Imperial College London
National Heart and Lung Institute

Multi-omics molecular profiling of lung tumours

Clara Domingo-Sabugo

Thesis submitted to Imperial College London for the degree of
Doctor of Philosophy

December 21, 2021

Copyright Declaration

Abstract

Lung Cancer (LC) is one of the most common malignancies and is the leading cause of cancer death worldwide among both men and women. Current LC classifications are based on histopathological features which poorly reflect the molecular diversity of these tumours. Consequently, primary and secondary drug resistance are very frequent, and a high mortality is usual in LC patients. Despite the fact that LC has been intensively studied, there is a lack of effective biomarkers for early detection, stratification and prognosis. Integration of omics data is a powerful approach that can be used to identify molecular subgroups relevant in the clinical setting. This thesis addresses this challenge by characterising the molecular alterations accompanying LC at the genetic and DNA methylation level, using a combination of Whole-Exome Sequencing (WES), Targeted Capture Sequencing (TCS), Single Nucleotide Polymorphism (SNP) genotyping, Whole-Genome Bisulfite Sequencing and RNA-sequencing. The integration of different types of omics data first validated previous molecular alterations in frequently diagnosed LC tumours. This allowed comparison of the genomic and epigenomic landscapes between these common and rarer LC subtypes. Next, novel molecular subgroups of Non-Small Cell Lung Cancer (NSCLC) tumours with bad prognostic, as well as subgroups of Lung Carcinoids (L-CDs, an understudied LC subtype) have been identified and their molecular alterations and signatures characterised. Significant associations with histological features and gene expression programmes have been found by using several bioinformatic tools. These results show the value of multi-omics approaches to better understand the molecular mechanisms underlying LC and to identify new biomarkers. Importantly, some of these findings may be translatable and are likely to improve the detection, monitoring and stratification for targeted therapies in LC patients.

Acknowledgements

There are many people I wish to thank who have made this PhD both possible and enjoyable.

Firstly, I would like to thank my supervisors Professor Miriam Moffatt and Professor William Cookson for giving me the opportunity to be part of their team and their support, dedication, and guidance throughout my studentship.

I also wish to thank the members of the Asmarley Centre for Genomic Medicine lab, past and present, for all their help and support, especially to Amit Mandal and Saffron A G Willis-Owen who helped making my work possible. Thanks to Anca Nastase, Long Hoang, Leah F Cuthbertson, Michael T Olanipekun, Colin P Churchward, Lauren A Headley-Morris and Pamelbir S Ladhar. It has been an honour to be able to work with such talented scientists and wonderful people.

I also wish to thank Professor Mark Lathrop and Markus Munter for giving me the opportunity to visit them at the McGill Genome Centre in Canada and enabling me to interact with an amazingly talented group of scientists from the Canadian Centre for Computational Genomics.

I also wish to thank all the enthusiastic members of the NHLI Postgraduate Committee. You have all helped NHLI students to keep socially and physically active during the lockdown due to the Covid-19 pandemic. A special thanks goes to Julia Frankenberg, for being the best peer support and friend at the NHLI, and to Helena Lund Palau and Michelle Chiappi, for your wise advice, help and friendship.

My deepest thanks go to my father for all the words of encouragement and uncountable actions to give me always the best possible education, and to my sister for her unconditional love and support throughout the whole journey. I wouldn't be able to get through it without you.

Dedication

To my mother, thank you for what you did for my education and for making me the person I am today, I wish you were still here with us.

Declaration of Originality

I, Clara Domingo Sabugo, declare that all the work contained within this thesis is the original work of the author and that any work of another person has been appropriately credited and referenced. This work has not been submitted for any other degree or professional qualification. Contributions from other individuals towards the work contained in this thesis include:

Dr. Elizabeth Starren (NHLI) contributed to the project with nucleic acid extraction of samples for genomics and transcriptomics analyses. Prof. Mark Lathrop, Dr. Markus Munter and the Canadian Centre for Computational Genomics (C3G) team (McGill Genome Centre) generated and sequenced WES and WGBS libraries. Dr. R. Eveleigh (McGill Genome Centre) performed somatic mutation calling from WES sequencing data. Dr. S. Dwyer (Genomic Medicine Group) prepared RNA sequencing libraries.

The work related to articles 1 and 2 bellow is explained in Chapter 5 and Chapter 6, respectively. Article 1 was published in Nature Scientific Reports[1]. Manuscript authors had input on the analysis, paper design and wording. Article 2 is currently under peer-review process, and authors had input on the paper design and wording.

