cell Lung Cancer Links Transcriptional Phenotypes to the New Histological WHO 2015 Classification

Anna Karlsson, MSc,[a] Hans Brunnström, MD, PhD,[b,c] Patrick Micke, MD, PhD,[d] Srinivas Veerla, PhD,[a] Johanna Mattsson, PhD,[d] Linnea La Fleur, MSc,[d] Johan Botling, MD, PhD,[d] Mats Jönsson, PhD,[a] Christel Reuterswärd, MSc,[a] Maria Planck, MD, PhD,[a,e] Johan Staaf, PhD[a,*]

[a]*Division of Oncology and Pathology, Department of Clinical Sciences Lund, Lund University, Lund, Sweden*
[b]*Division of Oncology and Pathology, Department of Clinical Sciences Lund, Lund University, Lund, Sweden*
[c]*Department of Pathology, Regional Laboratories Region Skåne, Lund, Sweden*
[d]*Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden*
[e]*Department of Respiratory Medicine and Allergology, Skåne University Hospital, Lund, Sweden*

Received 1 March 2017; revised 26 April 2017; accepted 12 May 2017
Available online - 20 May 2017

ABSTRACT

Introduction: Large cell lung cancer (LCLC) and large cell neuroendocrine carcinoma (LCNEC) constitute a small proportion of NSCLC. The WHO 2015 classification guidelines changed the definition of the debated histological subtype LCLC to be based on immunomarkers for adenocarcinoma and squamous cancer. We sought to determine whether these new guidelines also translate into the transcriptional landscape of lung cancer, and LCLC specifically.

Methods: Gene expression profiling was performed by using Illumina V4 HT12 microarrays (Illumina, San Diego, CA) on samples from 159 cases (comprising all histological subtypes, including 10 classified as LCLC WHO 2015 and 14 classified as LCNEC according to the WHO 2015 guidelines), with complimentary mutational and immunohistochemical data. Derived transcriptional phenotypes were validated in 199 independent tumors, including six WHO 2015 LCLCs and five LCNECs.

Results: Unsupervised analysis of gene expression data identified a phenotype comprising 90% of WHO 2015 LCLC tumors, with characteristics of poorly differentiated proliferative cancer, a 90% tumor protein p53 gene (*TP53)* mutation rate, and lack of well-known NSCLC oncogene driver alterations. Validation in independent data confirmed aggregation of WHO 2015 LCLCs in the specific phenotype. For LCNEC tumors, the unsupervised gene expression analysis suggested two different transcriptional patterns corresponding to a proposed genetic division of LCNEC tumors into SCLC-like and NSCLC-like cancer on the basis of *TP53* and retinoblastoma 1 gene (*RB1*) alteration patterns.

Conclusions: Refined classification of LCLC has implications for diagnosis, prognostics, and therapy decisions. Our molecular analyses support the WHO 2015 classification of LCLC and LCNEC tumors, which herein follow different tumorigenic paths and can accordingly be stratified into different transcriptional subgroups, thus linking diagnostic immunohistochemical staining–driven classification with the transcriptional landscape of lung cancer.

Keywords: Lung cancer; Large cell lung carcinoma; LCNEC; Mutation; Gene expression; WHO classification

Introduction

NSCLC accounts for most lung cancers and is dominated by the histological subtypes adenocarcinoma, squamous cell carcinoma (SqCC), and large cell

*Corresponding author.

Disclosure: The authors declare no conflict of interest.

Address for correspondence: Johan Staaf, PhD, Division of Oncology and Pathology, Department of Clinical Sciences Lund, Lund University, Medicon Village, SE 22381 Lund, Sweden. E-mail: johan.staaf@med.lu.se

ISSN: 1556-0864
http://dx.doi.org/10.1016/j.jtho.2017.05.008

carcinoma with or without neuroendocrine features (LCNEC and LCLC, respectively) LCLC and LCNEC together account for 2% to 3% of all cases depending on cohort demographics and classification scheme.[1] In the WHO 2004 classification of lung cancer, LCLC was defined as an undifferentiated NSCLC lacking architectural and cytologic features of SCLC without glandular or squamous differentiation, whereas LCNEC was defined as an LCLC with neuroendocrine morphological features and at least one positive result of immunohistochemical (IHC) staining for a neuroendocrine marker.[2]

Controversy has existed as to whether LCLCs actually represent a truly distinct biological entity or are merely a group of very poorly differentiated tumors of other NSCLC histological subtypes (adenocarcinoma and/or SqCC).[3,4] The 2015 WHO classification scheme regrouped previously histologically defined LCLCs that express pneumocyte markers (thyroid transcription factor 1 [TTF1] and napsin A) as adenocarcinoma and those that test positive for a squamous marker (p40, CK5/6, or p63) as nonkeratinizing or basaloid SqCC, leaving surgically resected tumors lacking expression of these markers (referred to as *marker null*) as LCLCs.[5] Furthermore, LCNECs were separated from LCLCs because they instead share many similarities with neuroendocrine SCLC on the morphological, protein, mutational, DNA methylation, and transcriptional levels, albeit with some heterogeneity (see Rossi et al.,[6] Clinical Lung Cancer Genome Project and Network Genomic Medicine,[7] and Simbolo et al.[12] and the references therein).[6–12] Consistently, in the WHO 2015 update, LCNEC tumors have now been labeled *neuroendocrine* together with SCLC.[5]

Accurate distinction of the histological subtypes is of major clinical relevance. For SqCC, adenocarcinoma, and neuroendocrine tumors (LCNEC or SCLC), subtype-directed diagnostics and therapeutics are widely established. By the refined WHO 2015 classification, a minimization of the remaining marker-null LCLC group has been achieved. Recent studies have suggested poorer outcome for marker-null LCLC cases,[3,8] whereas others have not found this association.[13,14] Notably, all studies are still based on limited retrospective patient cohorts, which may explain the discrepancies. In advanced disease, the LCLC marker-null counterpart, NSCLC not otherwise specified, has been reported to have a worse patient outcome.[15] LCLC tumors remain fairly uncharacterized at the molecular level by high-dimensional genomic techniques, especially considering the otherwise strong molecular efforts made in the WHO 2015 classification scheme. Recent sequencing studies have demonstrated differences in oncogene mutation frequencies between WHO 2004 LCLC tumors expressing adenocarcinoma or SqCC markers and marker-null cases.[3,8,14,16] On the transcriptional level, no studies have thus far resolved the heterogeneity previously suggested for the LCLC group. Importantly, only with an improved molecular understanding can patients with marker-null LCLC benefit from the growing number of targeted treatments adopted in lung cancer.

In this study, we therefore aimed to investigate the transcriptional landscape of LCLC and LCNEC tumors in relation to other histological subgroups of lung cancer. On the basis of unsupervised analysis of global gene expression patterns in 159 surgically resected tumors we demonstrate a WHO 2015 LCLC transcriptional phenotype. Our results link the recent lung cancer classification scheme with the transcriptional landscape of the disease, now defining the poorly differentiated marker-null LCLC group as an entity with a specific gene expression phenotype (GEP).

Materials and Methods

Patient Material Discovery Cohort

A total of 159 patients with early-stage lung cancer surgically treated at the Skåne University Hospital in Lund, Sweden, were collected. A total of 116 cases have been described in previous studies[8,10,17] (Table 1, Supplementary Table 1, and Supplementary Methods). Classification of LCLC and LCNEC was originally performed according to the WHO 2004 scheme[2] and later updated to the WHO 2015 scheme.[5]

Ethics Statement

The study was approved by the Regional Ethical Review Board in Lund, Sweden (registration nos. 2004/762, 2008/702, and 2014/748).

Reclassification of LCLC by IHC Staining

Lung cancer cases with neuroendocrine morphological features that were classified as LCLC according to the WHO 2004 guidelines were evaluated for IHC staining of the neuroendocrine markers chromogranin A, synaptophysin, and CD56 (see Karlsson et al.[8] and Supplementary Methods). WHO 2004–classified LCLC cases without neuroendocrine features were analyzed for IHC staining of CK5/P40 (squamous cell markers) and TTF1/napsin A (adenocarcinoma markers) as described in the Supplementary Methods and in Karlsson et al.,[8] Micke et al.,[18] and Brunnström et al.[19] to classify them according to the WHO 2015 guidelines. No LCLC case was classified as the WHO 2015 uncertain phenotype thanks to successful stains of all cases. In addition, LCNEC cases were also analyzed by retinoblastoma 1 (RB1) immunohistochemistry (see Supplementary Methods).

Table 1. Patient Characteristics and Clinicopathological Data

Variable	Total Cohort	LCLC
No. of patients	159	47
Histological subtype		
Adenocarcinoma	83	–
Squamous cell carcinoma	26	–
SCLC	3	–
LCLC[a]	33	33
LCNEC	14	14
LCLC immunomarker profile (excluding LCNEC)		
Adenocarcinoma-like	–	19 (58%)
Squamous cell carcinoma-like	–	4 (12%)
Marker null	–	10 (30%)
Tumor stage		
I	120 (76%)	24 (51%)
II	27 (17%)	16 (34%)
III	8 (5%)	6 (13%)
IV	2 (1%)	1 (2%)
Smoking history		
Never-smokers	19	0
Smokers	114	23
Not available	26	24
Sex		
Female	85	24
Male	73	23
Median age (range), y	67 (34-84)	63 (34-77)
Patients evaluable for		
Gene expression	159	47
Mutations		

Note: LCLC refers to the WHO 2004 classification.
[a]Basaloid (n = 6) and lymphoepithelioma-like (n = 1) cases are included in the LCLC sample numbers.
LCLC, lung cell lung cancer; LCNEC, large cell neuroendocrine carcinoma.

Global Gene Expression Analysis Discovery Cohort

RNA and DNA from fresh frozen tissue were extracted with the Qiagen Allprep extraction kit (Qiagen, Hilden, Germany). Gene expression data for 43 of the cases were generated by using the Illumina HT12 V4 microarrays (Illumina, San Diego, CA) at the Swegene Center for Integrative Biology at Lund University. Gene expression data were pooled with previously reported data for 116 cases analyzed by the same expression platform[10] as described (see Supplementary Methods) and is available as Gene Expression Omnibus series GSE94601. Consistency in pooling of the two cohorts was confirmed by principal component analysis[20] and analysis of 16 overlapping technical replicate samples between the two cohorts (see Supplementary Fig. 1). Consensus clustering[21] was performed as previously described[22] on mean-centered data (centering across samples) by using probe sets with different log2 SD cutoffs (see Supplementary Table 1). Differentially expressed probe sets between subgroups were identified by significance analysis of microarrays with a false discovery rate threshold of 1%. For independent validation of identified transcriptional subgroups, gene expression centroids were created as mean averages for each gene across all samples in the respective subgroup, as described.[23]

In addition, tumors were also scored according to six expression metagenes in lung cancer representing different biological processes and reported GEPs.[10,24,25] Functional classification was performed as described in Supplementary Methods.

Gene Expression Analysis Validation Cohort

Validation of gene expression subgroups was performed in data provided by Djureinovic et al.[26] (see Supplementary Methods). Histological classification of the samples was updated according to the WHO 2015 guidelines as previously described.[18]

Mutational Analysis

All cases were analyzed by the Illumina TruSight Tumor 26-gene next-generation sequencing (NGS) panel (Illumina), as described.[8] In addition, LCNEC cases were screened for retinoblastoma 1 gene (*RB1*) mutations by using a custom-designed bidirectional NGS panel (Illumina).

Results

IHC Reclassification of WHO 2004 LCLC

Thirty-three lung cancer cases classified as LCLC according to the WHO 2004 guidelines were included in the discovery cohort, of which 70% were reclassified as variants of adenocarcinoma or SqCC on the basis of the WHO 2015 guidelines. Specifically, 19 cases (58%) were reclassified as adenocarcinoma on the basis of positive expression of TTF1/napsin A, four (12%) were reclassified as SqCC on the basis of positive expression of CK5/P40, and 10 (30%) did not express any of these IHC markers (hereon referred to as *marker-null cases*) (see Table 1).

Unsupervised Gene Expression Analysis Stratifies LCLC in Accordance with Immunomarker Expression

To investigate whether the WHO 2015 guidelines translated into a better transcriptional subgrouping of LCLC, we performed unsupervised consensus clustering of a discovery cohort comprising 159 lung cancers of all histological subtypes, including 33 lung cancer cases that were classified as LCLCs and 14 classified as LCNECs according to the WHO 2004 guidelines. We first performed iterative consensus clustering without respect to sample annotations by using variable number of genes (Illumina probes displaying large variation in expression

across tumors [range 300–12581]) and evaluated cluster solutions (k = 5–11) to investigate sample cluster stability versus gene selection. Stable sample clusters formed across a wide range of different gene sets for different cluster solutions, indicating that gene selection has less influence on sample grouping in this cohort (Supplementary Fig. 2*A*). A similar stability was also observed when clustering only WHO 2004 LCLC and LCNEC cases separately (Supplementary Fig. 2*B*).

Next, we performed an in-depth comparison of unsupervised transcriptional subgroups with sample molecular and clinicopathological data. Acknowledging that the histological subtypes of lung cancer strongly influence the transcriptional landscape[7] and that subgroups within the histological subtypes likely exist, we chose a 10-group consensus cluster solution (k = 10 with 2730 Illumina probes, corresponding to a log2ratio standard deviation cutoff of 0.5) to be able to also study characteristics for minor subgroups. Consistent with previous studies,[7,27] we observed a clear separation of adenocarcinoma, SqCC, and SCLC cases into subclusters driven by specific transcriptional programs (Fig. 1). In agreement with recent studies,[7] LCNEC tumors clustered strongly (79% of cases) with SCLC tumors, forming a neuroendocrine subcluster (see Fig. 1). For LCLC, 84% of the WHO 2004 cases reclassified as adenocarcinoma-like clustered in an adenocarcinoma-dominated subcluster, whereas 50% of the LCLC SqCC-like cases clustered in an SqCC-dominated subcluster (see Fig. 1). Notably, 90% of marker-null cases (nine of 10) aggregated in a separate transcriptional cluster (see Fig. 1 [*cluster 8*]), hereon referred to as the *marker-null–enriched subtype*.

To investigate the robustness of the marker-null–enriched subtype in the selected consensus clustering solution, we applied histological annotations to the previously performed iterative consensus clustering (varying number of genes and cluster solutions). Reassuringly, the marker-null subtype was present in all analyses when six consensus clusters or more were used (Supplementary Fig. 3*A*). Similarly, performing consensus clustering in only the LCLC (WHO 2004) and LCNEC tumors also identified the marker-null cases as a separate distinct cluster (Supplementary Fig. 3*B*). Together, these results indicate that we identified, within a general lung cancer population, stable transcriptional subgroups describing LCLC and LCNEC in accordance with the WHO 2015 classification.

Molecular and Clinicopathological Characteristics of LCNEC Tumors

The LCNEC/SCLC-dominated tumor cluster (cluster 3 in Fig. 1) was on the transcriptional level characterized by high expression of neuroendocrine genes, proliferation-related genes, and gene clusters 11 and 8 metagene/Metacore (potentially representing an SRY-box 2 transcription factor–driven gene cluster based on Metacore analysis) and by lower expression of the napsin A/surfactant metagene and gene clusters 2 and 9 metagene/Metacore (potentially representing a MYC transcription factor driven gene cluster based on Metacore analysis) (Figs. 1 and 2 and Supplementary Fig. 4). In tumor cluster 3, tumor protein p53 gene (*TP53*) and *RB1* mutations were found in 91% and 82% of LCNEC tumors, respectively, with *RB1* mutations always concurrently with *TP53* mutations, and 91% of cases showed absent RB1 protein expression.

Only three LCNEC tumors were not present in the neuroendocrine cluster despite IHC expression of one or more tested neuroendocrine markers. Interestingly, two of three outlier cases showed distinct napsin A immunostaining, with the third case showing some focal positive cells (all three cases were TTF1 positive but without any clear histological adenocarcinoma component). All three outlier cases showed *TP53* mutations but positive RB1 protein expression, although one case (located in cluster 7) had a concurrent c.841C>A *RB1* mutation (COSM5658729) together with a serine/threonine kinase 11 gene (*STK11*) frameshift mutation (c.164_165insG). Notably, cluster 7 was strongly enriched for *STK11*-mutated adenocarcinomas (see Fig. 1). Consistently, the average expression of a 16-gene STK11 gene loss signature,[28] indicating serine/threonine kinase 11 inactivation, was higher in both LCNEC cases located in cluster 7 (see Fig. 2*B*). Finally, the three non–cluster 3 cases showed lower expression of the proliferation metagene and generally lower expression of the neuroendocrine metagene compared to the 11 LCNEC cases in tumor cluster 3 (see Fig. 2*C* and *D*).

Molecular and Clinicopathological Characteristics of LCLCs Reclassified as Adenocarcinoma and SqCC

As evident from Figure 1, adenocarcinoma-like or SqCC-like cases classified as LCLC according to the WHO 2004 guidelines did not form separate clusters. This finding indicates that despite their undifferentiated morphological features, these tumors retain transcriptional similarities with different reported adenocarcinoma or SqCC gene expression subtypes.[24,25]

Of the four LCLCs reclassified as SqCC (see Fig. 1), two grouped in a cluster dominated by secretory GEP–classified SqCCs[24] (cluster 4 in Fig. 1) whereas the remaining two cases fell in the neuroendocrine cluster (n = 1 [cluster 3]) and in the marker-null LCLC cluster (n = 1 [cluster 8]). Although our transcriptional analysis suggests that gene expression profiling might add to current SqCC IHC classification, the numbers are too low to allow any definite

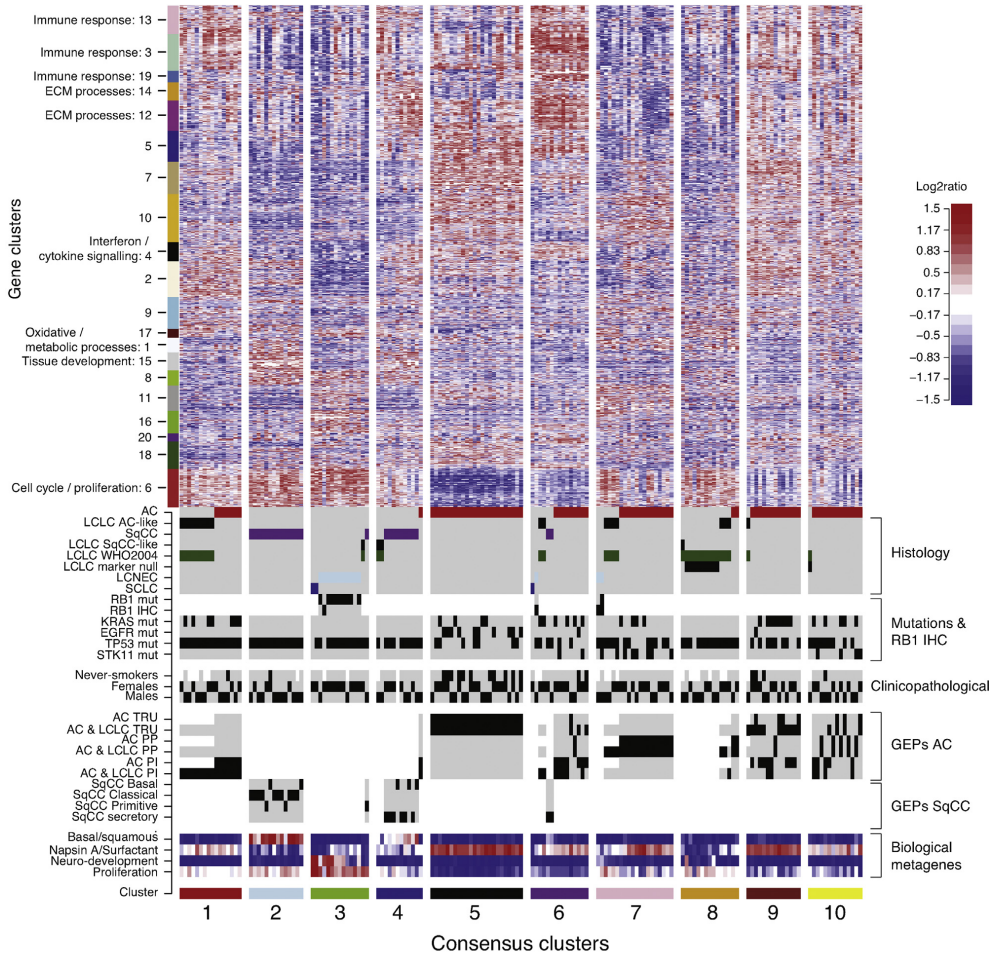


Figure 1. Unsupervised gene expression analysis stratifies large cell lung cancer (LCLC) and large cell neuroendocrine carcinoma (LCNEC) into molecular subgroups. Gene expression heatmap of 2730 Illumina probes across 159 lung cancers stratified by 10 specified consensus clusters. The 2730 probes correspond to a log2ratio standard deviation cutoff of more than 0.5 Annotations for histological subtypes, clinicopathological variables, selected mutations, retinoblastoma 1 immunohistochemistry (RB1 IHC), classification according to reported gene expression phenotypes (GEPs) for adenocarcinoma (AC) and squamous cell carcinoma (SqCC), and expression of selected biological metagenes are provided. For annotations, black corresponds to a positive/presence call, gray to a negative call, and white to not applicable or not available. Gene cluster functional annotations are provided for some specific clusters in the heatmap. ECM, extracellular matrix; mut, mutation; *RB1* mut, retinoblastoma 1 gene mutation; *TP53*, tumor protein p53 gene; *STK11*, serine/threonine kinase 11 gene; TRU, terminal respiratory unit; PP, proximal proliferative; and PI, proximal inflammatory.

conclusions to be drawn about the heterogeneity of SqCC- or SqCC-reclassified LCLC in this cohort.