1. Saffron A. G. Willis-Owen, Clara Domingo-Sabugo, Elizabeth Starren, Liming Liang, Maxim B. Freidin, Madeleine Arseneault, Youming Zhang, Shir Kiong Lu, Sanjay Popat, Eric Lim, Andrew G. Nicholson, Yasser Riazalhosseini, Mark Lathrop, William O. C. Cookson, Miriam F. Moffatt. Y disruption, autosomal hypomethylation and poor male lung cancer survival. *Sci. Rep.* **11**, 12453 (2021).
2. Clara Domingo-Sabugo, Saffron A.G. Willis-Owen, Amit Mandal, Anca Nastase, Sarah Dwyer, Cecilia Brambilla, José Héctor Gálvez, Qinwei Zhuang, Sanjay Popat, Robert Eveleigh, Markus Munter, Eric Lim, Andrew G. Nicholson, Mark Lathrop, William O.C. Cookson, Miriam F. Moffatt. Distinct pancreatic and neuronal Lung Carcinoid molecular subtypes revealed by integrative omic analysis. *medRxiv* (2021).

The work related to the below research article published in Molecular Oncology[2] encompasses part of the results presented in Chapter 3.

3. Long T Hoang, Clara Domingo-Sabugo, Elizabeth S Starren, Saffron A G Willis-Owen, Deborah J Morris-Rosendahl, Andrew G Nicholson, William O C M Cookson, Miriam F Moffatt. Metabolomic, transcriptomic and genetic integrative analysis reveals important roles of adenosine diphosphate in haemostasis and platelet activation in non-small-cell lung cancer. *Mol. Oncol.* **13**, 2406–2421 (2019).

Contents

List of Figures

List of Tables

List of Abbreviations

3′UTR	Three prime Untranslated Region
5′UTR	Five prime Untranslated Region
A	Adenine
AC	Atypical Carcinoid
AD	Average distance
ADM	Average distance between means
AF	Allele Frequency
AFB1	Aflatoxin B1
AID	Activation-Induced cytidine Deaminase
ALT	Alternative Lengthening of Telomeres
AP-1	Activating Protein 1
APC	Adenomatous Polyposis Coli
ARRB1	B-assrestin-1
AS	Alternative Splicing
ASCAT	Allele-Specific Copy number Analysis of Tumours
ATR	Ataxia telangiectasia and Rad3-related protein
ATRi	ATR inhibitors
BAF	B allele frequency
BAM	Binary form of SAM
BCL	Binary Base Call
BED	Browser Extensible Data
BER	Base Excision Repair
BMP	Bitmap Image File
Bp	Base Pair
BP	Biological Process
BQSR	Base Quality Score Recalibration
BRU	Biomedical Research Unit
C	Cytosine
C3G	Centre for Computational Genomics
CADD	Combined Annotation Dependent Depletion
CBP	CREB Binding Protein
CBS	Circular Binary Segmentation
CC	Cellular Component
CGIs	CpG Islands
CIN	Chromosome Instability
CLL	Chronic Lymphocytic Leukaemia
CMS	COSMIC mutational signature
CN	Copy Number

CNA	Copy Number Alteration
CNB	Copy Number Burden
CNV	Copy Number Variant
COSMIC	Catalogue of Somatic Mutations in Cancer
CpG	5'-C-phosphate-G-3'
CPM	Counts Per Million
CSN	Clinical Sequencing Nomenclature
CT	Computed Tomography
CTCs	Circulating Tumour Cells
CXRs	Chest Radiographies
dbscSNV	Splice-site consensus Single-Nucleotide Variants database
dbSNP	Single Nucleotide Polymorphism Database
DDR	DNA damage response
DISC	Death-Inducing Signalling Complex
DM	Differential Methylation
DMC	Differentially Methylated Cytosine
DMRs	Differentially Methylated Regions
***dn*CMS**	*De novo* mutational signatures
DNMT1	DNA Methyltransferase 1
DNMTs	DNA Methyltransferases
DSB	Double-Strand Break
Dsh	Dishevelled
DSS-single	Dispersion Shrinkage for Sequencing data with single replicates
eDMR	Differentially Methylated enhancer
EGF	Epidermal Growth Factor
EGFR	Epidermal Growth Factor Receptor
EM	Expectation-Maximization algorithm
EMT	Epithelial to Mesenchymal Transition
ENCODE	Encyclopedia of DNA Elements
ERV	Endogenous Retroviru
ERV1	Endogenous Retrovirus 1 family
ES	Enrichment Score
ExAC	Exome Aggregation Consortium
FDA	Food and Drug Administration
FDR	False Discovery Rate
FEV1	Forced Expired Volume in 1 second
FGFR	Fibroblast Growth Factor Receptor
FOM	Figure of merit
Fz	Frizzled
FZD	Frizzled Class Receptor
GATK	Genome Analysis Tool Kit

GC	GenCall
GCT	Gene Cluster Text
GEMINI	Genome MINIng
GERP	Genomic Evolutionary Rate Profiling
Gistic2	Genomic Identification of Significant Targets In Cancer
GnomAD	Genome Aggregation Database
GO	Gene Ontology
GREAT	Genomic Regions for Enrichment of Annotations Tool
GSEA	Gene Set Enrichment Analysis
GTF	General Transfer Format
GTP	Guanosine Triphosphate
H3K27me3	Trimethylation of H3K27
H3K27me3	Trimethylation of histone H3 lysine 27
HATs	Histone Acetyltransferases
HCA	Hierarchical Clustering Analysis
HDACs	Histone Deacetylases
HDMs	Histone Demethylases
HMG	High-Mobility Group
HMTs	Histone Methyltransferases
HP	Human Phenotype Ontology
HPA	Human Protein Atlas
HR	Homologous Recombination
ICGC	International Cancer Genome Consortium
ID	Identifier
IFN	Interferon
IGV	Integrated Genome Viewer
InDels	Insertions and Deletions
IR	Ionizing Radiation
IS	Immune System
KDM5C	Lysine Demethylase 5C
KDM5D	Lysine Demethylase 5D
KEGG	Kyoto Encyclopedia of Genes and Genomes
KIRC	Kidney Renal Clear Cell Carcinoma
L-CD	Lung Carcinoid
LC	Lung cancer
LCNEC	Large Cell Neuroendocrine Carcinoma
lincRNAs	Long intergenic non-coding RNAs
lncRNAs	long non-coding RNAs
LNETs	Lung Neuroendocrine Tumours
Log2 CPM	Log 2 counts per million reads
LRR	Log R ratio

LTR	Long Terminal Repeat
LUAD	Lung Adenocarcinoma
LUSC	Lung Squamous Cell Carcinoma
LYE	low expression of Y chromosome
MAF	Mutation Annotation Format
Mb	Mega base
MBD	Methyl CpG binding Domain
MF	Molecular Function
miRNAs	MmicroRNAs
ML	Machine Learning
MM	Module Membership
MMR	Mismatch Repair
MPT	Mitochondrial Permeability Transition
MRI	Magnetic Resonance Imaging
MSI	Microsatellite Instability
MSigDB	Molecular Signatures Database
ncRNAs	Noncoding RNAs
NECs	Neuroendocrine Carcinomas
NENs	Neuroendocrine Neoplasms
NER	Nucleotide Excision Repair
NES	Normalised Enrichment Score
Nf-κB	Nuclear Factor Kappa-light-chain-enhancer of activated B cells
NGS	Next-Generation Sequencing
NMF	Non-negative Matrix Factorization
NO2	Nitrogen Dioxide
OCGs	Oncogenes
OR	Olfactory Receptor
PanNETs	Pancreatic Neuroendocrine Tumours
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PCs	Principal Components
PDZD2	PDZ domain-containing 2
PET	Positron Emission Tomography
PI3K	Phosphoinositide 3-Kinase
PIP2	Phosphatidylinositol 4,5-bisphosphate
piRNAs	PIWI-interacting RNAs
PKB	Protein Kinase B
PPI	Protein-Protein Interaction
PRC	Polycomb repressive complex
PTEN	Phosphatase and Tensin homolog
PTMs	Post-Translational Modifications

QC	Quality Control
RASSF1	RAS-association domain family 1
RE	Response Element
rfDMR	Repeat-free Differentially Methylatd Region
RNAi	RNA interference
ROS	Reactive Oxygen Species
rrDMR	Repeat-rich Differentially Methylated Region
SAM	Sequence Alignment Mapping
SBS	Sequencing by Synthesis
SCLC	Small Cell Lung Carcinoma
scSNVs	Single Nucleotide Variants within splicing consensus regions
SD	Standard Deviations
SIFT	Sorting Intolerant From Tolerant
SINE	Long Interspersed Nuclear Element
SINE	Short Interspersed Nuclear Element
siRNAs	Small interfering RNAs
SLC	Solute-Carrier
snoRNAs	Small nucleolar RNAs
SNPs	Single-Nucleotide Polymorphisms
SNV	Single-Nucleotide Variant
SSBs	Single-strand DNA breaks
SSE	Sum-Squared Error
STAR	Spliced Transcripts Alignment to a Reference
T	Thymine
TC	Typical Carcinoid
TCGA	The Cancer Genome Atlas
TCS	Targeted Capture Sequencing
TERT	Telomerase Reverse Transcriptase
TEs	Transposable Elements
TET	Ten-Eleven Translocation
TF	Transcription Factor
TGF-B	Transforming Growth Factor Beta
TKIs	Tyrosine Kinase Inhibitors
TMB	Tumour Mutational Burden
TNF	Tumour Necrosis Factor
TSG	Tumour Suppressor Gene
U	Uracil
UCSC	University of California Santa Cruz
UV	Ultraviolet
VAF	Variant Allele Frequency
VCF	Variant Call Format