For adenocarcinoma-like WHO 2004 LCLCs (n = 19), these clustered primarily in adenocarcinoma-dominated clusters (84% of cases [clusters 1, 6, 7, and 9 in Fig. 1]). Importantly, these clusters showed distinct clinicopathological, mutational, and GEP subtype characteristics shared by both the original adenocarcinomas and the reclassified WHO 2004 LCLCs. Three adenocarcinoma-like LCLCs fell in the LCLC marker-null–dominated gene expression cluster (cluster 8 in Fig. 1). However, in-depth analysis of

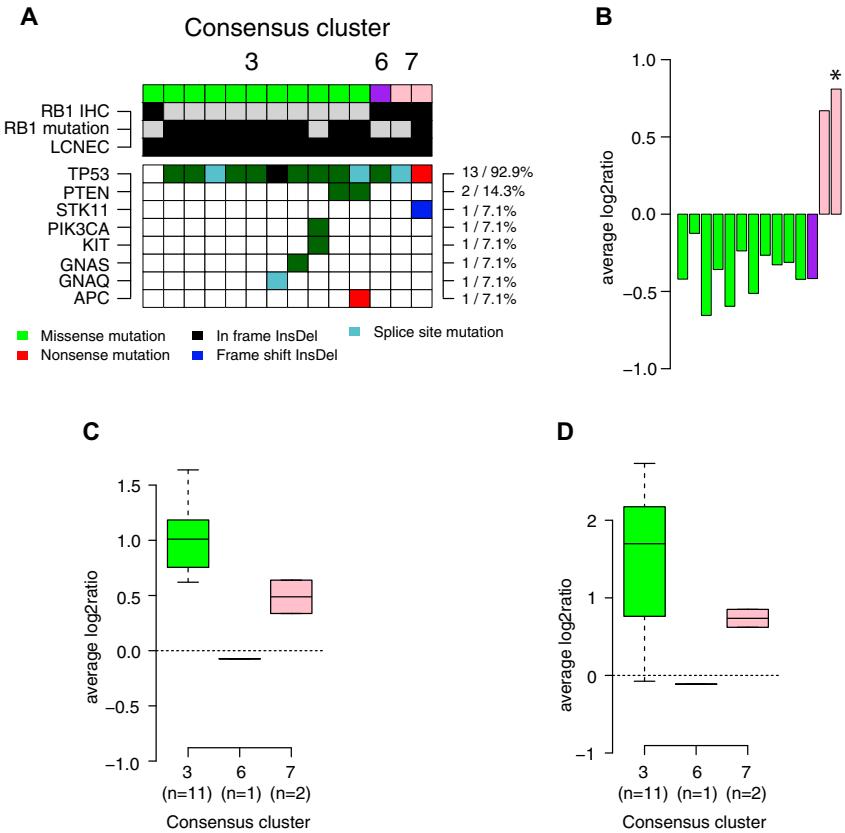


Figure 2. Mutations and gene expression characteristics of large cell neuroendocrine carcinoma (LCNEC) tumors with respect to consensus clusters. (*A*) Mutational map of detected mutations in 26 tumor suppressor and oncogenes for 14 LCNEC tumors stratified by consensus cluster assignment. Only genes with one or more mutations are shown. (*B*) Average log2ratio expression of a 16-gene serine/threonine kinase 11 gene (*STK11*) loss gene signature[28] for LCNEC tumors. Tumors are colored by their consensus cluster as in (*A*). Asterisk indicates a case with *STK11* mutation (*C*) Expression of the proliferation metagene for LCNEC tumors stratified by consensus cluster. (*D*) Expression of the neuroendocrine metagene for LCNEC tumors stratified by consensus cluster. Statistical testing in *B*–*D* was not performed on account of the small group sizes. In *A*–*D*, consensus cluster 3 cases are labeled green, cluster 6 cases purple, and cluster 7 cases pink. RB1 IHC, retinoblastoma 1 immunohistochemistry; *RB1* mutation, retinoblastoma 1 gene mutation; *TP53*, tumor protein p53 gene; *PTEN*, phosphatase and tensin homolog gene; *PIK3CA*, phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha gene; *KIT*, KIT proto-oncogene receptor tyrosine kinase gene; *GNAS*, GNAS complex locus gene; GNAQ, G Protein Subunit Alpha, Q; *APC*, adenomatous polyposis coli gene; APC, WNT signaling pathway regulator gene; InsDel, insertion/deletion.

mutational and gene expression patterns (expression of biological metagenes and gene cluster metagenes) in these few samples compared with adenocarcinoma-like LCLCs in other expression clusters failed to demonstrate any specific or clear differences.

Molecular and Clinicopathological Characteristics of Marker-Null LCLC

Nine out of 10 marker-null–classified LCLCs aggregated in a specific gene expression cluster, accounting for 60% of all tumors in this cluster (cluster 8 in Fig. 1). On the transcriptional level, this cluster was characterized mainly by high expression of proliferation-related genes (see Fig. 1 and Supplementary Fig. 4 for extensive comparison of metagenes and gene clusters). On the genomic level, besides *TP53* mutations in 90% of cases, only one *BRAF* mutation (p.Q456K), one phosphatase and tensin homolog gene (*PTEN*) mutation (p.F257L), and one APC, adenomatous polyposis coli gene (*APC*) mutation (p.E1080Ter) were found (Fig. 3*A*). A total of 445 genes were identified

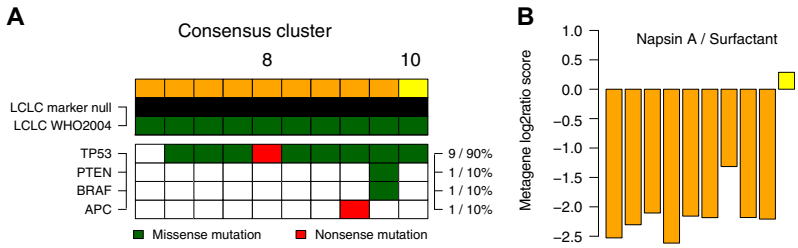


Figure 3. Genomic characteristics of large cell lung cancer (LCLC) marker-null cases. (*A*) Mutational map of detected mutations among 26 tumor suppressor and oncogenes in the 10 LCLC marker-null cases stratified by the 10 defined consensus clusters. Only genes with one or more mutations are shown. (*B*) Expression of the napsin A/surfactant biological metagene for the 10 marker-null cases stratified by consensus clusters. Colors of the individual samples correspond to cluster colors in (*A*). LCLC, large cell lung cancer; *TP53*, tumor protein p53 gene; *PTEN*, phosphatase and tensin homolog gene; *APC*, APC, adenomatous polyposis coli gene; WNT signaling pathway regulator gene.

by significance analysis of microarray to be differentially expressed between tumors in cluster 8 versus all other tumors (false discovery rate <1%). Consistent with our observations from biological metagenes and gene clusters, 79% of genes were down-regulated in cluster 8 tumors, whereas the 95 up-regulated genes (21%) were enriched for biological gene ontology processes such as nucleosome assembly, mitotic cell cycle, DNA replication, cell division, and cellular response to stress and oxidation-reduction processes (Panther overrepresentation test[29] [Bonferroni $p < 0.05$]).

The marker-null LCLC case that did not cluster in the marker-null expression cluster (instead located in cluster 10 in Fig. 1) was further analyzed for transcriptional characteristics by using biological metagene and gene cluster metagene expression patterns. One specific feature was observed, namely, higher expression of the napsin A/surfactant biological metagene than in the remaining cases in cluster 8 (Fig. 3*B*). This finding suggests that this case has some adenocarcinoma-like characteristics consistent with its clustering.

Validation of Transcriptional LCLC Subgroups

To validate our findings we analyzed a reported 199-sample NSCLC cohort profiled by RNA sequencing,[26] including 19 WHO 2004 LCLC cases and five LCNECs. WHO 2015 reclassification identified a similar proportion of LCLC marker-null cases in this cohort as in our discovery set (32% [n = 6]). The remaining cases (68%) were reclassified as adenocarcinoma (37%), SqCC (5%), adenosquamous (16%), or sarcomatoid (10%). To investigate the reproducibility of our discovery cohort findings, we created gene expression centroids for each consensus cluster in the discovery cohort (see Fig. 1) and classified the 199 tumors by a nearest centroid approach (Fig. 4 and Supplementary Table 1). Three of five LCNEC and four of six LCLC marker-null cases were classified into the LCNEC and LCLC marker-null clusters, respectively. The two LCNEC cases not in the neuroendocrine cluster did not express high mRNA levels of neuroendocrine marker genes. For the two outlier LCLC marker-null cases, one case was found in predicted cluster 10 (i.e., similar to the outlier in the discovery set), whereas the second was found in predicted cluster 3 (the neuroendocrine cluster) despite not showing increased neuroendocrine metagene expression (see Fig. 4).

To compare our supervised classification with unsupervised analysis of the validation cohort we performed an independent broad consensus clustering (similar to the discovery cohort). For higher standard deviations (using ~1000 to 2000 genes in the clustering), we did observe aggregation of marker-null LCLC tumors in specific clusters that was comparable to our classification results (≥67% of marker-null tumors in a single cluster). In these instances, marker-null cases comprised approximately one-third of cases in these clusters (see Supplementary Fig. 5). This finding is again similar to our results of classification of this cohort (LCLCs comprise 36% of tumors in predicted cluster 8). In contrast, in the unsupervised analysis of the validation cohort we did not find a similar enrichment of LCNEC tumors (<50% aggregation in a cluster by consensus clustering versus 60% for supervised classification) (see Supplementary Fig. 5).

Discussion

Previous extensive gene expression analyses have not discriminated LCLC as a distinct expression phenotype in NSCLC. Instead, LCLC tumors have often been dispersed in various adenocarcinoma, SqCC, or LCNEC subgroups (for example, see Clinical Lung Cancer Genome Project et al.,[7] Djureinovic et al.,[26] and Botling et al.[30]). In the current study, we explored the transcriptional and genomic spectrum of LCLC and LCNEC in the context of

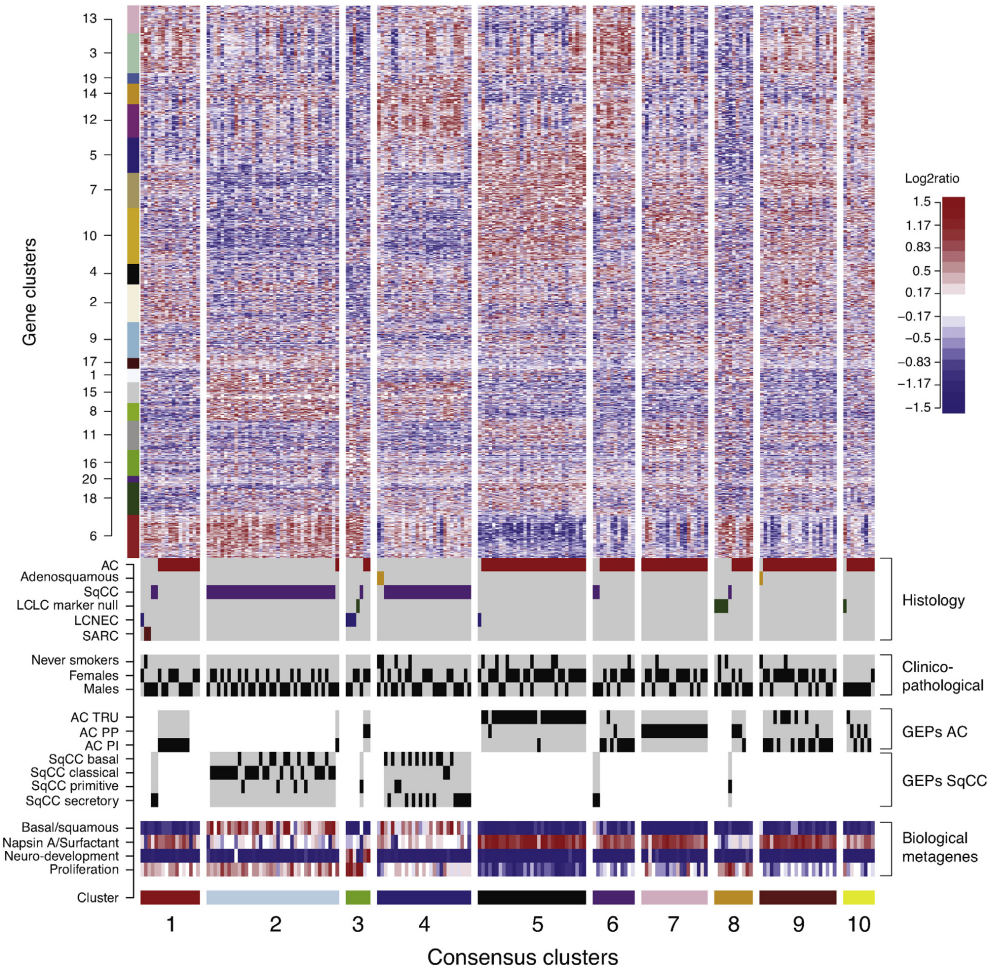


Figure 4. Validation of the large cell neuroendocrine carcinoma (LCNEC) and large cell lung cancer (LCLC) marker-null gene expression phenotypes (GEPs). Gene expression heatmap of 2196 genes across 199 lung cancers from Djureinovic et al.[26] classified by a nearest centroid predictor into 10 consensus clusters. Six cases were set as unclassified and are excluded from the heatmap (n = 193 cases in the heatmap). Annotations for histological subtypes, clinicopathological variables, classification according to reported GEPs, for adenocarcinoma (AC) and squamous cell carcinoma (SqCC), and expression of selected biological metagenes are provided. For annotations, black corresponds to a positive/presence call, gray to a negative call, and white to not applicable or not available. SARC, sarcomatoid, TRU, terminal respiratory unit; PP, proximal proliferative; PI, proximal inflammatory.

recent molecular findings, refined pathological classification schemes, and other histological subgroups of lung cancer. Through gene expression profiling of 159 primary lung tumors of all histological subtypes and subsequent independent validation in 199 additional cases, we found that the WHO 2015 guidelines translated into a better transcriptional subgrouping of LCLC. Our results endorse the recently refined definition of LCLC on a transcriptional level, demonstrating the presence of a GEP highly enriched for poorly differentiated highly proliferative tumors that do not express diagnostic immunomarkers for adenocarcinoma or SqCC.

An important conclusion from this study is that the updated WHO 2015 classification translates to the

transcriptional landscape, as the transcriptional phenotypes mimic the WHO classification on a general level. On a detailed level, transcriptional phenotypes likely provide additional stratification within histological subtypes that may be associated with prognosis and therapeutic options. Within both adenocarcinoma and SqCC different GEPs have been proposed (e.g., by Wilkerson et al.,[24] Wilkerson et al.,[25] Bhattacharjee et al.,[27] Takeuchi et al.,[31] Garber et al.,[32] and Raponi et al.[33]), although the consensus between phenotypes is not absolute in independent multicohort analysis.[34] Our data suggest that additional subgroup stratification appears possible in at least lung adenocarcinoma compared with the most used gene expression-based classification.[25,35]

The new WHO 2015 classification system has substantially reduced the proportion of LCLC (marker-null) NSCLC cases. The reclassification frequencies observed in the discovery and validation cohorts (70% and 68%, respectively) are consistent with previous studies (59%–90%).[3,14,16,36–40] Our transcriptional analysis of reclassified WHO 2004 LCLC cases demonstrates that these cases have heterogeneous profiles matching different molecular subsets/GEPs of adenocarcinoma and SqCC, providing additional biological information compared to current diagnostic immunomarkers. Generally, reclassified WHO 2004 LCLCs followed the more aggressive transcriptional phenotypes, such as proximal proliferative or proximal inflammatory in adenocarcinoma.[35] Specifically, no LCLC according to the WHO 2004 guidelines that has been reclassified as adenocarcinoma fell in the cluster characterized by *EGFR* mutations, never-smokers, and low-proliferative terminal respiratory unit–like tumors[35] (cluster 5). These observations appear consistent with the generally undifferentiated morphological state of WHO 2004 LCLC tumors and the well-known association of LCLC with smoking.

In this study, marker-null LCLC cases formed a specific, reproducible, transcriptional cluster linking the immunomarker classification with a transcriptional phenotype featuring characteristics of poorly differentiated proliferative cancer (see Figs. 1 and 4), which typically is linked to poorer patient outcome. In the validation cohort, the enrichment of marker-null tumors in the proposed phenotype was lower than in the discovery cohort (67% versus 90%, respectively). This might be due to low sample numbers in the former cohort causing strong proportional shifts by individual samples (16% and 20% shifts per sample for marker-null and LCNEC comparisons, respectively). The clearer identification of the marker-null phenotype in the discovery cohort may thus be due to a higher enrichment of marker-null tumors than in the validation cohort (6% versus 3%, respectively). Considering the overall low sample numbers, we acknowledge that additional validation is warranted.

Three main differentiation lineages (bronchioid/glandular, epidermoid/squamoid/squamous/keratinizing, and neuroendocrine) are generally recognized in lung cancer (for example, see the references in Pelosi et al.[14]). Recent mutational analyses of marker-null LCLC cases have suggested that genetic profiling could further reduce the marker-null group beyond IHC scoring by detection of typical adenocarcinoma or SqCC mutations, often favoring an adenocarcinoma lineage.[3,8,14,16] Although recent NGS-based studies[14,16] have reported a notably higher frequency of (especially) adenocarcinoma-linked mutations in marker-null cases than in our cohort, it may be noted that 83% of non–marker-null cases in our marker-null–enriched gene expression cluster (cluster 8) were adenocarcinoma or LCLCs according to the WHO 2004 guidelines that have been reclassified as adenocarcinoma. This observation may lend some support to a hypothesis that marker-null LCLCs represent a variant of undifferentiated TTF1-negative adenocarcinoma.[14,41] Clearly, deeper molecular/genetic characterization of marker-null LCLC may further reduce the WHO 2015 marker-null group, benefitting both clinical patient management and basic understanding of differentiation lineages in lung cancer.

On the basis of massive parallel sequencing studies, it is becoming evident that a subset of LCNEC tumors share mutational patterns with SCLC, whereas others carry mutations typically altered in non-neuroendocrine tumors.[7,9,11,12] Rekhtman et al.[9] recently hypothesized a genetic division of LCNECs into SCLC-like and NSCLC-like subgroups. The SCLC-like group was defined by concomitant *TP53* and *RB1* alterations, reflecting their ubiquitous inactivation in SCLC.[42] In contrast, the NSCLC-like subset was characterized by the lack of concomitant *TP53* and *RB1* alterations and occurrence of other NSCLC-type mutations (e.g., *KRAS* and *STK11* mutations). Interestingly, our data indicate that this stratification may potentially be mimicked also on the transcriptional level, providing a speculative link between the mutational and transcriptional landscape of LCNEC. An earlier gene expression study has suggested the presence of a good and poor outcome group in neuroendocrine tumors independent of LCNEC and SCLC status.[43] However, how these subgroups relate to the division of LCNEC by Rekhtman et al. is not clear,[9] as genetic data from the former study are not available. Acknowledging the limited number of LCNEC cases in the current study, consensus cluster 3 and non–cluster 3 cases still share intriguingly many features of the proposed SCLC-like and NSCLC-like LCNEC groups, respectively. These features include (1) frequent co-occurrence of *TP53* mutations and *RB1* mutations/protein loss in cluster 3, (2) higher

expression of proliferation-related genes and SCLC marker genes in cluster 3, (3) more pronounced expression of a potential SRY-box 2 (a suggested lineage survival oncogene in SCLC) transcription factor–driven gene cluster in cluster 3 (see Rudin et al.[44]), (4) indication of frequent *STK11* inactivation in non–cluster 3 cases, (5) distinct napsin A staining in two of three non–cluster 3 cases with few additional focal positive cells in the third case (all cases TTF1-positive), and (6) a trend of worse patient outcome in cluster 3 (with 55% of patients dead within 5 years versus 33% of non–cluster 3 patients). Notably, LCNEC has hitherto been consistently reported with negative napsin A expression,[4,45] and how this relates to the proposed NSCLC-like subgroup and whether the speculative association with transcriptional patterns indicated in the current study stands in larger validation studies remain to be determined.