VEP	Variant Effect Predictor
WES	Whole Exome Sequencing
WGCNA	Weighted Gene Co-expression Network Analysis
WGS	Whole Genome Sequencing
WHO	World Health organization
WP	Wiki Pathway
ZNFs	Zinc-Finger proteins

Chapter 1: General Introduction

1.1 Lung Cancer

1.1.1 Introduction: Lung Cancer

Lung cancer (LC) is one of the most common malignancies and the leading cause of cancer death worldwide with an estimated 1.8 million deaths[3]. Although people who have never smoked can develop lung cancer, smoking is the major risk factor of this disease accounting for over 85% of cases. In 2015, the World Health Organization (WHO) published a new histological classification of tumours of the lung[4,5] maintaining three major histological types[6]: Lung Adenocarcinoma (LUAD), Lung Squamous Cell Carcinoma (LUSC) and Lung Neuroendocrine Tumours (LNETs). LNETs comprise Small Cell Lung Carcinoma (SCLC), Large Cell Neuroendocrine Carcinoma (LCNEC), and Atypical and Typical Lung Carcinoid (L-CD).

1.1.2 Lung Cancer Causes and Risk Factors

Smoking is by far the main aetiological factor for developing LC and directly accounts for 82% of the cases[7]. Variation of LC rates and trends largely reflects the maturity of the tobacco epidemic[8], and a stable decrease has been observed in men. Nevertheless, this decline in LC incidence is two times slower in woman compared to men, and the majority of countries are continuing to observe a rising incidence of LC[9] among women. For instance, Denmark, Iceland, and Sweden show even higher incidence rates in females (ages 35-64 years) than in males. Smoking behaviours reflecting historical differences in tobacco uptake however do not seem to fully explain the higher incidence observed in women born since 1960[10].

Increased LC risk has also been reported after consumption of cigar, bidi and hookah in India, khii yoo in Thailand, water pipe in China and other tobacco products[11]. Roll-your-own cigarettes and de-nicotinized cigarettes have been shown to still be toxic despite manipulation of tobacco composition. People who have never smoked or consumed tobacco in other forms can, however, also develop LC. Factors such as genetic susceptibility, poor diet, alcohol consumption, occupational exposures and air pollution can act independently or in

concert and may predispose to LC. In particular, the identification of molecular alterations has started to bring light into the aetiology of LC in never smokers with several Single-Nucleotide Polymorphisms (SNPs) and mutations being linked to LC development[12,13].

Further, the exposure to second-hand tobacco smoke – this is a stream of smoke released from the burning cone mixed with the exhaled mainstream smoke in combination with the air in an indoor environment to which both smokers and non-smokers are exposed - is also well known to be involved in LC. This complex mixture leads to voluntary (passive) smoking. Low ventilation together with second-hand smoke can result in concentrations of toxic and carcinogenic agents above those found in urban areas[14].

Outdoor ambient air pollution, also known as fine particulate matter, is also a major risk factor with 3.4 million deaths being attributed to it in 2017[15]. Other inhalable agents, such as household burning of solid fuels for heating and cooking, have been suggested as further risk factors[16,17], as well as volatile organic compounds and Nitrogen Dioxide (NO_2) released from cooking, cleaning and other indoor air contaminants (candles, incense, shower gels and fragrances, glues, inks and air fresheners).

Additional risk factors include asbestos, radon and other ionizing radiation, as well as arsenic, nickel, silica and chloromethyl ethers. Infectious agents are also emerging as players in the development of LC[18] with lung microbiota dysbiosis being shown to correlate with LC[19]. Exposure to microbial oncogenes, toxins and/or Reactive Oxygen Species (ROS) from microbial activities can lead to mutations and dysregulation of important biological processes, such as cell cycle, proliferation or apoptosis, that can also contribute to carcinogenesis. A summary of several risk factors is shown in Fig. 1.1.

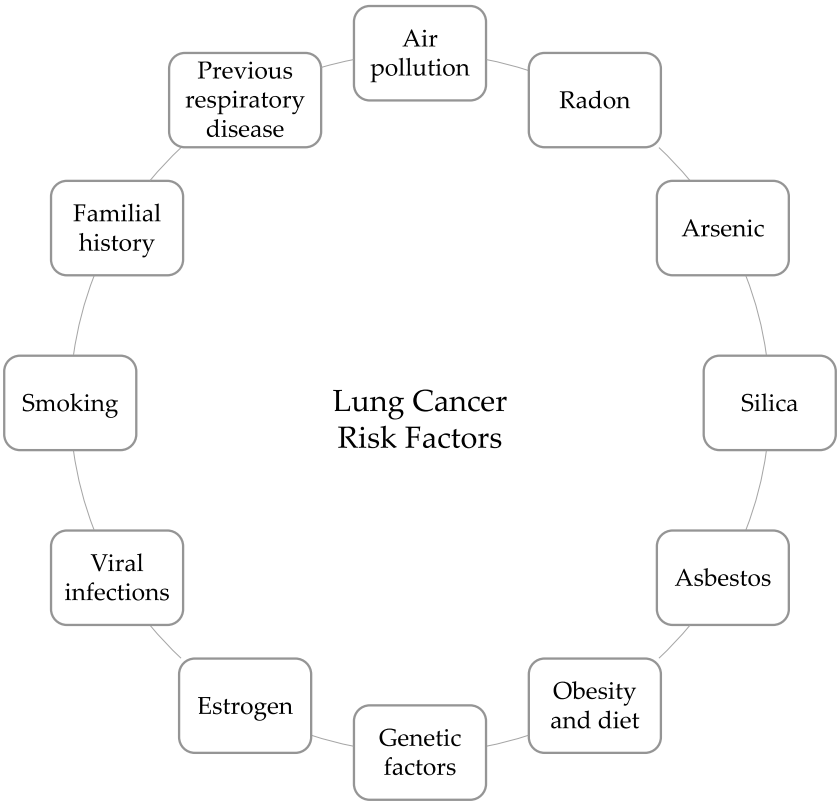


Figure 1. 1| Risk Factors associated with lung cancer. [Adapted from Akhtar, N. & Bansal, J. G. Risk factors of Lung Cancer in nonsmoker. *Curr. Probl. Cancer* 41, 328–339 (2017)][20].

1.1.3 Lung Cancer Detection and Diagnosis

Prognosis of LC patients is reliant upon detection at the initial stages of the disease. Several detection methods are currently being used in the clinic including Chest Radiographies (CXRs), Computed Tomography (CT) scans, Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), and cytology sputum and breath analyses.

The fact that LC is typically diagnosed when at an advanced stage, when survival is poor, makes it an ideal candidate for screening methods. Mass screening of individuals at high risk from LC could potentially be of benefit but have not as yet appeared to reduce mortality[21]. Advances in CT imaging techniques have allowed its application to become the most effective method for early LC detection[22] as it provides more detailed information regarding tumour physical location and nodule size than chest radiography. Compared to traditional radiography techniques, CT scans have been shown to reduce mortality from LC by 20%[23].

Additionally, molecular markers in sputum, bronchial brushings and blood have been investigated but are not currently being used in the clinic. Moreover, there is a lot of research

investment (both time and money) into the growing field of circulating tumour DNA (ctDNA) and Circulating Tumour Cells (CTCs). Although not yet successful, specifically for early detection of LC, a combination of different cancer biomarkers is in ongoing research and could further improve early diagnosis as well as being used to assess response to treatment, the monitoring of tumour burden and detection of relapse.

1.1.4 Lung Cancer Treatment

The type of treatment for LC patients depends on the histological subtype and staging, and includes surgery, stereotactic body radiation, thoracic radiotherapy and conventional cisplatin-based chemotherapy[24,25]. Currently, chemotherapy remains as the standard treatment for LC patients.

Emerging technology platforms are however allowing the molecular alterations that each cancer subtype undergo to be identified, highlighting the importance of genetic profiling of patient tumours. For example, the discovery of *EGFR* mutations[26] and *ALK, ROS1* and *cMET* rearrangements[27,28,29] as effective targets for patients with advanced NSCLC have changed clinical practice and therapeutics. In addition, the accumulation of molecular knowledge has allowed the rapid development of new drugs that specifically target less common molecular abnormalities, such as *HER2*, *RET*, *NTRK,* as well as the *KRAS* G12C mutation. Since very frequently tumours become resistant, and second and third-line treatments are often required, of relevant importance are new studies looking at resistance mechanisms and the newer generation of targeted therapies.