In summary, the current study provides a novel link between the recent WHO 2015 diagnostic classification scheme and the transcriptional landscape of lung cancer. Our study confirms that WHO 2004–classified LCLCs may be refined by genomic patterns into clinically relevant subgroups that may have implications for diagnosis, predictive testing, and therapy decisions. Specifically, we demonstrate a gene expression profile that further defines a phenotype associated with poor patient outcome and comprises undifferentiated LCLC cases not expressing adenocarcinoma or SqCC markers. A continued search for molecular targets for therapeutic inhibition in WHO 2015 LCLC cases is highly warranted.

Acknowledgments

Financial support for this study was provided by the Swedish Cancer Society, the Mrs. Berta Kamprad Foundation, the Gunnar Nilsson Cancer Foundation, the Crafoord Foundation, BioCARE a Strategic Research Program at Lund University, the Gustav V's Jubilee Foundation, and governmental funding (ALF). Dr. Staaf, Mr. Karlsson, and Dr. Planck conceived of the study. Mr. Karlsson performed the experimental microarray and next-generation sequencing analyses with the assistance of Mr. Reuterswärd. Dr. Mattsson performed the validation experiments. Drs. Micke, Botling, Mr. La Fleur, and Mattsson provided the validation cases. Dr. Brunnström performed the pathological evaluations, and Drs. Brunnström, Jönsson, and Planck evaluated the immunohistochemical staining results. Mr. Karlsson and Dr. Staaf performed statistical analyses with support of Dr. Veerla. Dr. Staaf drafted and wrote the manuscript together with Mr. Karlsson. All authors approved the final manuscript.

Supplementary Data

Note: To access the supplementary material accompanying this article, visit the online version of the *Journal of Thoracic Oncology* at www.jto.org and at http://dx.doi.org/10.1016/j.jtho.2017.05.008.

References

1. Howlader N, Noone AM, Krapcho M, et al, eds. SEER cancer statistics review, 1975-2013. http://seer.cancer.gov/csr/1975_2013/. Accessed 18 Jan 2017.
2. Travis WD, Brambilla E, Muller-Hermelink HK, Harris CC, eds. *World Health Organization Classification of Tumours. Pathology and Genetics of Tumours of the Lung, Pleura, Thymus and Heart*. Lyon, France: IARC Press; 2004.
3. Rekhtman N, Tafe LJ, Chaft JE, et al. Distinct profile of driver mutations and clinical features in immunomarker-defined subsets of pulmonary large-cell carcinoma. *Mod Pathol*. 2014;26:511–522.
4. Rossi G, Mengoli MC, Cavazza A, et al. Large cell carcinoma of the lung: clinically oriented classification integrating immunohistochemistry and molecular biology. *Virchows Arch*. 2014;464:61–68.
5. Travis WD, Brambilla E, Burke AP, Marx A, Nicholson AG, eds. *WHO Classification of Tumours of the Lung, Pleura, Thymus and Heart*. Lyon, France: IARC Press; 2015.
6. Rossi G, Cavazza A, Marchioni A, et al. Role of chemotherapy and the receptor tyrosine kinases KIT, PDGFRalpha, PDGFRbeta, and Met in large-cell neuroendocrine carcinoma of the lung. *J Clin Oncol*. 2005;23:8774–8785.
7. Clinical Lung Cancer Genome Project (CLCGP), Network Genomic Medicine (NGM). A genomics-based classification of human lung tumors. *Sci Transl Med*. 2013;5:209ra153.
8. Karlsson A, Brunnstrom H, Lindquist KE, et al. Mutational and gene fusion analyses of primary large cell and large cell neuroendocrine lung cancer. *Oncotarget*. 2015;6:22028–22037.
9. Rekhtman N, Pietanza MC, Hellmann MD, et al. Next-generation sequencing of pulmonary large cell neuroendocrine carcinoma reveals small cell carcinoma-like and non-small cell carcinoma-like subsets. *Clin Cancer Res*. 2016;22:3618–3629.
10. Karlsson A, Jonsson M, Lauss M, et al. Genome-wide DNA methylation analysis of lung carcinoma reveals one neuroendocrine and four adenocarcinoma epitypes associated with patient outcome. *Clin Cancer Res*. 2014;20:6127–6140.
11. Miyoshi T, Umemura S, Matsumura Y, et al. Genomic profiling of large-cell neuroendocrine carcinoma of the lung. *Clin Cancer Res*. 2016;23:757–765.
12. Simbolo M, Mafficini A, Sikora KO, et al. Lung neuroendocrine tumours: deep sequencing of the four WHO histotypes reveals chromatin remodelling genes as major players and a prognostic role for TERT, RB1, MEN1 and KMT2D. *J Pathol*. 2017;241:488–500.
13. Hwang DH, Szeto DP, Perry AS, Bruce JL, Sholl LM. Pulmonary large cell carcinoma lacking squamous differentiation is clinicopathologically indistinguishable from solid-subtype adenocarcinoma. *Arch Pathol Lab Med*. 2014;138:626–635.

14. Pelosi G, Fabbri A, Papotti M, et al. Dissecting pulmonary large-cell carcinoma by targeted next generation sequencing of several cancer genes pushes genotypic-phenotypic correlations to emerge. *J Thorac Oncol*. 2015;10:1560–1569.
15. Righi L, Vavala T, Rapa I, et al. Impact of non-small-cell lung cancer-not otherwise specified immunophenotyping on treatment outcome. *J Thorac Oncol*. 2014;9:1540–1546.
16. Driver BR, Portier BP, Mody DR, et al. Next-generation sequencing of a cohort of pulmonary large cell carcinomas reclassified by World Health Organization 2015 criteria. *Arch Pathol Lab Med*. 2016;140:312–317.
17. Salomonsson A, Jonsson M, Isaksson S, et al. Histological specificity of alterations and expression of KIT and KITLG in non-small cell lung carcinoma. *Genes Chromosomes Cancer*. 2013;52:1088–1096.
18. Micke P, Mattsson JS, Djureinovic D, et al. The impact of the fourth edition of the WHO classification of lung tumours on histological classification of resected pulmonary NSCCs. *J Thorac Oncol*. 2016;11:862–872.
19. Brunnström H, Johansson L, Jirström K, Jönsson M, Jönsson P, Planck M. Immunohistochemistry in the differential diagnostics of primary lung cancer: an investigation within the Southern Swedish Lung Cancer Study. *Am J Clin Pathol*. 2013;140:37–46.
20. Lauss M, Visne I, Kriegner A, Ringnér M, Jönsson G, Höglund M. Monitoring of technical variation in quantitative high-throughput datasets. *Cancer Inform*. 2013;12:193–201.
21. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*. 2010;26:1572–1573.
22. Karlsson A, Ringnér M, Lauss M, et al. Genomic and transcriptional alterations in lung adenocarcinoma in relation to smoking history. *Clin Cancer Res*. 2014;20:4912–4924.
23. Planck M, Isaksson S, Veerla S, Staaf J. Identification of transcriptional subgroups in EGFR-mutated and EGFR/KRAS-wild type lung adenocarcinoma reveals gene signatures associated with patient outcome. *Clin Cancer Res*. 2013;19:5116–5126.
24. Wilkerson MD, Yin X, Hoadley KA, et al. Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types. *Clin Cancer Res*. 2010;16:4864–4875.
25. Wilkerson MD, Yin X, Walter V, et al. Differential pathogenesis of lung adenocarcinoma subtypes involving sequence mutations, copy number, chromosomal instability, and methylation. *PLoS One*. 2012;7:e36530.
26. Djureinovic D, Hallstrom BM, Horie M, et al. Profiling cancer testis antigens in non-small-cell lung cancer. *JCI Insight*. 2016;1:e86837.
27. Bhattacharjee A, Richards WG, Staunton J, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci U S A*. 2001;98:13790–13795.
28. Kaufman JM, Amann JM, Park K, et al. LKB1 loss induces characteristic patterns of gene expression in human tumors associated with NRF2 activation and attenuation of PI3K-AKT. *J Thorac Oncol*. 2014;9:794–804.
29. Mi H, Poudel S, Muruganujan A, Casagrande JT, Thomas PD. PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res*. 2016;44:D336–D342.
30. Botling J, Edlund K, Lohr M, et al. Biomarker discovery in non-small cell lung cancer: integrating gene expression profiling, meta-analysis and tissue microarray validation. *Clin Cancer Res*. 2012;19:194–204.
31. Takeuchi T, Tomida S, Yatabe Y, et al. Expression profile-defined classification of lung adenocarcinoma shows close relationship with underlying major genetic changes and clinicopathologic behaviors. *J Clin Oncol*. 2006;24:1679–1688.
32. Garber ME, Troyanskaya OG, Schluens K, et al. Diversity of gene expression in adenocarcinoma of the lung. *Proc Natl Acad Sci U S A*. 2001;98:13784–13789.
33. Raponi M, Zhang Y, Yu J, et al. Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. *Cancer Res*. 2006;66:7466–7472.
34. Ringner M, Staaf J. Consensus of gene expression phenotypes and prognostic risk predictors in primary lung adenocarcinoma. *Oncotarget*. 2016;7:52957–52973.
35. The Cancer Genome Atlas Network, Collisson EA, Campbell JD, et al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*. 2014;511:543–550.
36. Au NH, Cheang M, Huntsman DG, et al. Evaluation of immunohistochemical markers in non-small cell lung cancer by unsupervised hierarchical clustering analysis: a tissue microarray study of 284 cases and 18 markers. *J Pathol*. 2004;204:101–109.
37. Barbareschi M, Cantaloni C, Del Vescovo V, et al. Heterogeneity of large cell carcinoma of the lung: an immunophenotypic and miRNA-based analysis. *Am J Clin Pathol*. 2011;136:773–782.
38. Monica V, Ceppi P, Righi L, et al. Desmocollin-3: a new marker of squamous differentiation in undifferentiated large-cell carcinoma of the lung. *Mod Pathol*. 2009;22:709–717.
39. Monica V, Scagliotti GV, Ceppi P, et al. Differential thymidylate synthase expression in different variants of large-cell carcinoma of the lung. *Clin Cancer Res*. 2009;15:7547–7552.
40. Pardo J, Martinez-Penuela AM, Sola JJ, et al. Large cell carcinoma of the lung: an endangered species? *Appl Immunohistochem Mol Morphol*. 2009;17:383–392.
41. Pelosi G, Fraggetta F, Pasini F, et al. Immunoreactivity for thyroid transcription factor-1 in stage I non-small cell carcinomas of the lung. *Am J Surg Pathol*. 2001;25:363–372.
42. George J, Lim JS, Jang SJ, et al. Comprehensive genomic profiles of small cell lung cancer. *Nature*. 2015;524:47–53.
43. Jones MH, Virtanen C, Honjoh D, et al. Two prognostically significant subtypes of high-grade lung neuroendocrine tumours independent of small-cell and large-cell neuroendocrine carcinomas identified by gene expression profiles. *Lancet*. 2004;363:775–781.
44. Rudin CM, Durinck S, Stawiski EW, et al. Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. *Nat Genet*. 2012;44:1111–1116.
45. Masai K, Tsuta K, Kawago M, et al. Expression of squamous cell carcinoma markers and adenocarcinoma markers in primary pulmonary neuroendocrine carcinomas. *Appl Immunohistochem Mol Morphol*. 2013;21:292–297.

Study IV

Mutational and gene fusion analyses of primary large cell and large cell neuroendocrine lung cancer

Anna Karlsson[1], Hans Brunnström[2,3], Kajsa Ericson Lindquist[3], Karin Jirström[2,3], Mats Jönsson[1], Frida Rosengren[1], Christel Reuterswärd[1], Helena Cirenajwis[1], Åke Borg[1,6], Per Jönsson[4], Maria Planck[1,5], Göran Jönsson[1,6], Johan Staaf[1,6]

[1]Division of Oncology and Pathology, Department of Clinical Sciences Lund, Lund University, Medicon Village, SE 22381 Lund, Sweden

[2]Division of Oncology and Pathology, Department of Clinical Sciences Lund, Lund University, SE 22185 Lund, Sweden

[3]Department of Pathology, Regional Laboratories Region Skåne, SE 22185 Lund, Sweden

[4]Department of Thoracic Surgery, Lund University, Skåne University Hospital, SE 22185 Lund, Sweden

[5]Department of Oncology, Skåne University Hospital, SE 22185 Lund, Sweden

[6]Create Health Strategic Center for Translational Cancer Research, Lund University, Medicon Village, SE 22381 Lund, Sweden

Correspondence to:
Johan Staaf, **e-mail:** johan.staaf@med.lu.se

Keywords: *large cell lung cancer, LCNEC, mutation, gene fusion, ALK*

Received: *March 1, 2015* ***Accepted:*** *June 05, 2015* ***Published:*** *June 17, 2015*

ABSTRACT

Large cell carcinoma with or without neuroendocrine features (LCNEC and LC, respectively) constitutes 3–9% of non-small cell lung cancer but is poorly characterized at the molecular level. Herein we analyzed 41 LC and 32 LCNEC (including 15 previously reported cases) tumors using massive parallel sequencing for mutations in 26 cancer-related genes and gene fusions in *ALK*, *RET*, and *ROS1*. LC patients were additionally subdivided into three immunohistochemistry groups based on positive expression of TTF-1/Napsin A (adenocarcinoma-like, *n* = 24; 59%), CK5/P40 (squamous-like, *n* = 5; 12%), or no marker expression (marker-negative, *n* = 12; 29%). Most common alterations were *TP53* (83%), *KRAS* (22%), *MET* (12%) mutations in LCs, and *TP53* (88%), *STK11* (16%), and *PTEN* (13%) mutations in LCNECs. In general, LCs showed more oncogene mutations compared to LCNECs. Immunomarker stratification of LC revealed oncogene mutations in 63% of adenocarcinoma-like cases, but only in 17% of marker-negative cases. Moreover, marker-negative LCs were associated with inferior overall survival compared with adenocarcinoma-like tumors (*p* = 0.007). No *ALK*, *RET* or *ROS1* fusions were detected in LCs or LCNECs. Together, our molecular analyses support that LC and LCNEC tumors follow different tumorigenic paths and that LC may be stratified into molecular subgroups with potential implications for diagnosis, prognostics, and therapy decisions.

INTRODUCTION

Non-small cell lung cancer (NSCLC) accounts for the majority of diagnosed lung cancers and is dominated by the adenocarcinoma, squamous cell carcinoma (SqCC) and large cell carcinoma with or without neuroendocrine features (LCNEC and LC, respectively) histological subtypes. In NSCLC, LC and LCNEC together account for 3–9% of all cases depending on cohort demographics and classification scheme, with a generally poor prognosis compared to other NSCLC subgroups [1, 2]. In the 2004 WHO classification of lung cancer LC is defined as an undifferentiated NSCLC lacking architectural and cytologic features of small-cell carcinoma, glandular or squamous differentiation, whereas LCNEC is defined as an LC with neuroendocrine morphological features and at least one positive neuroendocrine immunohistochemical (IHC) marker [3]. LCNEC tumors share many similarities with small-cell lung cancer (SCLC) on the morphological, IHC and molecular level [4] (and references therein). Based on advances in

immunomarkers for classification of adenocarcinoma and SqCC there is today significant controversy on whether LC actually represent a truly distinct biological entity, or merely a group of very poorly differentiated tumors of other NSCLC groups (adenocarcinoma and/or SqCC) [5, 6]. In fact, in the most recent 2015 WHO classification of lung cancer LCs that are mucin-positive or expresses pneumocyte markers should now be classified as adenocarcinoma, and the squamous marker-positive cases as nonkeratinizing squamous cell carcinoma [7].

In comparison to other NSCLC subgroups, LC and LCNEC tumors remain fairly uncharacterized at the molecular level by modern genomic techniques. Recent studies have investigated copy number alterations (CNAs) in LC and LCNEC [8, 9], highlighted the transcriptional similarity between LCNEC and SCLC [8], and identified a neuroendocrine DNA methylation subgroup in lung cancer [10]. In contrast, studies of the genome-wide mutational landscape in LC and LCNEC using massive parallel sequencing methods (NGS) are scarce. A recent analysis of 15 LCNEC tumors using whole-exome sequencing associated mainly mutations in *TP53*, *RB1*, and *EP300* with LCNEC (and SCLC) tumor histology, with additional mutations in LCNEC also found in adenocarcinomas and SqCCs [8]. Studies of smaller gene sets have identified abnormal *TP53* expression in both LC and LCNEC tumors and *KRAS* mutations predominantly in LCs [6, 11]. Mutations in *EGFR* and *ALK* gene fusions represent current molecular treatment predictive alterations for targeted therapy in lung cancer [12], but rarely appear in LC or LCNEC tumors with only a few reported cases in the literature [5, 6, 13–16]. Clearly, better characterization of the mutational landscape in LC and LCNEC is needed to take advantage of the growing number of targeted treatments and our emerging understanding of treatment resistance factors in lung cancer.

In this study, we aimed to investigate the mutational landscape of LC and LCNEC tumors using a panel of 26 well-established oncogenes and tumor suppressor genes in combination with *ALK*, *RET*, and *ROS1* gene fusion analysis and copy number analysis of targeted genes. To this end, we analyzed 41 LC and 17 LCNEC cases by massive parallel sequencing and combined our results with 15 whole-exome sequenced LCNEC cases [8] and previously reported copy number data [8, 10].

MATERIALS AND METHODS

Patient material

DNA and total RNA were extracted from 57 early stage lung cancer patients surgically treated at the Skåne University Hospital in Lund, Sweden (Table 1 and Supplementary Table S1). Patients in this retrospective study had not received any neoadjuvant treatment before surgery. One patient harbored a mixed cancer, with one LC and one LCNEC tumor component, treated as two individual tumors in the analysis. In total, 41 LC and 17 LCNEC samples were included from this patient cohort. For all cases, relevant pathological slides were re-evaluated and clinicopathological characteristics were updated to be in line with recent international criteria and guidelines [3, 17]. Thirteen cases have been described in previous studies [10, 18]. From Seidel et al. [8], we included whole-exome sequencing and copy number data on genes investigated in the experimental Lund cohort from 15 additional LCNEC cases (Table 1).

Ethics statement

The study was approved by the Regional Ethical Review Board in Lund, Sweden (Registration no. 2004/762, 2008/702, 2007/445, and 2014/748).

Immunohistochemistry

Cases with neuroendocrine morphological features were evaluated for IHC staining of the neuroendocrine markers chromogranin A, synaptophysin and CD56 (Supplementary Methods). At least 10% positive tumor cells were required for positive staining for these markers. In addition, LC cases were analyzed for IHC staining of CK5/P40 (squamous cell markers) and TTF-1/Napsin A (adenocarcinoma markers). Staining intensities for these markers were categorized as 0 (<1% positive tumor cells), 1 (1–10%), 2 (11–25%), 3 (26–50%), and 4 (>50% positive tumor cells). Similar to the recent 2015 WHO update on lung cancer [7], we classified a categorized intensity of ≥1 as positive for TTF-1 or Napsin A, and ≥ 2 as positive for CK5 or P40. A LC sample was classified as adenocarcinoma-like if a positive TTF-1 and/or Napsin A staining was observed. A LC case was classified as squamous-like if a positive CK5 and/or P40 staining was observed. IHC analyses are further described in Brunnström et al. [19] and Supplementary Methods.

Mutational analysis

All Lund cases were analyzed by the NGS-based Illumina TruSight Tumor gene panel on a MiSeq instrument according to manufacturer's instructions (Illumina, San Diego, CA, US). Analyzed regions included a selected set of complete exons in 26 genes: *AKT1* (exon 2), *ALK* (exon 23), *APC* (exon 15), *BRAF* (exons 11, 15), *CDH1* (exons 8, 9, 12), *CTNNB1* (exon 2), *EGFR* (exons 18, 19, 20, 21), *ERBB2* (exon 20), *FBXW7* (exons 7, 8, 9, 10, 11), *FGFR2* (exon 6), *FOXL2* (exon 1), *GNAQ* (exons 4, 5, 6), *GNAS* (exons 6, 8), *KIT* (exons 9, 11, 13, 17, 18), *KRAS* (exons 1, 2, 3, 4), *MAP2K1* (exon 2), *MET* (exons 1, 4, 13, 15, 16, 17, 18, 20), *MSH6* (exons 5), *NRAS* (exons 1, 2, 3, 4), *PDGFRA* (exons 11, 13, 17), *PIK3CA* (exons 1, 2, 7, 9, 20), *PTEN* (exons 1, 2, 3, 4, 5, 6, 7, 9), *SMAD4* (exons 8, 11), *SRC* (exon

Table 1: Patient characteristics and clinicopathological data

	Lund	CLCGP [8]	All cases
Histology			
LC (basaloid)	41 (6) *, **	-	41 (6)
LCNEC	17*	15	32
LC immunomarker profile			
Adenocarcinoma-like	24 (59%)	-	24 (59%)
Squamous cell carcinoma-like	5 (12%)	-	5 (12%)
Marker null	12 (29%)	-	12 (29%)
Tumor stage			
I	29	7	36
II	19	6	25
III	8	2	10
IV	2	0	2
Smoking history			
Never-smokers	0	0	0
Smokers	34	11	45
Not available	24	4	28
Gender			
Female	31	6	37
Male	27	9	36
Age (median & range)	66 (47–82)	67 (47–80)	66 (47–82)
Patients evaluable for			
Mutations	58	15	73
ALK, *RET*, *ROS1* fusions	46	1***	47
Copy number alterations	46	10	56

*One patient had a mixed tumor with both an LC and LCNEC component.
** Basaloid (n = 6) and lymphoepithelioma-like (n = 1) cases are included in the LC sample numbers.
*** Evaluated for *ALK*/*RET*/*ROS1* fusions by FISH.