Furthermore, cancer cells often use "checkpoint" proteins or immune cells to avoid being attacked by the Immune System (IS). Drugs that target these checkpoints (checkpoint inhibitors), are being researched in numerous clinical trials for use alone or in combination with chemotherapy. Hence, immunotherapy can help an individual's own IS to recognise and destroy cancerous cells more effectively and have revolutionized LC therapy (Fig. 1.2). For example, immunotherapy with anti-PD1/PD-L1 drugs have become the gold standard for patients with no driver mutations.

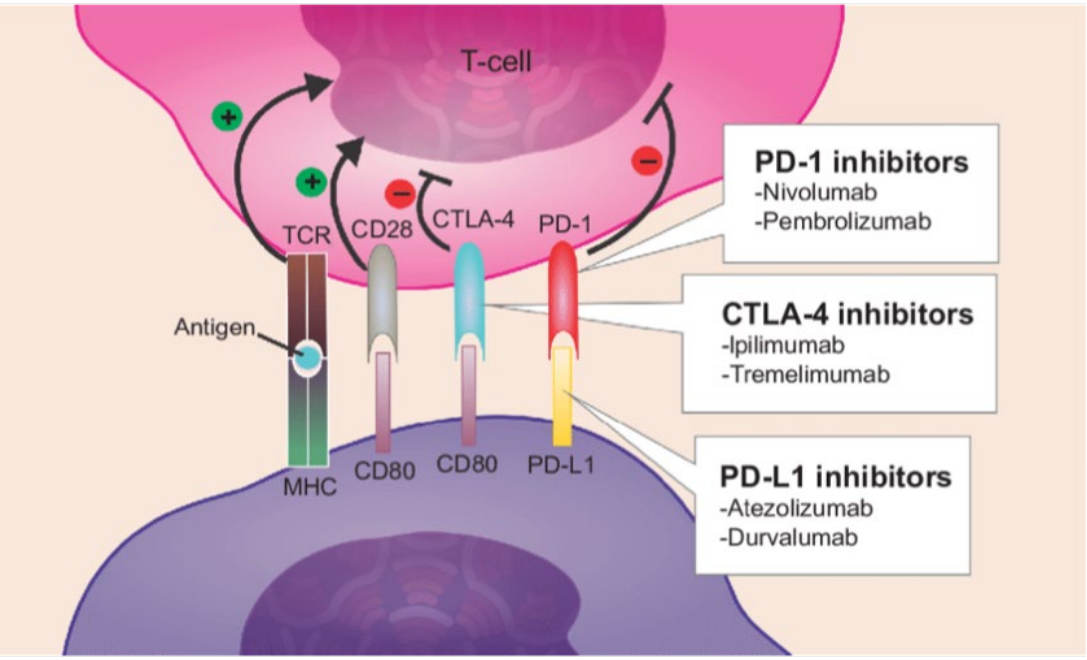


Figure 1. 2| Mechanism of action of immune checkpoint inhibitors in advanced NSCLC. The activity of several proteins on T cells (CTLA-4, PD-1, and PD-L1) function as immune checkpoints since they help in downregulating immune responses. Ipilimumab and tremelimumab are monoclonal antibodies that inhibit CTLA-4, while nivolumab, pembrolizumab, atezolizumab, and durvalumab inhibit PD-1 and PD-L1 as indicated. These drugs act by reducing immune checkpoint activity, thus diminishing tumour evasion. Abbreviations: TCR, T-cell receptor; MHC, major histocompatibility complex. [Taken from: Wakelee, H. Evaluating the Role of Targeted Therapy in Lung Cancer. *Oncology (Williston Park, N.Y.)* vol. 33 (2019)][30].

The current immunotherapy treatments however are effective in only 15%-20% of NSCLC patients[31] and, similar to what is happening for targeted therapies, a lot of ongoing research aims to understand and identify therapies for when an immune therapy does not work for a patient, or when it initially was effective and then stopped having efficacy. Deciding which patients are eligible for immunotherapy with or without chemotherapy still depends on PD-L1 expression levels: immunotherapy is being given to LC patients with PD-L1 high levels, and a combination of both chemo and immunotherapy is being given for those with medium PD-L1 expression levels. If known druggable mutations are detected, then checkpoint inhibitors are not normally given since resistance is very likely. Finally, for LCs with no driver mutations PD-L1 levels are used to decide the use of a single-agent immunotherapy, for patients that are not very symptomatic, or a combination for symptomatic patients because the response rates are higher.