10), *STK11* (exons 1, 4, 6, 8), and *TP53* (exons 2, 3, 4, 5, 6, 7, 8, 9, 10, 11). DNA extraction was performed using the Qiagen GeneRead (Qiagen, Hilden, Germany) kit for formalin-fixed paraffin embedded tissue (FFPE), or by the Qiagen AllPrep kit for fresh frozen tissue (Supplementary Methods). Macrodissection of FFPE cases were performed when possible prior to DNA extraction. Alignment, quality filtering, variant calling, and variant annotation were performed using the standard MiSeq Reporter and VariantStudio analysis pipeline (Illumina). Only nonsynonymous variants with a quality score equal to 100 that passed the bi-directional sequencing quality filter in TruSight Tumor were considered. Read depths (X) for genes with detected variants varied between 3024–143140X (median 18624X, interquartile range 28450X).

Gene fusion analysis

Analysis of *ALK*, *RET*, and *ROS1* gene fusions were performed using the RNA-based Archer FusionPlex ARR v2 kit (ArcherDX, Boulder, CO, US) and the MiSeq instrument (Illumina) (Supplementary Methods). The HCC78 (*ROS1-SLC34A2*), KARPAS-299 (*ALK-NPM1*), LC-2/ad (*CCDC6-RET*), and H2228 (*EML4-ALK*) cell lines were used as controls (Supplementary Methods). RNA from FFPE tissues (n = 11 samples) were extracted using the Qiagen Allprep FFPE extraction kit (Qiagen), while

RNA from cell lines or fresh frozen tissue were extracted using the non-FFPE Qiagen Allprep extraction kit. Data analysis was performed using software tools provided with the Archer kit (ArcherDX). Confirmatory ALK immunohistochemistry was performed using the D5F3 antibody (Ventana Medical Systems), and confirmatory *ALK* FISH analysis using the Vysis ALK break apart FISH probe (Vysis) according to manufacturers' instructions.

Copy number analysis

Calls of copy number gain, loss, amplification and focused copy number loss for genes included in the TruSight Tumor panel were made for 46 tumors in the Lund cohort based on data from ongoing or published studies on the same tumor cohort [10], and for ten cases from Seidel et al. [8] as described by [10, 20] and Supplementary Methods.

NGS validation analyses

Ten mutations in *KRAS* detected by the TruSight Tumor panel were selected for validation by the Therascreen® KRAS RGQ PCR Kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol. In addition, seven unrelated tumor FFPE specimens, including two melanomas, two lung adenocarcinomas, and three colon cancers were also used to validate the NGS platform. These samples had verified mutations in *BRAF* (the two melanomas and one colon cancer: V600E), *KRAS* (two colon cancers: G13D and G12S), and *EGFR* (the two lung adenocarcinomas: L858R and E746_A750del). Mutations in these cases were obtained from routine clinical diagnostics based on pyrosequencing or Q-PCR performed at the Skåne University Hospital in Lund, Sweden.

RESULTS

Tumor and patient characteristics

A cohort of 41 LC and 17 LCNEC cases (Lund cohort) were pooled with 15 reported LCNEC cases [8], thus rendering a total of 73 cases (Table 1). All patients with available chart data were (current or former) smokers. There were no statistical differences in the distribution of tumor stage, gender, or age of diagnosis between LC and LCNEC cases ($p > 0.05$, Fisher's exact test or Wilcoxon's test). Among the LC cases, six tumors were histopathologically subclassified as basaloid, and one as lymphoepithelioma-like. Protein expression of adenocarcinoma markers TTF-1 and Napsin A and squamous markers CK5 and P40 were investigated in the Lund cohort and the public LCNEC cohort (TTF-1 only). 77% of all analyzed LCNEC cases showed positive TTF-1 expression. 59% of LC cases in the Lund cohort were IHC positive for TTF-1, while 44% were IHC positive for Napsin A. 75% of the TTF-1 positive LC cases also showed positive Napsin A expression, while no case was Napsin A positive but TTF-1 negative. For the squamous markers CK5 and P40, 5% and 10% of LC cases showed positive staining, respectively.

The mutational spectrum of LC and LCNEC

The 73 LC and LCNEC cases were analyzed for mutations in 26 cancer-related genes through NGS-based analysis of fresh frozen or FFPE tumor tissues. In total, 117 nonsynonymous variants, with alternate variant frequencies (the fraction of all reads with the detected variant) between 3–91% (Lund cohort only), were identified in 13 genes, for which gene copy number status were also extracted (Figure 1 and Supplementary Figure S1 and Supplementary Tables S2–S3 listing explicit variant data). Median number of variants per sample was one and maximum was three. 72 out of 73 cases, including all Lund cases, showed variants in at least one gene.

TP53 mutations were the most dominant alteration in both LC and LCNEC tumors (83% and 88% of cases, respectively). *TP53* mutations typically manifested as missense mutations in active protein domains and nonsense mutations in between active domains (Figures 1 and Supplementary Figure S1). Remaining alterations were found in considerably lower numbers in both subgroups. In LC, *KRAS* and *MET* were the second and third most frequently mutated genes (22% and 12%, respectively), while corresponding genes in LCNEC were *STK11* and *PTEN* (16% and 13%, respectively) (Figure 1, Supplementary Table S3). These alterations highlight a more general difference between the two subgroups, regarding alterations in oncogenes versus tumor suppressors. Here, LC cases typically showed more alterations in oncogenes compared to LCNECs. Specifically, 20 oncogene alterations in *BRAF*, *GNAQ*, *GNAS*, *KRAS*, *KIT*, *MET*, *NRAS*, *MAP2K1/MEK1*, and *PIK3CA* were found in 44% of the LC cases as compared to eight alterations affecting 22% of LCNEC cases (Supplementary Table S3). Notably, for the two *KRAS* alterations observed in LCNEC cases, one was not in the active RAS protein domain (a *KRAS* M1I mutation). The second, a G12C mutation, was found in both the LCNEC and the LC component of the included multicomponent tumor, with different alternate allele frequencies (40.3% in the LCNEC and 7.9% and in the LC component of the tumor). Together, this suggests that activating *KRAS* mutations are in fact rare in LCNEC. Consistent with a general idea of a limited number of oncogene hits required to activate a tumorigenic pathway, we observed only one case in each histological subgroup with >1 mutation in any of the eight oncogenes.

Two additional differences regarding oncogene mutations may be noted. Firstly, *KIT* mutations were exclusively found in LCNEC cases ($n = 3$, 9%). Secondly, all *TP53* wild type LC cases (17% of all LC cases) harbored oncogene mutations in *KRAS*, *MET* or *PIK3CA* (Figure 1A). This observation suggests that these *TP53* wild type tumors may be more dependent on oncogene

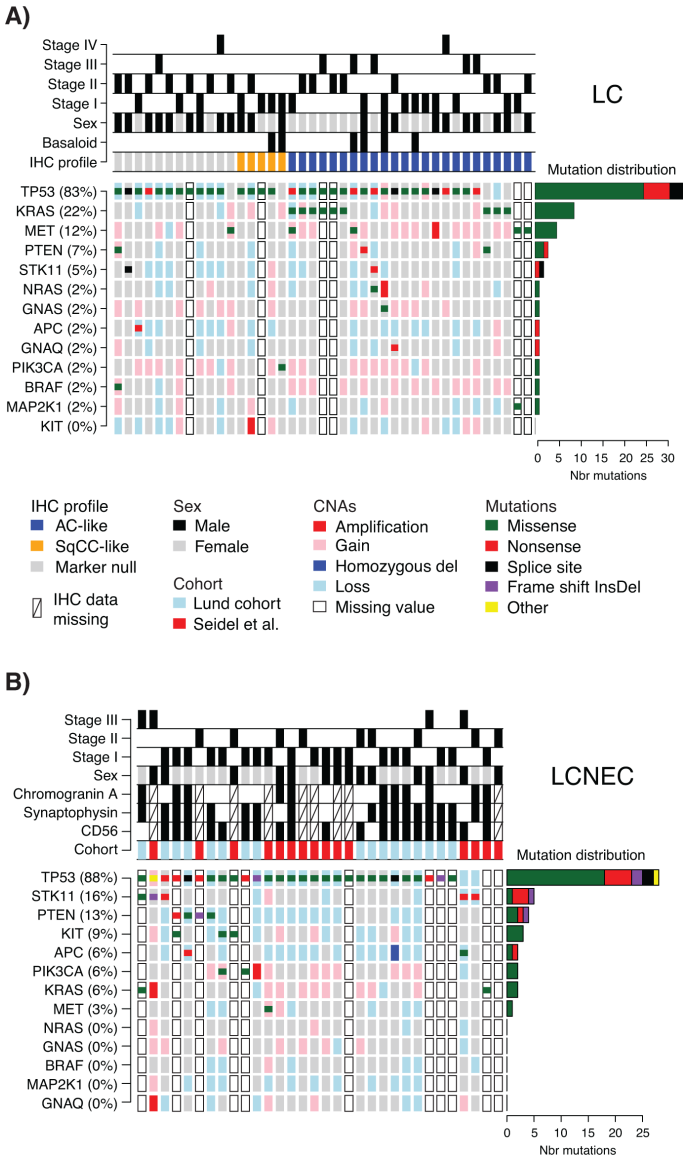


Figure 1: Detected mutations and copy number alterations in LC and LCNEC. A. Detected gene variants and copy number alterations (CNAs) (rows) in 41 LC cases (columns), ordered by immunomarker profile of adenocarcinoma-like (AC-like), squamous cell carcinoma like (SqCC-like), or marker null phenotype (TTF-1/Napsin A and CK5/P40 negative). Copy number status is shown as larger background rectangles and mutations as squares for each sample and gene. Right side bar plot summarizes the distribution of the different mutation types for each gene. **B.** Detected variants and copy number alterations in 32 LCNEC cases displayed as in A. Samples are ordered according to gene variant frequency.

activation alone for sustained tumor development. However, in this relatively small retrospective cohort we did not find support for differences related to tumor stage or gender ($p = 0.69$ and $p = 0.42$, respectively, Fisher's exact test), or patient outcome (overall survival, log-rank $p > 0.05$) between these tumors and *TP53*-mutated LCs.

Finally, high-level copy number gain (amplifications), or low copy deletions (putative homozygous deletions) were very scarce in analyzed cases (only single cases with *NRAS*, *KRAS*, *GNAQ*, *MET*, *KIT*, or *PIK3CA* amplifications, Figure 1), and no distinct cases of monoallelic amplification of mutated oncogene alleles were observed. For several tumor suppressors (*TP53*, *STK11*, *PTEN*, and *APC*), in especially LCNEC cases, we observed apparent support of Knudson's multiple-hit hypothesis [21], with DNA mutation and associated copy number loss (Figure 1).

DNA mutations in histopathological and immunohistological subgroups of LC and LCNEC

Basaloid tumors represent a rare histopathological subgroup of LC characterized by specific cytological and tissue architectural characteristics [3]. In our LC cohort, six cases were subclassified as basaloid cancer. When viewed as two subgroups, i.e., basaloid versus non-basaloid LC, there were no differences in mutation frequencies between the groups for two of the most commonly mutated genes, *TP53* and *MET*. In contrast, no *KRAS* mutations were observed in basaloid cases, consistent with Rossi et al. [6].

Recently, patient outcome and specific oncogene mutations in LC tumors have been associated with tumor subgroups defined by positive expression of adenocarcinoma (TTF-1/Napsin A) or squamous cell carcinoma (CK5/P40) immunohistochemistry markers [5]. In our LC cohort, 24 tumors (59%) were positive by immunohistochemistry for TTF-1/Napsin A (referred to as adenocarcinoma-like), 5 tumors (12%) were CK5/P40 positive (squamous-like), whereas 12 tumors (29%) did not express any of these IHC markers (marker null cases). Stratification of identified mutations by IHC subgroup revealed a striking enrichment of oncogene mutations in adenocarcinoma-like LC tumors (85% of all oncogene mutations, affecting 63% of these cases), including all nine *KRAS* mutations and the single *NRAS* (G12D) and *MAP2K1*/*MEK1* (K57N) mutations (Figure 1A, Supplementary Tables S2 and S3). These *KRAS* mutations were all typical driver mutations located in codon 12 (one G12S, two G12C, three G12V mutations), 13 (one G13C and one G13D mutation), and 61 (one Q61K mutation), suggesting that these represent likely driver events in the affected tumors. In contrast, 92% of marker null cases carried a *TP53* mutation, but only 17% of cases had an oncogene mutation (one *BRAF* Q456K and one *MET* T1010I mutation). Moreover, marker null cases showed a poorer overall survival compared to adenocarcinoma-like cases ($p = 0.007$, log-rank

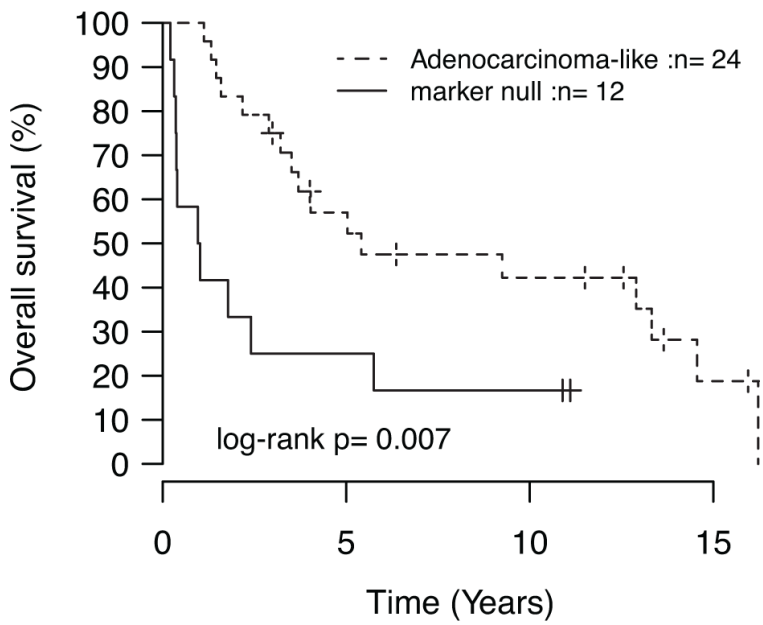


Figure 2: Kaplan-Meier analysis of the association with overall survival for immunomarker-defined subgroups of LC. *P*-value calculated using the log-rank test.

test, Figure 2). The poorer outcome of marker null LC cases compared to adenocarcinoma-like cases was significant also in multivariate analysis including immunomarker stratification, tumor stage, and gender as covariates and overall survival as clinical endpoint (Hazard ratio = 4.4, 95% Confidence interval = 1.5–12.5, $p = 0.006$ for marker null stratification).

For LCNEC cases, we observed no association between DNA mutations and IHC expression of the chromogranin A, synaptophysin or CD56 neuroendocrine markers.

Validation analyses of NGS DNA mutation results

To validate the NGS platform we first analyzed seven independent tumor FFPE samples from lung, colon and melanoma with known mutations in *BRAF* (V600E), *KRAS* (G12S, G13D), and *EGFR* (L858R and E746_A750del). The variant allele frequency for these alterations by pyrosequencing ranged between 27.7–41.8%. All mutations could successfully be identified by the NGS platform.

Secondly, we selected the ten LC and LCNEC cases from the Lund cohort harboring *KRAS* mutations (variant allele frequencies between 3.1–41%), and validated them using quantitative PCR (Qiagen Therascreen). Eight *KRAS* mutations were correctly identified, while the remaining two, Q61K and G13C, were not covered by the Therascreen assay.

ALK, RET, and ROS1 gene fusion analysis in LC and LCNEC

The ability of the Archer FusionPlex assay to identify gene fusions in *ALK*, *RET* and *ROS1* was successfully validated in four cell lines with known fusion gene rearrangements, HCC78 (*ROS1-SLC34A2*), KARPAS-299 (*ALK-NPM1*), LC-2/ad (*CCDC6-RET*), and H2228 (*EML4-ALK*).

Fusion gene analysis was performed on all 58 Lund cases using RNA from fresh frozen ($n = 47$) or FFPE ($n = 11$) tumor tissues. However, only 46 fresh frozen tumors passed the initial Archer data quality analysis steps after sequencing (35 LC, 11 LCNEC). The failure of the FFPE cases is likely due to extensive RNA degradation in the tissue blocks caused by the fixation process and subsequent storage. In the 46 analyzable cases, we identified no *RET* or *ROS1* fusions. Only one analyzed tumor, a LCNEC case, showed a candidate *ALK* gene fusion event (*DNBL-ALK*) based on NGS data, however just with the minimum number of reads required for reporting (Supplementary Methods). However, confirmatory ALK IHC and FISH analysis could not confirm protein overexpression or an actual gene fusion event in this case.

DISCUSSION

In the current study, we have explored the mutational spectrum of 26 well-established cancer-related genes and *ALK*, *RET*, and *ROS1* gene fusions by massive parallel sequencing in a large panel of thoroughly histopathologically classified primary LC and LCNEC lung cancers. In comparison to existing methods, NGS-based methods for DNA variant detection generally offer higher sensitivity in detecting low-frequency variants. Together with the specific feature of bi-directional sequencing in the Illumina TruSight Tumor assay this allows for sensitive variant detection also in FFPE samples. Besides the presented molecular characterization of LC and LCNEC tumors, the current study also supports the feasibility of using NGS-based methods for analysis of treatment predictive DNA alterations in routine clinical lung FFPE tumor tissues.

In LC as a whole, our findings of frequent *KRAS* mutations and less frequent alterations in *BRAF*, *MAP2K1*, and *PIK3CA* are in agreement with previous studies [5, 6]. Similarly, in LCNEC the high mutation rate of *TP53* and the scarcity of *KRAS* mutations have also been reported before (see, e.g., Rossi et al. [4]). The similar frequency of *TP53* mutations between the LC and LCNEC group mimics findings of similar p53 protein expression by Iyoda et al. [11]. Due to a paucity of studies, the roles of *PTEN*, *STK11* and *MET* mutations in LC and LCNEC are largely unknown. In our study, alterations in the tumor suppressors *PTEN* and *STK11* were mainly observed together with *TP53* mutations in both LC and LCNEC, while *MET* mutations were more often found in *TP53* wild type LC cases. Although LCNEC tumors have been shown to strongly express receptor tyrosine kinases such as *KIT*, *PDGFRA*, *PDGFRB* and *MET*, compared to other NSCLC groups, there is less support of mutations being the underlying cause for the elevated expression [4, 18]. Supporting these results, we identified no *PDGFRA* mutations in any of the tumors, only one *MET* mutation in the LCNEC group, whereas three *KIT* mutations were found exclusively in LCNEC cases. However, the impact of some of these mutations is difficult to assess without functional characterization, as all do not occur in active protein domains (see Supplementary Figure S1).

In this study, alterations in oncogenes, with exception of *KIT* alterations, are generally more frequent in LCs when considered a single entity compared to LCNEC. However, it is becoming apparent that the mutational spectrums in LC and LCNEC are different based on recent whole-exome sequencing studies. Specifically, LCNEC has been suggested to be more similar to SCLC [8], in line with the similarity of LCNEC and SCLC on the morphological, immunohistochemical, transcriptional, copy number, and epigenetic levels [4, 8–10]. Consistently, SCLC have recently been reported to harbor high frequencies of *TP53* mutations and similar frequencies of *KIT*, *PIK3CA* and *KRAS* alterations as for the LCNEC cases in the current study [22]. Together, our observations further support that LC and LCNEC follow different evolutionary paths.

Stratification of LC cases based on immunomarkers for adenocarcinoma (TTF-1/Napsin A) and squamous cell carcinoma (CK5/P40) revealed that 71% of the cases could be classified as variants of adenocarcinoma or SqCC. This observation is in line with previous reports (59–90%) [5, 23–27], although the observation of 29% of LC as marker null is on the higher end compared to the literature. One reason for this could be that we in this retrospective cohort used the TTF-1 8G7G3/1 clone that compared to the SPT24 clone is slightly less sensitive, whereas the SPT24 clone yields more cases positive for both CK5/P40 and TTF-1. Likewise, we used CK5 and P40 as markers of squamous cell carcinoma, while P63 may be more sensitive (but also less specific). Irrespectively, our data demonstrates the value of using multiple immunomarkers for undifferentiated lung cancers. While no apparent differences in oncogene amplification frequency could be observed between immunomarker-defined subgroups, the subgroups showed a distinctively different spectrum of especially oncogene mutations. Adenocarcinoma-like LCs (59% of all analyzed LC cases) harbored the overwhelming majority of detected oncogene mutations (85%), affecting 63% of these tumors. By comparison, only 17% of marker null LC cases showed oncogene mutations. Thus, adenocarcinoma-like LC appears to represent a more oncogene driven subgroup compared to CK5/P40 positive tumors and marker null LCs. These findings are in excellent agreement with Rekhtman et al. [5], including the observation of the single *PIK3CA* mutation in a CK5/P40 positive case, and a poorer overall survival for marker null patients compared to TTF-1/Napsin A positive LC patients. Despite the retrospective nature of the patient material, our results in combination with other recent molecular studies clearly challenge LC as an independent tumor entity on the molecular level, supporting that the current LC definition rather includes a heterogeneous collection of poorly differentiated tumors from other NSCLC subgroups [5, 6, 8]. In fact, supported by both clinicopathological and molecular studies the recent 2015 WHO classification of lung cancer now stress that the term LC should now only be used for undifferentiated tumors not expressing pneumocyte or squamous markers [7]. Importantly, in the 2004 WHO classification the LC definition provides little molecular information for a predictive molecular testing strategy to guide individualized treatment for this patient cohort [5]. A refined stratification of LC based on molecular characteristics may therefore have considerable impact on diagnosis, predictive molecular testing and in the end, therapy selection [5, 6, 8].

In contrast to the immunomarker-defined LC subgroups, less is known whether LCNEC tumors may be divided into similar subgroups. In the current study we found no associations of the neuroendocrine markers used to identify LCNEC tumors with specific mutations. This lack of association may be because these markers do not represent putative subgroups at all, and/or, as indicated by our analyses, that the mutational landscape in LCNEC is different in respect to, especially, oncogene drivers compared to non-neuroendocrine NSCLC. Clearly, further genomic studies of both marker null LC (undifferentiated LC) and LCNEC tumors including large scale sequencing approaches, gene expression profiling, and DNA methylation profiling are needed to further characterize these tumor groups.

EGFR mutations and *ALK* gene fusions are the key molecular treatment predictive alterations for targeted therapy in lung cancer today [12]. However, both alterations are scarce in LC and LCNEC tumors [5, 6, 14–16], consistent with our findings of no *EGFR* mutations or validated *ALK* gene fusions in either LC or LCNEC tumors. Specifically, the absence of *ALK* rearrangements in our LC cohort compared to the few *ALK* rearranged cases reported by Rekhtman et al. [5] is consistent with that our cohort comprises only of known smokers, while *ALK* rearrangements in the former study were found in never or light smokers. In recent studies, lung cancer patients with tumors harboring *ROS1* or *RET* gene fusions have shown notable responses to ALK or other multi-target kinase inhibitors [28, 29]. Similar to *ALK* fusions, the frequency of these alterations in LC and LCNEC is largely unknown, but may be expected to be very low. Consistently, we found no *RET* or *ROS1* gene fusions in LC or LCNEC tumors based on targeted RNA sequencing.

In summary, the current study adds further insights into the mutational landscape of LC and LCNEC, supporting that these tumor subgroups follow different tumorigenic paths. Moreover, our study supports that LC may be refined by molecular and immunomarkers into clinically relevant subgroups that may have implications for diagnosis, and therapy decisions. Despite the identification of adenocarcinoma-like LC as a subset of tumors with a potentially high frequency of forthcoming therapeutically relevant driver mutations, a continued search for additional molecular targets for therapeutic inhibition in non-adenocarcinoma NSCLC is warranted.

Authors' contributions

JS, AK and MP conceived of the study. AK, JS, FR, and CR performed the mutational analyzes and gene fusion experiments. HC performed the cell culturing. KJ and KEL provided validation samples and data. MJ and KEL assisted in validation experiments. AK and JS performed statistical analyses with support from GJ. HB performed pathological evaluations and IHC analyses. PJ and MP included patients. JS drafted and wrote the manuscript with support from AK and GJ. The study was methodologically and scientifically supported by ÅB. All authors approved the final manuscript.

FUNDING

Financial support for this study was provided by the Swedish Cancer Society, the Knut & Alice Wallenberg Foundation, the Foundation for Strategic Research through the Lund Centre for Translational Cancer Research (CREATE Health), the Mrs Berta Kamprad Foundation, the Gunnar Nilsson Cancer Foundation, the Swedish Research Council, the Lund University Hospital Research Funds, the Gustav V:s Jubilee Foundation, and governmental funding (ALF).

CONFLICTS OF INTEREST

Authors declare that they have no competing interests.

REFERENCES

1. Asamura H, Kameya T, Matsuno Y, Noguchi M, Tada H, Ishikawa Y, Yokose T, Jiang SX, Inoue T, Nakagawa K, Tajima K, Nagai K. Neuroendocrine neoplasms of the lung: a prognostic spectrum. J Clin Oncol. 2006; 24:70–76.
2. Battafarano RJ, Fernandez FG, Ritter J, Meyers BF, Guthrie TJ, Cooper JD, Patterson GA. Large cell neuroendocrine carcinoma: an aggressive form of non-small cell lung cancer. J Thorac Cardiovasc Surg. 2005; 130:166–172.
3. Travis W.D. BE, Muller-Hermelink H.K., Harris C.C. (Eds.). (2004). World Health Organization Classification of Tumours. Pathology and Genetics of Tumours of the Lung, Pleura, Thymus and Heart. (Lyon: IARC Press).
4. Rossi G, Cavazza A, Marchioni A, Longo L, Migaldi M, Sartori G, Bigiani N, Schirosi L, Casali C, Morandi U, Facciolongo N, Maiorana A, Bavieri M, et al. Role of chemotherapy and the receptor tyrosine kinases KIT, PDGFRalpha, PDGFRbeta, and Met in large-cell neuroendocrine carcinoma of the lung. J Clin Oncol. 2005; 23:8774–8785.
5. Rekhtman N, Tafe LJ, Chaft JE, Wang L, Arcila ME, Colanta A, Moreira AL, Zakowski MF, Travis WD, Sima CS, Kris MG, Ladanyi M. Distinct profile of driver mutations and clinical features in immunomarker-defined subsets of pulmonary large-cell carcinoma. Mod Pathol. 2014; 26:511–522.
6. Rossi G, Mengoli MC, Cavazza A, Nicoli D, Barbareschi M, Cantaloni C, Papotti M, Tironi A, Graziano P, Paci M, Stefani A, Migaldi M, Sartori G, et al. Large cell carcinoma of the lung: clinically oriented classification integrating immunohistochemistry and molecular biology. Virchows Arch. 2014; 464:61–68.
7. Travis W.D. BE, Burke A.P., Marx A, Nicholson A.G. (Eds.). (2015). WHO Classification of Tumours of the Lung, Pleura, Thymus and Heart. (Lyon: IARC Press).
8. CLCGP. A genomics-based classification of human lung tumors. Sci Transl Med. 2013; 5: 209ra13.
9. Staaf J, Isaksson S, Karlsson A, Jonsson M, Johansson L, Jonsson P, Botling J, Micke P, Baldetorp B, Planck M. Landscape of somatic allelic imbalances and copy number alterations in human lung carcinoma. International journal of cancer. 2012; 1:2020–2031.
10. Karlsson A, Jonsson M, Lauss M, Brunnstrom H, Jonsson P, Borg A, Jonsson G, Ringner M, Planck M, Staaf J. Genome-wide DNA Methylation Analysis of Lung Carcinoma Reveals One Neuroendocrine and Four Adenocarcinoma Epitypes Associated with Patient Outcome. Clin Cancer Res. 2014; 20:6127–6140.
11. Iyoda A, Hiroshima K, Moriya Y, Mizobuchi T, Otsuji M, Sekine Y, Shibuya K, Iizasa T, Saitoh Y, Fujisawa T. Pulmonary large cell neuroendocrine carcinoma demonstrates high proliferative activity. Ann Thorac Surg. 2004; 77:1891–1895; discussion 1895.
12. Leighl NB, Rekhtman N, Biermann WA, Huang J, Mino-Kenudson M, Ramalingam SS, West H, Whitlock S, Somerfield MR. Molecular Testing for Selection of Patients With Lung Cancer for Epidermal Growth Factor Receptor and Anaplastic Lymphoma Kinase Tyrosine Kinase Inhibitors: American Society of Clinical Oncology Endorsement of the College of American Pathologists/International Association for the Study of Lung Cancer/Association for Molecular Pathology Guideline. J Clin Oncol. 2014.
13. Nakamura H, Tsuta K, Yoshida A, Shibata T, Wakai S, Asamura H, Furuta K, Tsuda H. Aberrant anaplastic lymphoma kinase expression in high-grade pulmonary neuroendocrine carcinoma. J Clin Pathol. 2013; 66:705–707.
14. Gainor JF, Varghese AM, Ou SH, Kabraji S, Awad MM, Katayama R, Pawlak A, Mino-Kenudson M, Yeap BY, Riely GJ, Iafrate AJ, Arcila ME, Ladanyi M, et al. ALK rearrangements are mutually exclusive with mutations in EGFR or KRAS: an analysis of 1, 683 patients with non-small cell lung cancer. Clin Cancer Res. 2013; 19:4273–4281.
15. De Pas TM, Giovannini M, Manzotti M, Trifiro G, Toffalorio F, Catania C, Spaggiari L, Labianca R, Barberis M. Large-cell neuroendocrine carcinoma of the lung harboring EGFR mutation and responding to gefitinib. J Clin Oncol. 2011; 29:e819–822.
16. Aroldi F, Bertocchi P, Meriggi F, Abeni C, Ogliosi C, Rota L, Zambelli C, Bna C, Zaniboni A. Tyrosine Kinase Inhibitors in EGFR-Mutated Large-Cell Neuroendocrine Carcinoma of the Lung? A Case Report. Case Rep Oncol. 2014; 7:478–483.
17. Travis WD, Brambilla E, Noguchi M, Nicholson AG, Geisinger K, Yatabe Y, Powell CA, Beer D, Riely G, Garg K, Austin JH, Rusch VW, Hirsch FR, et al. International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society: international multidisciplinary classification of lung adenocarcinoma: executive summary. Proc Am Thorac Soc. 2011; 8:381–385.

18. Salomonsson A, Jonsson M, Isaksson S, Karlsson A, Jonsson P, Gaber A, Bendahl PO, Johansson L, Brunnstrom H, Jirstrom K, Borg A, Staaf J, Planck M. Histological specificity of alterations and expression of KIT and KITLG in non-small cell lung carcinoma. Genes Chromosomes Cancer. 2013; 52:1088–1096.

19. Brunnstrom H, Johansson L, Jirstrom K, Jonsson M, Jonsson P, Planck M. Immunohistochemistry in the differential diagnostics of primary lung cancer: an investigation within the Southern Swedish Lung Cancer Study. Am J Clin Pathol. 2013; 140:37–46.

20. Karlsson A, Ringner M, Lauss M, Botling J, Micke P, Planck M, Staaf J. Genomic and transcriptional alterations in lung adenocarcinoma in relation to smoking history. Clin Cancer Res. 2014; 20:4912–4924.

21. Knudson AG Jr. Mutation and cancer: statistical study of retinoblastoma. Proceedings of the National Academy of Sciences of the United States of America. 1971; 68:820–823.

22. Ross JS, Wang K, Elkadi OR, Tarasen A, Foulke L, Sheehan CE, Otto GA, Palmer G, Yelensky R, Lipson D, Chmielecki J, Ali SM, Elvin J, et al. Next-generation sequencing reveals frequent consistent genomic alterations in small cell undifferentiated lung cancer. J Clin Pathol. 2014; 67:772–776.

23. Au NH, Cheang M, Huntsman DG, Yorida E, Coldman A, Elliott WM, Bebb G, Flint J, English J, Gilks CB, Grimes HL. Evaluation of immunohistochemical markers in non-small cell lung cancer by unsupervised hierarchical clustering analysis: a tissue microarray study of 284 cases and 18 markers. The Journal of pathology. 2004; 204:101–109.

24. Barbareschi M, Cantaloni C, Del Vescovo V, Cavazza A, Monica V, Carella R, Rossi G, Morelli L, Cucino A, Silvestri M, Tirone G, Pelosi G, Graziano P, et al. Heterogeneity of large cell carcinoma of the lung: an immunophenotypic and miRNA-based analysis. Am J Clin Pathol. 2011; 136:773–782.

25. Monica V, Ceppi P, Righi L, Tavaglione V, Volante M, Pelosi G, Scagliotti GV, Papotti M. Desmocollin-3: a new marker of squamous differentiation in undifferentiated large-cell carcinoma of the lung. Mod Pathol. 2009; 22:709–717.

26. Monica V, Scagliotti GV, Ceppi P, Righi L, Cambieri A, Lo Iacono M, Saviozzi S, Volante M, Novello S, Papotti M. Differential Thymidylate Synthase Expression in Different Variants of Large-Cell Carcinoma of the Lung. Clin Cancer Res. 2009; 15:7547–7552.

27. Pardo J, Martinez-Penuela AM, Sola JJ, Panizo A, Gurpide A, Martinez-Penuela JM, Lozano MD. Large cell carcinoma of the lung: an endangered species? Applied immunohistochemistry & molecular morphology: AIMM / official publication of the Society for Applied Immunohistochemistry. 2009; 17:383–392.

28. Shaw AT, Ou SH, Bang YJ, Camidge DR, Solomon BJ, Salgia R, Riely GJ, Varella-Garcia M, Shapiro GI, Costa DB, Doebele RC, Le LP, Zheng Z, et al. Crizotinib in ROS1-rearranged non-small-cell lung cancer. The New England journal of medicine. 2014; 371:1963–1971.

29. Drilon A, Wang L, Hasanovic A, Suehara Y, Lipson D, Stephens P, Ross J, Miller V, Ginsberg M, Zakowski MF, Kris MG, Ladanyi M, Rizvi N. Response to Cabozantinib in patients with RET fusion-positive lung adenocarcinomas. Cancer Discov. 2013; 3:630–635.

Study V

Biology of Human Tumors

Clinical Cancer Research

Genome-wide DNA Methylation Analysis of Lung Carcinoma Reveals One Neuroendocrine and Four Adenocarcinoma Epitypes Associated with Patient Outcome

Anna Karlsson[1], Mats Jönsson[1], Martin Lauss[1], Hans Brunnström[1], Per Jönsson[2], Åke Borg[1,3], Göran Jönsson[1,3], Markus Ringnér[1,3], Maria Planck[1], and Johan Staaf[1,3]

Abstract

Purpose: Lung cancer is the worldwide leading cause of death from cancer. DNA methylation in gene promoter regions is a major mechanism of gene expression regulation that may promote tumorigenesis. However, whether clinically relevant subgroups based on DNA methylation patterns exist in lung cancer remains unclear.

Experimental Design: Whole-genome DNA methylation analysis using 450K Illumina BeadArrays was performed on 12 normal lung tissues and 124 tumors, including 83 adenocarcinomas, 23 squamous cell carcinomas (SqCC), 1 adenosquamous cancer, 5 large cell carcinomas, 9 large cell neuroendocrine carcinomas (LCNEC), and 3 small-cell carcinomas (SCLC). Unsupervised bootstrap clustering was performed to identify DNA methylation subgroups, which were validated in 695 adenocarcinomas and 122 SqCCs. Subgroups were characterized by clinicopathologic factors, whole-exome sequencing data, and gene expression profiles.

Results: Unsupervised analysis identified five DNA methylation subgroups (epitypes). One epitype was distinctly associated with neuroendocrine tumors (LCNEC and SCLC). For adenocarcinoma, remaining four epitypes were associated with unsupervised and supervised gene expression phenotypes, and differences in molecular features, including global hypomethylation, promoter hypermethylation, genomic instability, expression of proliferation-associated genes, and mutations in *KRAS*, *TP53*, *KEAP1*, *SMARCA4*, and *STK11*. Furthermore, these epitypes were associated with clinicopathologic features such as smoking history and patient outcome.

Conclusions: Our findings highlight one neuroendocrine and four adenocarcinoma epitypes associated with molecular and clinicopathologic characteristics, including patient outcome. This study demonstrates the possibility to further subgroup lung cancer, and more specifically adenocarcinomas, based on epigenetic/molecular classification that could lead to more accurate tumor classification, prognostication, and tailored patient therapy. *Clin Cancer Res; 20(23); 6127–40. ©2014 AACR.*

Introduction

Lung cancer is currently the leading cause of death from cancer worldwide (1). The disease is broadly divided into small-cell lung cancer (SCLC; ∼15% of all cases) and non–small cell lung cancer (NSCLC). NSCLC is further divided into adenocarcinoma, squamous cell carcinoma (SqCC), and large-cell carcinoma with or without neuroendocrine features (LCNEC and LC, respectively). Lung cancer is a molecularly heterogeneous disease involving different alterations that drive tumorigenesis, including DNA sequence alterations, copy number alterations (CNAs), and epigenetic modifications, such as DNA methylation and histone/chromatin modifications. DNA methylation at CpG dinucleotides in gene promoter regions is a major mechanism of gene expression regulation, and aberrant promoter hypermethylation may lead to inactivation of tumor suppressor genes, thereby promoting tumorigenesis (2).

Genome-wide DNA methylation profiling of NSCLC have identified epigenetic subgroups (epitypes) of tumors associated with characteristic molecular alterations and prognosis (3–7). NSCLC/adenocarcinomas with increased promoter methylation levels have been highlighted, and termed CpG island methylator phenotype (CIMP),

[1]Division of Oncology and Pathology, Department of Clinical Sciences Lund, Lund University, Lund, Sweden. [2]Department of Thoracic Surgery, Lund University and Skåne University Hospital, Lund, Sweden. [3]CREATE Health Strategic Center for Translational Cancer Research, Lund University, Lund, Sweden.

Note: Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

Corresponding Author: Johan Staaf, Division of Oncology and Pathology, Department of Clinical Sciences Lund, Lund University, Medicon Village, Lund SE-22381, Sweden. Phone: +46462221444; Fax: +4646147327; E-mail: johan.staaf@med.lu.se

doi: 10.1158/1078-0432.CCR-14-1087

Translational Relevance

DNA methylation in gene promoters is a major mechanism of gene expression regulation that may promote tumorigenesis. DNA methylation of specific genes, and/or patterns of DNA methylation in lung cancers have been associated with patient outcome. However, hitherto, neither the existence of reproducible DNA methylation-based subgroups of potential clinical relevance nor the DNA methylation pattern across multiple histologic subgroups has been carefully investigated in lung cancer. On the basis of a multicohort approach, we conducted a comprehensive survey of genome-wide DNA methylation patterns in lung cancer identifying one neuroendocrine and four reproducible adenocarcinoma epitypes. Importantly, epitypes were associated with specific clinicopathologic and molecular characteristics, gene expression phenotypes, and patient outcome. These findings shed light on the epigenetic characteristics and molecular diversity underlying lung cancer. Moreover, they highlight the possibility to further subgroup the disease based on epigenetic/molecular classification, which could lead to improvements in tumor classification, prognostication, and tailored patient therapy.

stratifying tumors into CIMP-high, CIMP-low/negative, and CIMP-intermediate subgroups, in analogy to findings from other cancer forms (3, 4, 8). In addition, gene expression phenotypes like the bronchioid, magnoid, and squamoid subtypes in adenocarcinoma (9, 10) have also been associated with specific DNA methylation patterns (9). However, the proposed NSCLC epitypes have not been independently replicated. Moreover, genome-wide epigenetic patterns across multiple lung cancer histotypes have not yet been reported.

Herein, we investigated the landscape of DNA methylation in different histologic subgroups of lung cancer with the intention to derive methylation-based subgroups of clinical and molecular relevance. On the basis of a discovery cohort of 124 primary lung cancers, including all major histologic subgroups, we found a specific DNA methylation pattern of neuroendocrine tumors and identified four epitypes of adenocarcinoma that were subsequently validated in 817 independent NSCLC cases. Epitypes were associated with molecular and clinicopathologic differences, and linked to gene expression phenotypes based on integration with DNA sequencing and gene expression data. Together, our findings highlight the possibility to further subgroup lung cancer based on epigenetic/molecular classification, providing a clear refinement of previously suggested models and a more accurate tumor classification, which could lead to new targets for diagnostics, therapeutic intervention, and prognostication of the disease.