1.1.5 Lung Cancer Challenges

Current key trends in LC research are directed by the major challenges in the management of LCs. Fundamentally, research is being focused on prevention and early detection, improving cure rates in early stages, and management and treatment of advanced and metastatic LC disease.

Although significant advancements have been made in recent years for most other cancer types, there have only been small improvements in the 5-year survival rate among LC patients. The high mortality observed in LCs is associated with diagnosis at advanced stages when symptoms start to appear, which are heterogenous and often mistaken for other problems. The 5-year relative survival rate for all LCs (NSCLC and SCLC combined) is 19%, with the 5-year survival being higher for NSCLC (23%) than for SCLC patients (6%)[32,33].

In addition, lung tumours are characterised by a high degree of molecular heterogeneity because of the multistage carcinogenesis with a combination of genetic and epigenetic processes happening alone or in combination. Different levels of heterogeneity have been recognised in all cancer types: heterogeneity at the genetic level, since mutations are located at different genomic locations and with different timings; heterogeneity at the cellular level within a patient's tumour (intra-tumour heterogeneity) or tumour by tumour (inter-tumour heterogeneity); and finally, at the histological level with spatial heterogeneity, referring to the different regions where tumours are detected (Fig. 1.3).

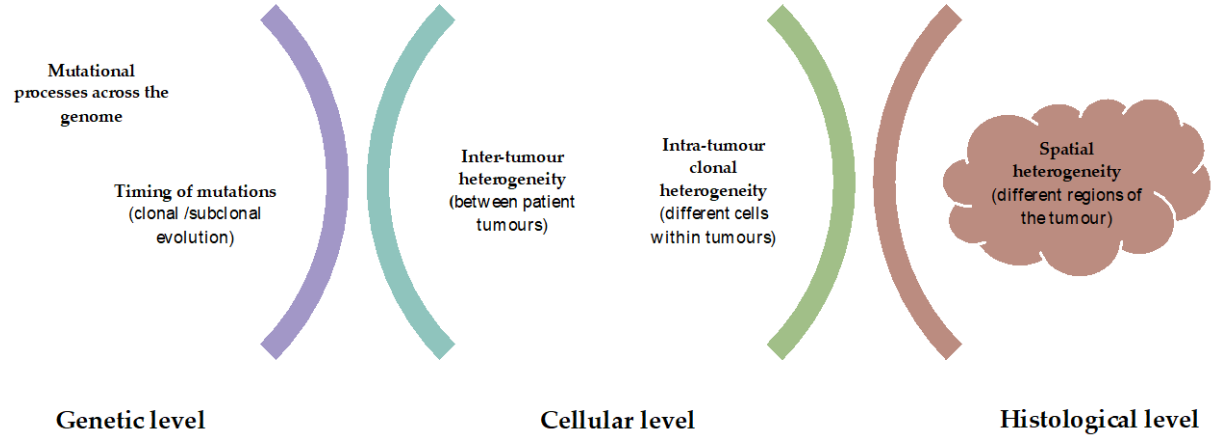


Figure 1. 3| Cancer heterogeneity levels in Lung Cancer (LC). Multiple types of heterogeneity can be found in LC. At the genetic level, with mutations occurring at different genomic locations and rates during tumour evolution; at the cellular level, between patient tumours (inter-tumour heterogeneity) or between different cells within the same tumour (intra-tumour heterogeneity); and at the histological level, when microenvironmental factors shape cancer cell phenotypes and cancer progression.

The genetic background within each individual has been shown to be highly heterogeneous[34,35]. For instance, Campbell *et al*. showed that individuals carry several different clones in healthy skin with two or three driver mutations[36]. It is presumed that environmental exposures together with ageing-related processes and the individual's genetic background could have a combined impact sufficient to initiate tumorigenesis. Thus, the most advantageous genetic and/or epigenetic modifications within a gene can differ between individuals as well as within the same individual when different clones arise simultaneously. Consequently, distinct cellular populations within a tumour can show a diverse spectrum of features: from the expression of cell markers, to the genetic and/or epigenetic alterations, as well as from various non-genetic mechanisms, including stem cell populations and the immune microenvironment of the tumour[37].

Furthermore, primary and metastatic tumours invariably show different molecular lesions and the treatment given can determine the course of future molecular alterations that would favour cancer cells to proliferate. This fact makes primary and secondary drug resistance very frequent among LC patients and, for example, resistance to Tyrosine Kinase Inhibitors (TKIs) in patients with *EGFR*-mutant LC remains a big concern.