Materials and Methods

Patient material

DNA and total RNA were extracted from the same tissue piece for 124 tumors and 12 matched normal lung tissue specimens from patients with early-stage lung cancer operated at the Skåne University Hospital (Lund, Sweden; Table 1, discovery cohort). The study was approved by the Regional Ethical Review Board in Lund, Sweden (Registration no. 2004/762 and 2008/702). The 12 normal specimens originated from patients with adenocarcinoma and were mixed in gender (3 males, 9 females), smoking status (6 never-smokers, 6 smokers), and patient age (57–82 years). Three-hundred and seventy three adenocarcinomas from The Cancer Genome Atlas (TCGA) project (11) and 444 NSCLC cases from Sandoval and colleagues (ref. 5; Sandoval cohort) were used as validation cohorts (Table 1).

Global methylation analysis

All cases were analyzed by the Illumina Human Methylation 450K v1.0 platform (Illumina) according to manufacturer's instructions (Supplementary Materials and Methods). Signal intensities were obtained from GenomeStudio (Illumina), converted to β-values, filtered, and normalized to remove biases between Infinium I and II probes (Supplementary Materials and Methods). CpG probes with aberrant methylation in tumors compared with normal lung tissue in the discovery set were identified as described in Supplementary Materials and Methods (Supplementary Fig. S1A), and annotated through the human embryonic stem cell (H1hESC) chromatin state track (12) and the Illumina CpG island track. CpGs in repetitive elements were identified through the "repeats_rmsk_hg19" table from the UCSC Genome Browser. Unsupervised class-discovery was performed using bootstrap clustering (ref. 13; Supplementary Materials and Methods). Principal component analysis (14), including clinicopathologic and technical factors, and comparison of bisulfite conversion plate and beadchip id against unsupervised bootstrap clusters were performed to assess that no technical artifacts influenced methylation data, or bootstrap groups, for the 124-sample discovery cohort (Supplementary Fig. S1B–S1D). DNA methylation centroids representing bootstrap clusters were created from the average β-value for each CpG probe in respective cluster. Samples in validation cohorts were assigned to the centroid with the smallest Euclidean distance for matching CpGs. Methylation data for the discovery cohort is available as GSE60645 (15).

Copy number analysis

Log2 copy number estimates and CNAs for CpG probes in the discovery and Sandoval cohorts were generated and identified as described in Supplementary Materials and Methods from 450K methylation beadchip data. For the TCGA cohort, copy number estimates and CNAs were obtained from level 3 Affymetrix SNP6 data as described (16, 17). Complex arm-wise aberration index (CAAI) scores were calculated similar to Russnes and colleagues (ref. 18; Supplementary Materials and Methods).

Table 1. Patient characteristics and clinicopathologic data for included cohorts

	Lund cohort	Sandoval et al. (5)	TCGA (11)
Usage	Discovery	Validation	Validation
Total number of patients	124	444	373
Histology			
Adenocarcinoma	83	322	373
SqCC	23	122	—
LC	5	—	—
LCNEC	9	—	—
SCLC	3	—	—
Other	1 (Adenosquamous)	—	—
Tumor stage			
I	110	237	206
II	10	94	87
III	1	102	63
IV	—	11	16
Not available	3	—	1
Smoking history			
Never-smokers	20	47	57
Smokers	104	334	304
Gender			
Male	53	254	173
Female	71	190	200
Mutation status			
EGFR-mutated	12	—	49[a]
KRAS-mutated	24	—	100[a]
Patient outcome			
Outcome type[b]	OS (121/83)	RFS (198/155)	—
Evaluable for			
CNAs	X	X[c]	X
Gene expression	X	—	X
Mutation spectrum	—	—	X

NOTE: X, data available for analysis.

[a]Nonsilent mutations from Mutation Annotation Format (MAF) file.

[b]OS, overall survival; RFS: relapse-free survival. Number of patients with outcome data (NSCLC/adenocarcinoma).

[c]Only CN-FGA.

Global gene expression analysis

Gene expression analysis was performed on 117 tumors from the discovery cohort using Illumina Human HT-12 V4 microarrays, available as GSE60645 (15). TCGA adenocarcinoma expression data were obtained as RNASeq V2 data. Six correlated gene expression modules in lung cancer, representing different tumor and/or tumor environment associated processes, were derived as originally described by Fredlund and colleagues in GSE29016 (refs. 19, 20; Supplementary Materials and Methods; Supplementary Table S1). These expression modules included an immune response, a neuroendocrine, and a stroma/extra cellular matrix module. Data processing steps, including adenocarcinoma and SqCC molecular subtype classification (9, 21), correlation of methylation and expression data, and calculation of different expression metagenes are further described in Supplementary Materials and Methods.

Functional classification

Gene Ontology enrichment were performed using the DAVID Functional Annotation Tool (22) with the default human population background and a Bonferroni-adjusted $P < 0.05$ as significance threshold.

Results

Genome-wide DNA methylation patterns in lung cancer

We analyzed 124 lung tumors from five histologic subgroups for global DNA methylation patterns using Illumina 450K methylation arrays (Table 1, discovery cohort). Overall, DNA methylation in the tumors followed a distinct pattern along the gene coding sequence, with low methylation levels near the transcription start site and high methylation levels at gene bodies, 3′UTRs, and intergenic regions (Fig. 1A). Correlation analyses of DNA methylation and gene expression revealed a pattern of negative correlations at transcription start sites and more positive correlations in

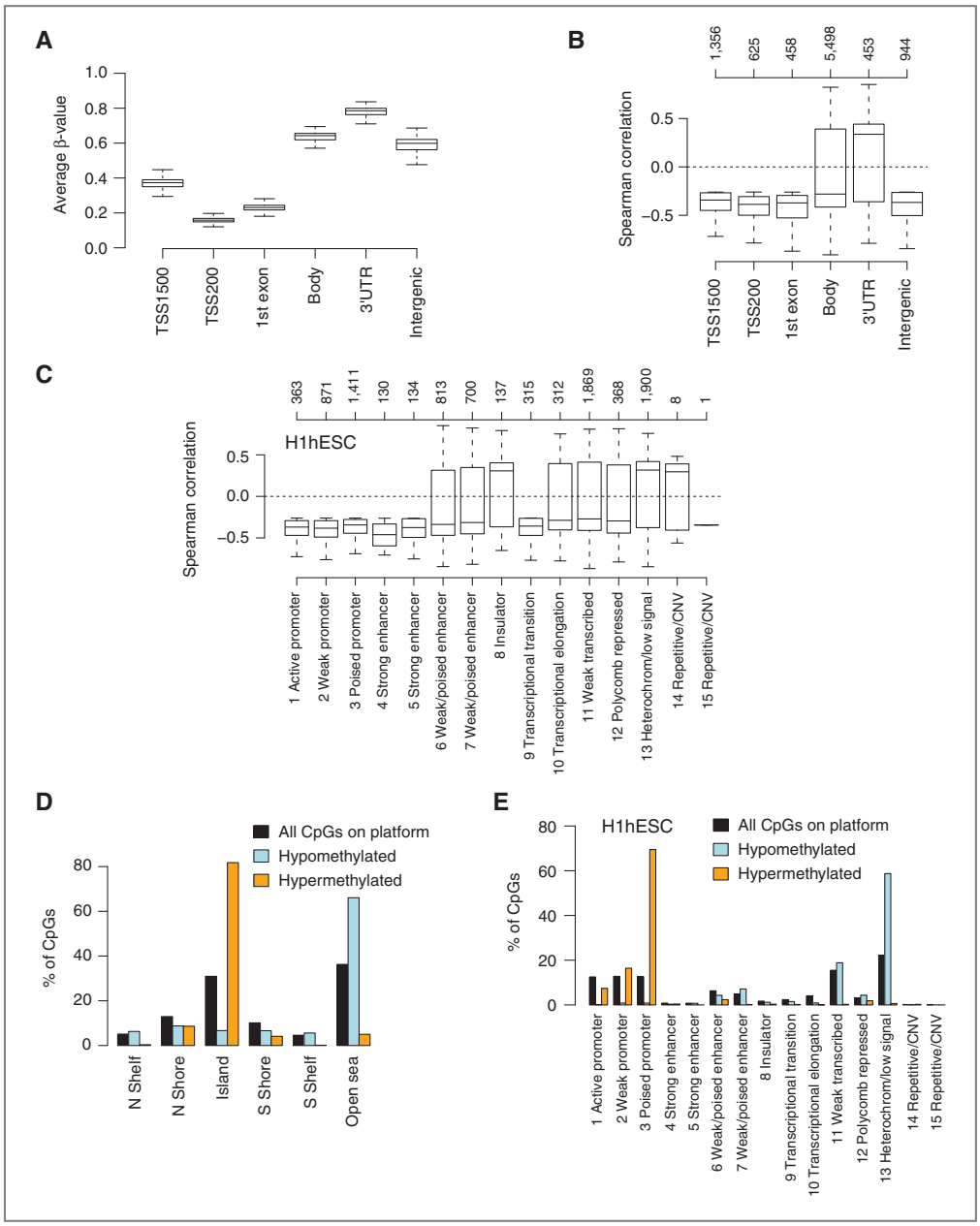


Figure 1. DNA methylation patterns in lung cancer. A, distribution of average β-values for 473864 CpGs stratified by Illumina gene location across the 124 tumors in the discovery cohort. TSS, transcription start site. B, Spearman correlation of DNA methylation and gene expression for 9,334 gene matching CpGs in 77 adenocarcinomas from the discovery cohort, stratified by Illumina gene location. Top axis indicates number of CpGs per group. C, Spearman correlation for the 9334 CpGs grouped according to the human embryonic stem cell (H1hESC) chromatin state track (12). (*Continued on the following page.*)

open sea/heterochromatin regions and gene bodies (Fig. 1B and C; Supplementary Table S2). We identified 4136 CpGs with aberrant methylation in >10% ($n = 13$) of tumors compared with normal lung tissue, including multiple *HOX* genes, Wnt signaling pathway genes, *APC*, *CDH13*, *GATA4*, *GATA5*, and *RASSF1* consistent with previous studies (3, 6) (Supplementary Fig. S1A; Supplementary Table S2). Hypomethylated CpGs in tumors were enriched in open sea/heterochromatin regions, whereas hypermethylated CpGs were typically located in transcription start sites, CpG islands, and poised promoters in human embryonic stem cells, H1hESC cells, consistent with previous reports (refs. 6, 23; Fig. 1D and E). Hypomethylated CpGs were enriched in repetitive regions (LINE, SINE, LTR elements) compared with hypermethylated CpGs (21% vs. 4%, respectively, Fisher exact $P = 9e{-}54$). Importantly, changing the number of CpGs with aberrant methylation by lowering or increasing the number of required tumors with aberrant methylation ($n = 2$–20 tumors equaling CpG sets between ~1,000 and 44,000 CpGs; Supplementary Fig. S1A) yielded the same enrichment pattern of hypomethylated and hypermethylated CpGs.

Functional annotation analysis of genes with hypermethylated CpGs in the 4136 CpG set showed enrichment of biologic processes such as regulation of transcription, neural development, and cell morphogenesis corroborating previous studies (24, 25), whereas hypomethylated genes showed a much less clear functional enrichment (Supplementary Table S2).

Unsupervised class discovery based on genome-wide DNA methylation patterns identifies five epitypes

Unsupervised bootstrap analysis based on the 4136 CpGs highlighted five tumor clusters in the discovery cohort, hereafter referred to as epitypes (ES1, ES2, ES3, ES4, and ES5; Fig. 2A and Supplementary Fig. S1E). Importantly, epitype association for individual samples was robust across different CpG sets (numbers between 1,282–17,710 CpGs) in exploratory bootstrap analysis (Supplementary Fig. S1F). ES1 showed a global hypomethylation pattern, ES4 a promoter methylation pattern, ES5 a methylation pattern resembling normal lung tissue, whereas ES2 had a pattern in between ES1 and ES4 (Fig. 2). Consistent with the global DNA hypomethylation pattern, ES1 tumors also showed more hypomethylation of CpGs in repetitive elements (Fig. 2C and Supplementary Fig. S1G). Notably, 89% of ES3 cases were either SCLC ($n = 2$) or LCNEC ($n = 6$) tumors. Consistent with the dominance of neuroendocrine cases in ES3, we found distinct overexpression of a neuroendocrine gene expression metagene compared with the other epitypes ($P = 5e{-}05$, Kruskal–Wallis test). Hence, we refer to ES3 as a neuroendocrine epitype. On the other hand, SqCC tumors clustered in ES1 (17%), ES2 (57%), and ES5 (22%) (Fig. 2A). A distinct association of SqCC cases in ES2 with the reported classical SqCC gene expression subtype (21) was found, with >86% of classical subtype classified SqCC cases present in this epitype. Adenocarcinomas ($n = 83$) were divided into ES1 (12%), ES2 (14%), ES4 (36%), and ES5 (37%; Fig. 3A).

Validation of lung cancer epitypes

To validate the identified epitypes from the discovery cohort, we created DNA methylation centroids for each epitype based on the 4136 CpGs. Next, we classified two independent cohorts analyzed by the same methylation platform (Sandoval and TCGA) comprising 122 SqCC tumors and 695 adenocarcinomas (Table 1). Principal component analysis performed in the validation cohorts confirmed that the centroid classification explained most of the total variation in DNA methylation compared with available clinicopathologic, technical (batch and beadchip data), and molecular factors, including clinical smoking history, sex, tumor stage, tumor size, histology (adenocarcinoma or SqCC), *EGFR*, *KRAS*, and *TP53* mutations (Supplementary Figs. S2A and S3A–S3C). Notably, most of these factors (e.g., smoking status) contributed little to the total variation in DNA methylation. Moreover, the classification of the validation cohorts was robust across different sets of CpGs, and overlapped extensively with independently derived unsupervised bootstrap groups in these cohorts (Supplementary Figs. S2B–S2D and S3D–S3F).

In both validation cohorts, $\leq$1% of cases were classified as ES3, supporting that this epitype is highly distinct for lung cancers expressing neuroendocrine marker genes. Similar to the discovery cohort, SqCC tumors in the Sandoval cohort were primarily classified as ES2 (49% of SqCC cases) or ES5 (33%). Although LC, LCNEC, and SqCC tumors were present in different clusters in the discovery set, this cohort is underpowered to robustly claim existence of different epitypes within these histologic subgroups. Moreover, there currently exist no comparable LC, LCNEC, or SCLC cohorts suitable for validation of novel epitypes within these subgroups. Consequently, we hereafter focus the characterization and validation of the epitypes only on lung adenocarcinomas in the three cohorts (excluding ES3), using clinicopathologic factors, gene expression data, CNAs, and mutational data. Fig. 3A shows the distribution of adenocarcinomas between epitypes in all investigated cohorts.

Adenocarcinoma epitypes are associated with reproducible clinicopathologic characteristics including smoking history, *EGFR*, and *KRAS* mutations

The epitypes showed differences in the composition of never-smokers and smokers. ES5 was enriched for never-smokers in both the discovery and Sandoval cohorts (63–68% of all never-smokers), while less in the TCGA cohort

(*Continued.*) Digits correspond to track state id. The number of CpGs per group is indicated on top of the panel. D, 4,136 CpGs were selected on the basis of variation in at least 13 of 124 tumor cases compared with 12 normal lung tissues and grouped according to Illumina CpG island annotations. Orange bars correspond to CpGs hypermethylated in tumors, blue bars correspond to CpGs hypomethylated in tumors, and black bars correspond to distribution of all CpGs on the Illumina platform. E, the 4,136 CpGs were grouped according to the human embryonic stem cell (H1hESC) chromatin state track.

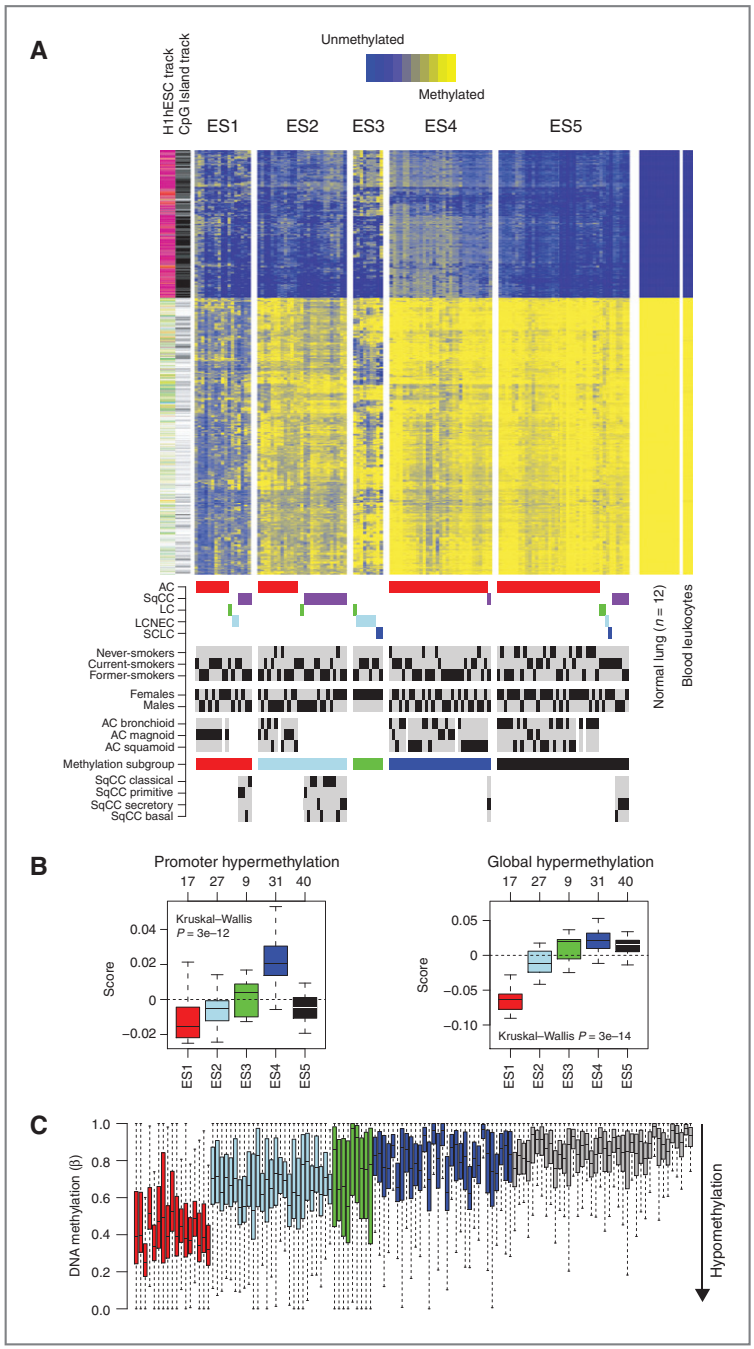


Figure 2. Identification of five DNA methylation subtypes in the discovery cohort. A, DNA methylation subtypes in 124 lung cancers based on bootstrap clustering of 4,136 variant CpGs. Heatmap displays beta values (rows) from unmethylated (blue) to methylated (yellow) for three sample groups (columns): 124 tumors divided into five subtypes by bootstrap clustering, 12 matched normal lung tissues, and blood leukocytes, with associated clinical characteristics and reported adenocarcinoma (AC) and SqCC gene expression phenotypes (9, 21). Left hand CpG tracks, CpG island track; black, island; gray, shore/shelf; white, open sea, H1hESC track (ref. 12; embryonic stem cell chromatin state): purple, poised promoter; red, active promoter; yellow, enhancer; green, transcribed; blue, insulator; white, heterochromatin. Sample annotations: black, yes; gray = no. B, global promoter hypermethylation (left) and global hypermethylation (right) score for methylation clusters (based on all filtered CpGs on the platform). C, box plots of DNA methylation for 629 CpGs matching repetitive elements from the set of 4,136 for each tumor in the discovery cohort across epitypes. Tumors are colored according to epitype as in A, with exception for ES5 (gray).

(35%; Fig. 3B). In contrast, never-smokers were rarely classified as ES1 in any cohort (0%–5% of all never-smokers). However, in exploratory analysis, we identified only 513 CpGs (1.1% of analyzed CpGs) to be statistically associated with clinical smoking status in adenocarcinomas across all three cohorts (false discovery rate adjusted Wilcoxon $P < 0.05$ and >0.05 difference in average β-value between groups). Notably, only 21 of these CpGs showed a more stringent difference in DNA methylation (>0.1 average β-value difference).

Consistent with the distribution of never-smokers, *EGFR* mutations were often found in ES5 tumors in the discovery and TCGA cohorts (58% and 30% of all mutations, respectively), but rarely in ES1 cases (4%–8%; Fig. 3C). Another notable difference between the epitypes was a similar enrichment of *KRAS*-mutated cases in the ES4 promoter hypermethylated epitype in both the discovery and TCGA cohorts (50%–54% of all *KRAS* mutations; Fig. 3C).

Adenocarcinoma epitypes are associated with adenocarcinoma gene expression phenotypes

In both the discovery and TCGA cohorts, the epitypes were associated with the reported bronchioid (ES5), magnoid (ES1, ES2), and squamoid (ES4) adenocarcinoma gene expression phenotypes (9) (Fig. 3D). The association of the epitypes with gene expression phenotypes was further supported by an extensive overlap between epitypes and gene expression subgroups derived from individual unsupervised consensus clustering of the discovery and TCGA cohorts (Fig. 3E). Together, these results provide a strong link between genome-wide DNA methylation and the transcriptional landscapes in lung adenocarcinoma.

Gene expression signatures associated with adenocarcinoma epitypes

The epitypes were associated with consistent differences in various gene expression metagenes in both the discovery and TCGA cohorts. For instance, ES1 had the highest expression of proliferation-associated genes (the CIN70; ref. 26, metagene), while ES5 the lowest ($P = 0.00005$ in the discovery cohort and $P = 9e{-}17$ in TCGA, Kruskal–Wallis test). The opposite pattern was found for expression of a terminal respiratory unit (TRU; ref. 27) gene signature ($P = 0.0002$ and $P = 6e{-}18$, respectively, Kruskal–Wallis test). The epitypes also differed in expression of an immune response–associated metagene and a stroma/extracellular matrix–associated metagene. Notably, the expression of these two gene modules likely relates to infiltration of immune or stromal cells in the analyzed macrodissected tissue. ES1 consistently showed the lowest and ES5 the highest expression of both metagenes (Fig. 3F, data not shown for the TCGA cohort). These results suggest that ES5 is an epitype with considerable infiltration of nonmalignant cells consistent with the observed methylation pattern being most similar to normal lung tissue. In contrast, ES1 would represent tumors with high tumor cell content. ES2 showed a different pattern for these two metagenes compared with the other epitypes (Fig. 3F). Expression of the stromal metagene was similar in ES2, ES4, and ES5, whereas expression of the immune metagene was lower in ES2 compared with ES4 and ES5, but higher compared with ES1. Supporting these observations, we found similar patterns of stromal and immune expression scores between the epitypes using the Estimation of Stromal and Immune Cells in Malignant Tumors (ESTIMATE) method (28) in both cohorts (data not shown). Together, this suggests that differences in the cellular type and amount of infiltrating nonmalignant cells may exist between the epitypes.

To further investigate biologic processes differing between the epitypes, we identified differentially expressed genes between adenocarcinomas stratified by epitype in the discovery ($n = 1{,}824$ expression probes) and TCGA ($n = 5{,}726$ genes) cohorts (Supplementary Materials and Methods). Functional analysis revealed enrichment of biologic processes involved in immune response, cell proliferation, and cell adhesion (Supplementary Table S3), consistent with results from the metagene analyses (Fig. 3).

The mutational spectrum of adenocarcinoma epitypes

To further characterize the mutational spectrum in the epitypes, we analyzed whole-exome sequencing data for TCGA adenocarcinomas. Overall, ES1 cases harbored the highest number of mutations and ES5 the least (Fig. 4A), independent of patient smoking status. Moreover, the epitypes showed differences in the type of mutation transversions when stratified by smoking status (Supplementary Fig. S4). The largest differences were observed in the distributions of C>T and C>A transversions (recognized as a smoking-related signature; ref. 29), between the ES1 (more C>A, less C>T) and ES5 epitypes (less C>A, more C>T). Consistently, overlapping ES1 cases were more often classified as transversion-high (89%) in the recent TCGA study compared with the other epitypes (55%–70%, Fisher exact $P = 0.03$; ref. 3).

To search for individual mutations associated with the epitypes, we performed a permutation-based screen of 174 genes identified by MutSigCV (30) analysis of 402 TCGA adenocarcinomas as described in ref. (31). This analysis identified seven genes with false discovery rate $\leq$10%, including four well-known tumor suppressor genes (*KEAP1*, *TP53*, *STK11*, and *SMARCA4*) and three genes appearing as either false positives (*COL11A1*, and *LRRIQ3*), or with <10% mutation frequency in any epitype (*SNRPN*). For *TP53*, *STK11*, *KEAP1*, and *SMARCA4*, we observed notable differences in the mutation frequencies between the epitypes (Fig. 4B), but no differences in mutation type (missense, truncating, or in-frame indel; χ^2 $P > 0.05$). The latter result may partly be related to the overall low number of specific mutations, for example, 86% of *SMARCA4* mutations in ES4 were missense mutations compared with 30% to 50% in ES1, ES2, or ES5. In these analyses, *KRAS* mutations were borderline nonsignificantly associated with the epitypes, whereas the association of *EGFR* mutations with the epitypes was less strong (see Fig. 3C).

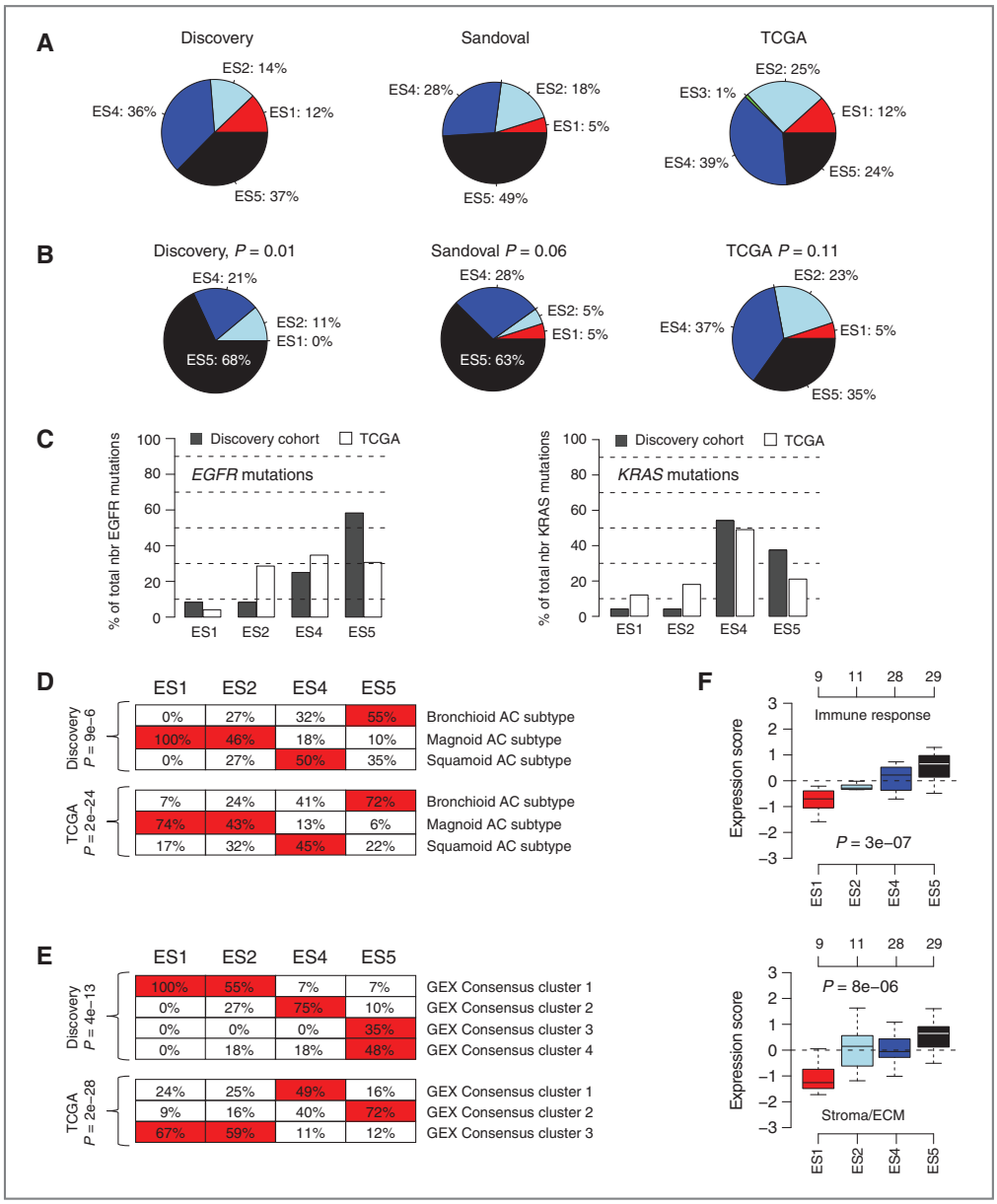


Figure 3. Clinicopathologic and gene expression characteristics of adenocarcinoma epitypes. A, distribution of adenocarcinomas in epitypes in the discovery, Sandoval, and TCGA cohorts. B, distribution of never-smokers with adenocarcinoma in epitypes in the discovery, Sandoval, and TCGA cohorts. *P* values calculated using the Fisher exact test. C, distribution of *EGFR* (left) and *KRAS* mutations (right) in the epitypes in the discovery and TCGA cohorts. D, association of reported gene expression adenocarcinoma (AC) subtypes (bronchioid, magnoid, and squamoid; ref. 9) with epitypes in the discovery (top) and TCGA (bottom) cohorts. Combinations with clear enrichment highlighted in red. *P* values were calculated using the χ^2 test. (*Continued on the following page.*)

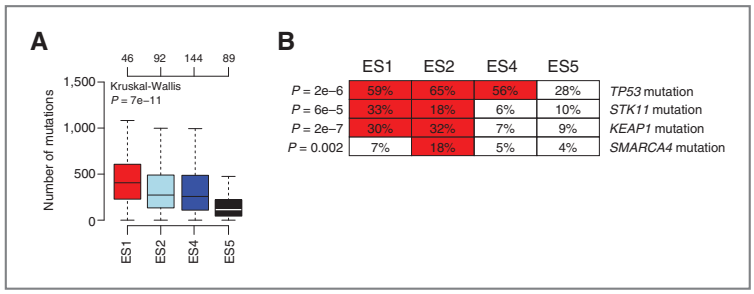


Figure 4. Association of adenocarcinoma epitypes with mutations. A, total number of mutations (silent and nonsilent) for TCGA adenocarcinomas stratified by epitype. B, association of specific mutations (only nonsilent) with epitypes in TCGA adenocarcinomas. Mutations were identified by a permutation-based approach applied to a list of significant mutations in adenocarcinoma identified through MutSigCV analysis as described in ref. (31) and Supplementary Materials and Methods. *P* values were calculated using the Fisher exact test.

Adenocarcinoma epitypes are associated with patient outcome

The four epitypes were associated with patient outcome (overall survival or relapse-free survival) in the discovery and Sandoval cohorts for NSCLC in general, and adenocarcinoma specifically (Fig. 5). Convincingly, in both cohorts, the ES2 and ES5 epitypes were associated with the best outcome in adenocarcinomas, whereas ES1 and ES4 were associated with the worst outcome. For stage I adenocarcinomas, the epitypes were associated with overall survival in the discovery cohort (relapse-free survival, log-rank $P = 0.005$), while borderline nonsignificant in the Sandoval cohort (relapse-free survival, log-rank $P = 0.06$). However, for NSCLC stage I tumors from Sandoval, the epitypes were associated with relapse-free survival (log-rank $P = 0.04$).

In univariate analysis of epitype association, patient age, smoking history, sex, *EGFR*, and *KRAS* mutation status in stage I adenocarcinomas from the discovery cohort, the epitypes were the only significant factor for overall survival ($P < 0.05$). In multivariate analysis including all these factors, the ES2 and ES5 epitypes remained significant ($P < 0.05$). In multivariate analysis of stage I adenocarcinomas from the Sandoval cohort, none of the factors (age, smoking history, gender, and epitype) reached significance.

Discussion

In the current study, we have explored the landscape of genome-wide DNA methylation across the major histologic subgroups of lung cancer, identifying five epitypes of tumors linked to different gene expression phenotypes. We demonstrate that aberrant DNA methylation in lung cancer is consistent with the classical view of hypermethylation in CpG islands, and hypomethylation in heterochromatin regions, including repetitive elements (32). Hypermethylated genes were enriched for developmental and differentiation-associated processes and polycomb targets premarked by histone H3K27 trimethylation in embryonic cells (24, 25). These results are consistent with a hypothesis that DNA methylation in lung cancer preferentially targets genes involved in morphogenetic processes and late stage differentiation of the lung epithelium, potentially contributing to establishment of an early undifferentiated cancer phenotype (24).

Through a multicohort approach, we demonstrate that LCNEC and SCLC tumors with neuroendocrine features represent a distinct lung cancer epitype compared with NSCLC, consistent with a similar association based on copy number and transcriptional alterations (33). Supporting ES3 as a distinct neuroendocrine epitype, centroid classification of 69 NSCLC cell lines (7) classified only the known LCNEC cell line, NCI-H1155, as ES3. Remaining cell lines were predominantly classified as ES1 (58%) or ES4 (36%). In both the discovery cohort and the Sandoval NSCLC validation cohort, DNA methylation epitypes identified by unsupervised bootstrap analysis comprised of a mix of adenocarcinomas and SqCCs. On the transcriptional and CNA level, adenocarcinomas and SqCCs display large differences (16, 27, 33). Here, additional studies (larger cohorts) are needed to pinpoint DNA methylation alterations that could explain such histology or cell type–specific expression patterns.

In the discovery cohort, we divided adenocarcinomas (91% stage I tumors) into four epitypes (ES1, ES2, ES4, and ES5), with marked differences in molecular and clinicopathologic characteristics, including patient outcome. Although resected stage I NSCLC patients have the most

(*Continued.*) E, association of epitypes with gene expression subgroups derived from individual unsupervised consensus clustering (50) of adenocarcinomas in the discovery cohort (top) and TCGA (bottom) cohort. For the discovery cohort, clusters were derived by consensus clustering of genes with log2ratio SD > 0.5 across all adenocarcinomas using a four-group cluster solution. The TCGA consensus clusters were derived by unsupervised consensus clustering of the most varying genes (SD of variation across tumors >1) using a three-group solution as described in ref. (31). *P* values were calculated using the χ^2 test. F, expression of an immune response associated metagene (top) and a stroma/extracellular matrix (ECM) metagene (bottom) in the discovery cohort for adenocarcinomas stratified by epitypes. *P* values were calculated using the Kruskal–Wallis test.

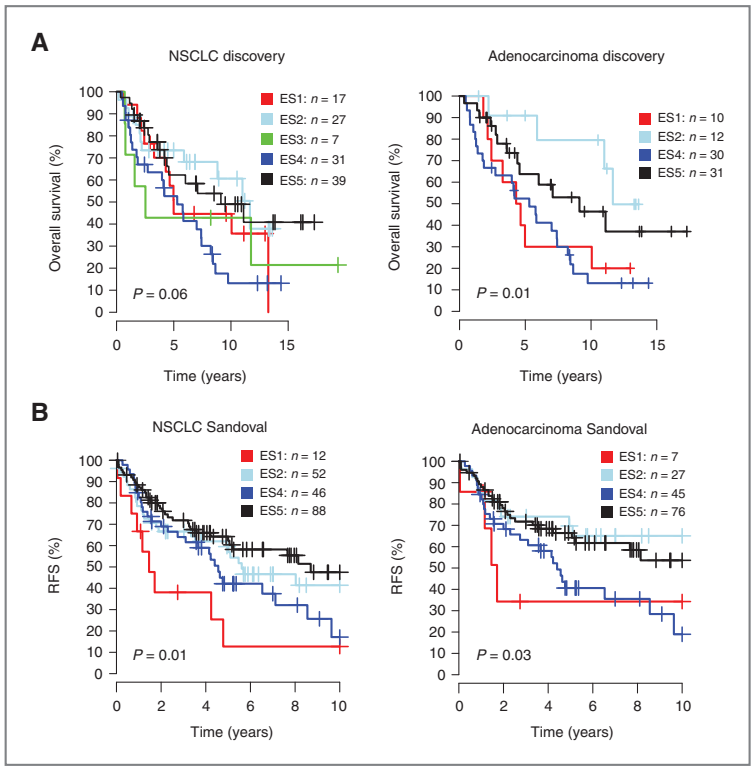


Figure 5. Association of adenocarcinoma epitypes with patient outcome. A, overall survival for patients with NSCLC (left) and patients with adenocarcinoma (right) stratified by epitype in the discovery cohort. B, relapse-free survival (RFS) for patients with NSCLC (left) and patients with adenocarcinoma (right) stratified by epitype in the Sandoval cohort. In this cohort, no patient included in survival analyses received adjuvant chemotherapy. *P* values were calculated using the log-rank test.

favorable prognosis, the 5-year survival rate is 52% to 89% (34). Thus, improved molecular subclassification of early-stage NSCLC is highly relevant. To date, only a few studies have reported DNA methylation epitypes in NSCLC or adenocarcinoma specifically (3–7). However, thorough validation of the reported epitypes has not been performed in any of these studies. In contrast, we validated our epitypes in 695 adenocarcinomas from two independent cohorts showing that they (i) provide powerful explanations of the total variation in DNA methylation compared with other clinicopathologic and molecular factors; (ii) are robust across a wide range of CpGs; (iii) have consistent clinicopathologic and molecular features in different cohorts; and (iv) could be recovered in validation cohorts by independent unsupervised analysis.

On the basis of extensive promoter hypermethylation, overrepresentation of *KRAS*-mutated adenocarcinomas, and poor patient outcome, the ES4 epitype shares features with the Shinjo and colleagues (4) adenocarcinoma CIMP-high phenotype. Supporting this association, 89% of all matching CIMP-high cases in the recent TCGA study were classified as ES4, whereas the remaining 11% were classified as ES1 (3). However, in contrast with the Shinjo and colleagues (4) CIMP-high phenotype, the ES4 epitype included never-smokers, *EGFR* mutations, and was not associated with gender (similar to the CIMP-high group in ref. 3; Fig. 3). These discrepancies may be because the CIMP definition in lung cancer is not standardized, illustrated by the differences in CIMP-high frequency between the TCGA and Shinjo and colleagues' studies (20.4% and 7.8%, respectively; refs. 3, 4). Notably, the enrichment of *KRAS*-mutated adenocarcinomas in a promoter hypermethylated cluster is consistent with previous reports (4, 6). This enrichment is intriguing given that *KRAS*-mutated adenocarcinomas have been reported to display less distinctive mRNA and CNA patterns compared with, for example, *EGFR*-mutated adenocarcinomas (17, 35). However, *KRAS* mutations have not been found to be the driver of such a promoter hypermethylated epitype in either lung adenocarcinoma or colorectal cancer, suggesting a more complex underlying mechanism (6, 36).

The ES1 epitype was characterized by global hypomethylation distant from CpG islands, hypomethylation of CpGs in repetitive elements, high expression of proliferation-associated genes, a non-TRU–like expression pattern, association with the magnoid subtype, a high

mutational burden including *TP53*, *KEAP1*, and *STK11* mutations, strong association with smoking, and poor patient outcome in both discovery and validation cohorts (Figs. 2–5). Hypomethylation in cancer have been associated with different repetitive elements that could contribute toward genomic instability (refs. 37, 38; and references therein). Accordingly, we found that ES1 tumors displayed not only more CNAs, but also that these alterations appeared more complex compared with the other epitypes based on the complex arm-wise aberration index (CAAI; ref. 18; Supplementary Fig. S5A–S5C). We also found that copy number breakpoints occurring in repetitive elements for copy number gain or loss regions were hypomethylated to a greater extent in ES1 tumors (Supplementary Fig. S5D and S5E). Together, these clinicopathologic and molecular characteristics suggest that tumor progression in ES1 may be primarily driven by genomic instability and less by classical oncogene activation (exemplified by a lower *EGFR* and *KRAS* mutation frequency). The latter is supported by the fact that ES1 cases were less often denoted oncogene-positive compared with tumors from other epitypes based on data from the recent TCGA study ($P = 0.02$, Fisher exact test; ref. 3). Moreover, from the same study, ES1 cases showed higher tumor purity and tumor ploidy compared with the other epitypes (Kruskal–Wallis $P < 0.0001$). Thus, ES1 appears to represent a poorly differentiated, aneuploid, and aggressive subset of adenocarcinomas with high tumor cellularity, less driven by oncogene activation. A smaller fraction of ES1 cases showed concomitant global hypomethylation and promoter methylation (more evident in the larger validation cohorts, Supplementary Figs. S2E and S3G). This subset of cases may better resemble the Shinjo and colleagues (4) CIMP-high group, as they were all smokers and did not harbor any *EGFR* mutations (data not shown). Irrespectively, our results support that a CIMP phenotype can occur in adenocarcinomas with markedly different epigenetic, transcriptional, and genetic make-up.

DNA methylation patterns may act as a fingerprint for different cell types (39). Compared with the DNA methylation pattern of ES1, the ES2 epitype appears more infiltrated by nonmalignant cells. Consistently, we observed differences in gene expression of metagenes associated with immune response and stroma/extracellular matrix between ES1 and ES2. Intriguingly, despite indicators of poor prognosis, including frequent CNAs, higher expression of proliferation-related genes, association with the magnoid expression subtype, and a high mutational burden (including *TP53*, *STK11*, and *KEAP1* mutations), ES2 adenocarcinoma cases (together with ES5 cases) showed the best outcome. While the generally better prognosis of ES5 cases may be attributable to their lower proliferation rate, the better prognosis of ES2 patients compared with ES1 could, hypothetically, be related to an altered and/or reduced immune cell infiltration in ES1, which have been shown to confer a poorer prognosis in multiple cancer types (19, 40–42). Whether the ES2 epitype represents an intermediate/transition state to ES1 for adenocarcinomas remains to be investigated. Although somatic alterations in specific epigenetic regulators were recently found in a notable proportion of adenocarcinomas, there were no associations with global DNA methylation patterns (3). Here, the association of *SMARCA4* (a nucleosome remodeler) mutations with ES2 is intriguing and warrants further investigation.

In contrast with the other epitypes, ES5 showed a DNA methylation pattern with similarities to blood leukocytes and normal lung tissue. Together, with its more TRU-like expression pattern, lower expression of proliferation-related genes, higher expression of immune and stroma-related metagenes, high frequency of bronchioid classified tumors, enrichment of never-smokers, and better patient outcome, ES5 matches a proposed TRU type of adenocarcinoma (43) but also shares characteristics with the CIMP-negative epitype reported by Shinjo and colleagues (4). Importantly, ES5 cannot be dismissed as an epitype merely due to sampling issues, as the analyzed tumor DNA carried both CNAs and mutations. For instance, for the 25 cancer hallmark genes defined by Imielinski and colleagues (44), 78% of ES5 cases in the TCGA cohort carried at least one alteration (mutation or CNA). Moreover, the lack of NSCLC cell lines classified as ES5 or ES2 (see above) does not dismiss these epitypes in clinical tumor specimens, as for instance the well established intrinsic molecular subtypes in breast cancer are not reproduced exactly in breast cancer cell lines (45).

DNA methylation of smaller sets of CpGs/genes (e.g. *CDKN2A*, *FHIT*, *APC*, and *RASSF1A*) have been associated with smoking in both genome-wide and gene-focused studies of lung cancer (6, 38, 46) (and references therein). However, on a genome-wide level our results suggest an overall less dominant effect of clinical smoking history on the DNA methylation landscape in primary adenocarcinomas, despite the enrichment of never-smokers in specific epitypes. This conclusion may be exemplified by results from the principal component analyses, the presence of never-smokers in all epitypes with exception of ES1, the low number of smoking associated CpGs in both the current and previous studies (6, 46), and that the patterns of global and promoter hypermethylation between epitypes were similar irrespective of smoking status (Figs. 2A, S1B, S2A, S3A and S6A and B). Combined with similar findings of intrinsically heterogeneous gene expression and CNA patterns in smoking-defined adenocarcinoma subgroups (20, 31), our results question whether never-smokers can be identified as a molecular subgroup of its own with transcriptional, DNA methylation, and CNA patterns clearly different from tumors arising in smokers. Instead, our study further supports that a majority of adenocarcinomas arising in never-smokers together with a specific subset of tumors from smokers represent a more distinct and relevant molecular/biologic entity of less aggressive and potentially more smoking-unrelated disease (20, 31). The clinical smoking definitions are intrinsically problematic due to their self-reported nature, but also because they do not capture the intensity and duration of cigarette exposure,

and the exposure to environmental tobacco smoke and other pollutants for never-smokers. Interestingly, the few TCGA never-smokers classified as ES1 display smoking characteristic C>A transversion frequencies similar to current-smokers, clearly different from, for example, ES2-classified never-smokers (Supplementary Fig. S4A). Thus, whether these never-smokers are "true" never-smokers remains unclear. This suggests that ES1 is in fact strongly related to patients with a smoking history and, importantly, presumably also distinct underlying tumor biology and/or tumorigenic events.

The question of whether the observed DNA methylation epitypes/alterations are driver or passenger events, and their position and role in the evolutionary tree of a tumor remains to be determined. Promoter hypermethylation of individual genes, notably tumor suppressors like *CDKN2A*, have been recognized as early events in lung tumorigenesis, while there is a lack of consensus over whether global hypomethylation is an early or late event in lung cancer (see refs. 37, 38). The impact of smoking on epigenetic modifications may further complicate the picture, as certain alterations have been associated with duration or amount of tobacco smoking and may thus be later events in the cancer development and progression (38). Whole-genome bisulfite sequencing combined with other profiling/sequencing techniques may be one potential way of reconstructing the evolution of a tumor in relation to driver mutations, CNAs, and DNA methylation, as recently described for DNA alterations in breast cancer (47).

Besides describing DNA methylation patterns in lung adenocarcinoma, our study strongly supports a link between adenocarcinoma gene expression phenotypes and genome-wide DNA methylation patterns (9). Importantly, this link brings further insights and explanation to the observed clinicopathologic characteristics, gene expression patterns, mutational signatures, and biologic pathways/processes associated with the epitypes (3, 9, 27, 43). However, the current study also extends the knowledge about genome-wide DNA methylation patterns in the adenocarcinoma gene expression phenotypes, for example, showing that the current definition of these phenotypes comprises of a mix of DNA methylation patterns (Fig. 3D). In contrast with Wilkerson and colleagues (9), we found that the magnoid subtype was strongly associated with a global DNA hypomethylation pattern in both the discovery and TCGA cohorts (Supplementary Fig. S6C and S6D). Furthermore, DNA methylation patterns in and between the epitypes were consistent irrespective of bronchioid, magnoid, or squamoid classification (Supplementary Fig. S6E and S6F). Together, our results suggest that further refinement of both the proposed gene expression phenotypes and the CIMP phenotype in lung adenocarcinoma should be possible through integrated analysis of transcriptional, copy number, and DNA methylation data.

Epigenetic alterations, including DNA methylation, are potentially reversible which offers an interesting therapeutic opportunity. For instance, DNA methyltransferase (DNMT) inhibitors can induce DNA hypomethylation at specific gene loci that can result in sustained gene reactivation (48). Currently, DNMT inhibitors and multiple histone deacetylase (HDAC) inhibitors are in clinical use and/or clinical testing in different malignancies, and a recent phase I/II trial reported an objective response to a combinatorial treatment with DNMT and HDAC inhibitors in recurrent metastatic NSCLC (49). Interestingly, in a recent NSCLC cell line experiment, cell lines with a CIMP-positive phenotype responded with growth inhibition to 5-Aza-dC (a DNMT inhibitor) treatment, while CIMP-negative cell lines did not (4). Whether the proposed epitypes in the current study define patient subgroups likely to benefit or not from such treatments remains to be investigated.

In summary, based on a multicohort approach, we have conducted a comprehensive survey of the genome-wide DNA methylation pattern in lung cancer involving the major histologic subgroups. Together, the current study adds further layers of information about the epigenetic characteristics and molecular diversity in lung cancer. Moreover, it highlights the possibility to further refine disease classification that may ultimately lead to improvements in detection, patient stratification, prognostication, and therapy.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: A. Karlsson, A. Borg, M. Planck, J. Staaf
Development of methodology: A. Karlsson, M. Jönsson, M. Lauss, M. Ringner, J. Staaf
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): H. Brunnström, P. Jönsson, M. Planck
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): A. Karlsson, M. Lauss, G. Jönsson, M. Ringner, J. Staaf
Writing, review, and/or revision of the manuscript: A. Karlsson, H. Brunnström, A. Borg, M. Ringner, M. Planck, J. Staaf
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): M. Jönsson, M. Planck
Study supervision: A. Borg, J. Staaf

Acknowledgments

The authors thank the editors at Elevate Scientific for helpful comments on the article.

Grant Support

Financial support for this study was provided by the Swedish Cancer Society, the Knut and Alice Wallenberg Foundation, the Foundation for Strategic Research through the Lund Centre for Translational Cancer Research (CREATE Health), the Mrs Berta Kamprad Foundation, the Gunnar Nilsson Cancer Foundation, the Swedish Research Council, the Lund University Hospital Research Funds, the Gustav V's Jubilee Foundation, and the IngaBritt and Arne Lundberg Foundation. The SCIBLU Genomics center is supported by governmental funding of clinical research within the national health services (ALF) and by Lund University.

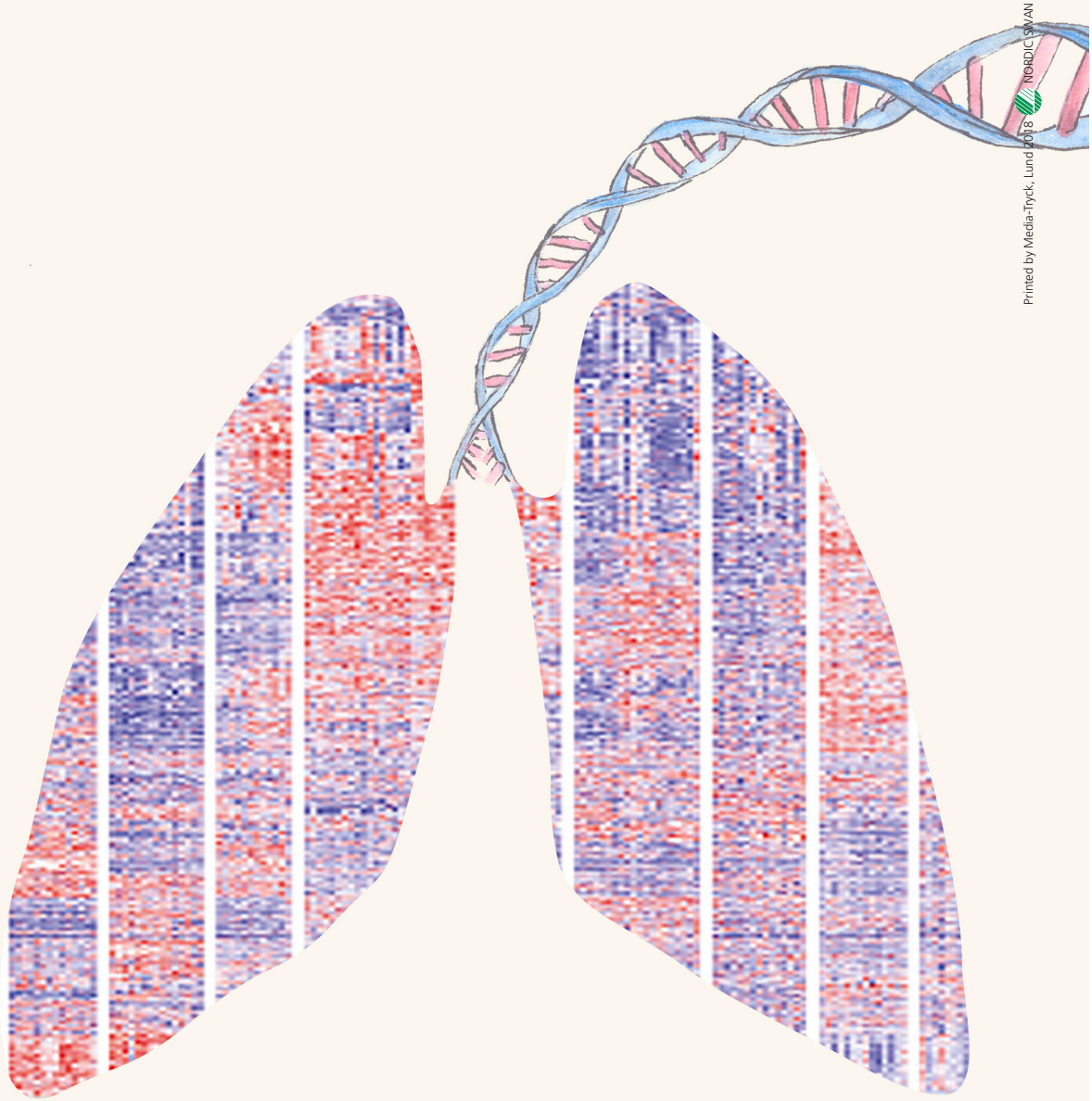
The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received April 29, 2014; revised September 18, 2014; accepted September 26, 2014; published OnlineFirst October 2, 2014.

References

1. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin 2011;61:69–90.
2. Jones PA, Baylin SB. The epigenomics of cancer. Cell 2007;128: 683–92.
3. The Cancer Genome Atlas Network A. Comprehensive molecular profiling of lung adenocarcinoma. Nature 2014;511:543–50.
4. Shinjo K, Okamoto Y, An B, Yokoyama T, Takeuchi I, Fujii M, et al. Integrated analysis of genetic and epigenetic alterations reveals CpG island methylator phenotype associated with distinct clinical characters of lung adenocarcinoma. Carcinogenesis 2012;33: 1277–85.
5. Sandoval J, Mendez-Gonzalez J, Nadal E, Chen G, Carmona FJ, Sayols S, et al. A prognostic DNA methylation signature for stage I non-small-cell lung cancer. J Clin Oncol 2013;31:4140–7.
6. Selamat SA, Chung BS, Girard L, Zhang W, Zhang Y, Campan M, et al. Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression. Genome Re 2012;22: 1197–211.
7. Walter K, Holcomb T, Januario T, Du P, Evangelista M, Kartha N, et al. DNA methylation profiling defines clinically relevant biological subsets of non-small cell lung cancer. Clin Cancer Res 2012;18: 2360–73.
8. Hughes LA, Melotte V, de Schrijver J, de Maat M, Smit VT, Bovee JV, et al. The CpG island methylator phenotype: what's in a name? Cancer Res 2013;73:5858–68.
9. Wilkerson MD, Yin X, Walter V, Zhao N, Cabanski CR, Hayward MC, et al. Differential pathogenesis of lung adenocarcinoma subtypes involving sequence mutations, copy number, chromosomal instability, and methylation. PLoS ONE 2012;7:e36530.
10. Hammerman PS, Lawrence MS, Voet D, Jing R, Cibulskis K, Sivachenko A, et al. Comprehensive genomic characterization of squamous cell lung cancers. Nature 2012;489:519–25.
11. The Cancer Genome Atlas. Available from: http://cancergenome.nih.gov/.
12. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 2011;473:43–9.
13. Lauss M, Aine M, Sjodahl G, Veerla S, Patschan O, Gudjonsson S, et al. DNA methylation analyses of urothelial carcinoma reveal distinct epigenetic subtypes and an association between gene copy number and methylation status. Epigenetics 2012;7:858–67.
14. Lauss M, Visne I, Kriegner A, Ringner M, Jonsson G, Hoglund M. Monitoring of technical variation in quantitative high-throughput datasets. Cancer Inform 2013;12:193–201.
15. Gene Expression Omnibus. Available from: http://www.ncbi.nlm.nih.gov/geo/.
16. Staaf J, Isaksson S, Karlsson A, Jonsson M, Johansson L, Jonsson P, et al. Landscape of somatic allelic imbalances and copy number alterations in human lung carcinoma. Int J Cancer 2012;1:2020–31.
17. Planck M, Edlund K, Botling J, Micke P, Isaksson S, Staaf J. Genomic and Transcriptional Alterations in Lung Adenocarcinoma in Relation to EGFR and KRAS Mutation Status. PLoS ONE 2013;8: e78614.
18. Russnes HG, Vollan HK, Lingjaerde OC, Krasnitz A, Lundin P, Naume B, et al. Genomic architecture characterizes tumor progression paths and fate in breast cancer patients. Sci Transl Med 2010;2:38ra47.
19. Fredlund E, Staaf J, Rantala JK, Kallioniemi O, Borg A, Ringner M. The gene expression landscape of breast cancer is shaped by tumor protein p53 status and epithelial-mesenchymal transition. Breast Cancer Res 2012;14:R113.
20. Staaf J, Jonsson G, Jonsson M, Karlsson A, Isaksson S, Salomonsson A, et al. Relation between smoking history and gene expression profiles in lung adenocarcinomas. BMC Med Genomics 2012;5:22.
21. Wilkerson MD, Yin X, Hoadley KA, Liu Y, Hayward MC, Cabanski CR, et al. Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types. Clin Cancer Res 2010;16:4864–75.
22. da Huang W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 2009;4:44–57.
23. Berman BP, Weisenberger DJ, Aman JF, Hinoue T, Ramjan Z, Liu Y, et al. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. Nat Genet 2012;44:40–6.
24. Helman E, Naxerova K, Kohane IS. DNA hypermethylation in lung cancer is targeted at differentiation-associated genes. Oncogene 2012;31:1181–8.
25. Easwaran H, Johnstone SE, VanNeste L, Ohm J, Mosbruger T, Wang Q, et al. A DNA hypermethylation module for the stem/progenitor cell signature of cancer. Genome Res 2012;22:837–49.
26. Carter SL, Eklund AC, Kohane IS, Harris LN, Szallasi Z. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. Nat Genet 2006; 38:1043–8.
27. Takeuchi T, Tomida S, Yatabe Y, Kosaka T, Osada H, Yanagisawa K, et al. Expression profile-defined classification of lung adenocarcinoma shows close relationship with underlying major genetic changes and clinicopathologic behaviors. J Clin Oncol 2006;24: 1679–88.
28. Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun 2013;4:2612.
29. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature 2013;500:415–21.
30. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013;499:214–8.
31. Karlsson A, Ringner M, Lauss M, Botling J, Micke P, Planck M, et al. Genomic and transcriptional alterations in lung adenocarcinoma in relation to smoking history. Clin Cancer Res 2014;20: 4912–24.
32. Ehrlich M. DNA methylation in cancer: too much, but also too little. Oncogene 2002;21:5400–13.
33. CLCGP. A genomics-based classification of human lung tumors. Sci Transl Med 2013;5:209ra153.
34. Crino L, Weder W, van Meerbeeck J, Felip E. Early stage and locally advanced (non-metastatic) non-small-cell lung cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2010;21 Suppl 5:v103–15.
35. Chitale D, Gong Y, Taylor BS, Broderick S, Brennan C, Somwar R, et al. An integrated genomic analysis of lung cancer reveals loss of DUSP4 in EGFR-mutant tumors. Oncogene 2009;28:2773–83.
36. Hinoue T, Weisenberger DJ, Lange CP, Shen H, Byun HM, Van Den Berg D, et al. Genome-scale analysis of aberrant DNA methylation in colorectal cancer. Genome Res 2012;22:271–82.
37. Chen C, Yin N, Yin B, Lu Q. DNA methylation in thoracic neoplasms. Cancer Lett 2011;301:7–16.
38. Langevin SM, Kratzke RA, Kelsey KT. Epigenetics of lung cancer. Transl Res 2014. S1931-5244(14)00085-1 [pii] 10.1016/j.trsl.2014.03.001 PUBMED: 24686037.
39. Fernandez AF, Assenov Y, Martin-Subero JI, Balint B, Siebert R, Taniguchi H, et al. A DNA methylation fingerprint of 1628 human samples. Genome Res 2012;22:407–19.
40. Jonsson G, Busch C, Knappskog S, Geisler J, Miletic H, Ringner M, et al. Gene expression profiling-based identification of molecular subtypes in stage IV melanomas with different clinical outcome. Clin Cancer Res 2010;16:3356–67.
41. Pages F, Berger A, Camus M, Sanchez-Cabo F, Costes A, Molidor R, et al. Effector memory T cells, early metastasis, and survival in colorectal cancer. N Engl J Med 2005;353:2654–66.
42. Zhang L, Conejo-Garcia JR, Katsaros D, Gimotty PA, Massobrio M, Regnani G, et al. Intratumoral T cells, recurrence, and survival in epithelial ovarian cancer. N Engl J Med 2003;348:203–13.

43. Yatabe Y. EGFR mutations and the terminal respiratory unit. Cancer Metastasis Rev 2010;29:23–36.
44. Imielinski M, Berger AH, Hammerman PS, Hernandez B, Pugh TJ, Hodis E, et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. Cell 2012;150:1107–20.
45. Prat A, Parker JS, Karginova O, Fan C, Livasy C, Herschkowitz JI, et al. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. Breast Cancer Res 2010;12: R68.
46. Tan Q, Wang G, Huang J, Ding Z, Luo Q, Mok T, et al. Epigenomic analysis of lung adenocarcinoma reveals novel DNA methylation patterns associated with smoking. Onco Targets Ther 2013;6: 1471–9.
47. Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, Greenman CD, Lau KW, et al. The life history of 21 breast cancers. Cell 2012;149:994–1007.
48. Raynal NJ, Si J, Taby RF, Gharibyan V, Ahmed S, Jelinek J, et al. DNA methylation does not stably lock gene expression but instead serves as a molecular mark for gene silencing memory. Cancer Res 2012;72:1170–81.
49. Juergens RA, Wrangle J, Vendetti FP, Murphy SC, Zhao M, Coleman B, et al. Combination epigenetic therapy has efficacy in patients with refractory advanced non-small cell lung cancer. Cancer Discov 2013;1:598–607.
50. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics 2010;26:1572–3.



Printed by Media-Tryck, Lund 2018 NORDIC SWAN ECOLABEL 3041 0903

LUND UNIVERSITY

FACULTY OF MEDICINE

Division of Oncology and Pathology
Department of Clinical Sciences, Lund

Lund University, Faculty of Medicine
Doctoral Dissertation Series 2018:122
ISBN 978-91-7619-690-8
ISSN 1652-8220