As highlighted above, LC heterogeneity remains a major challenge in relation to the detection and treatment of LC patients. Future studies therefore should ideally include comprehensive genomic characterisation of tumour specimens at the time of disease progression to better understand the clonal evolution of tumours under treatment-induced selection pressure.

1.1.6 Lung Cancer Molecular Landscapes

The current LC classification reflects the distinct histopathological features of the disease but has also incorporated, as evidence has emerged, molecular profiles since the latter have changed the way these diseases are treated with specific drugs. Importantly, LC has become a group of histologically and molecularly heterogeneous diseases even within the same histological subtype.

Tumour Mutational Burden (TMB) is defined as the number of somatic, coding, base substitution, and InDel mutations per Mega base (Mb) of a tumour genome examined. Cancers associated to DNA damage frequently appear highly mutated whilst paediatric and well differentiated tumours have usually low TMB[38]. By far, poorly differentiated tumours show the highest mutation rate when considering all changes or when considering only protein-altering changes, with mutation rates of 5.79 and 4.55 mutations per Mb respectively. Still NSCLC appeared as one of the tumours with the highest mutation rate of protein-altering mutations, with LUADs showing mutation rates of 3.5 per Mb and LUSCs 3.9 per Mb, in comparison with an average rate of 1.8 per Mb across all tumour types[39]. Similarly, TMB has been reported to be higher for LNETs with median TMB ranging from 9.9-12.2 mutations per Mb, followed by LUSCs (TMB=9) and LUADs (TMB=6.3). Finally, lung atypical carcinoids alone have been found to have a median of 1.8 mutations per Mb. These large-scale observations are summarised below in Tables 1.1a-b, and highlight the genomic burden variability in LC histotypes[38]. Of relevance, for the clinic, TMB has been shown to correlate with the number of neoantigens and hence identified as a predictive biomarker of immunotherapy response for NSCLC[40] and other tumours.

a

Cancer type	# samples	# mutations	# samples mutated	mutation rate (protein-altering)	mutation rate (synonymous)	mutation rate (all changes)
Breast Cancer	183	524	158	1.14	0.32	1.46
Breast Cancer (HER2+)	59	229	54	1.55	0.41	1.96
Breast Cancer (HR+)	65	141	54	0.87	0.26	1.12
Breast Cancer (Triple Neg)	59	154	50	1.04	0.3	1.34
Lung Cancer	134	1276	129	3.8	0.96	4.76
Lung Cancer (Adenocarcinoma)	57	503	55	3.52	0.89	4.41
Lung Cancer (Carcinoma Small Cell)	5	54	5	4.3	0.56	4.86
Lung Cancer (Carcinoma Squamous	63	620	60	3.92	1.02	4.94
Lung Cancer (Others)	9	99	9	4.55	1.24	5.79
Ovarian Cancer	58	174	54	1.19	0.36	1.55
Pancreatic Cancer	8	16	7	0.8	0.25	1.04
Prostate Cancer	58	48	29	0.33	0.08	0.4
All Cancers	441	2038	377	1.84	0.49	2.33

b

Disease type	Specimen count	Median TMB	Maximum TMB	Percent cases with >20 mutations/Mb	95% CI
Lung atypical carcinoid	83	1.8	180.2	1.2	0.1 - 6.5
Lung adenosquamous carcinoma	154	5.4	73.0	12.3	8 - 18.5
Lung adenocarcinoma	11855	6.3	755.0	12.3	11.8 - 12.9
Lung sarcomatoid carcinoma	130	7.2	165.2	19.2	13.4 - 26.8
Lung non-small cell lung carcinoma (nos)	2636	8.1	173.9	17.0	15.6 - 18.5
Lung squamous cell carcinoma (scc)	2102	9.0	521.6	11.3	10 - 12.7
Lung small cell undifferentiated carcinoma	913	9.9	227.9	9.0	7.3 - 11
Lung large cell neuroendocrine carcinoma	288	9.9	98.2	19.8	15.6 - 24.8
Lung large cell carcinoma	74	12.2	56.8	24.3	14.9 - 33.7

Table 1. 1| A) Mutation rates in different human cancers. [Taken from: Kan, Z. *et al.* Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature* 466, 869–873 (2010)][39]. **B) Tumour Mutational Burden (TMB) across LC histotypes.** [Adapted from: Chalmers, Z. R. *et al.* Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med.* 9, 1–14 (2017)][38].

1.1.6.1 Genetic Alterations in Lung Cancer

1.1.6.1.1 LUADs

LUAD is characterised by a high TMB, frequent cytosine to adenine mutations (C>A) (explained by carcinogen exposures) and regional heterogeneity, where mutation rates are lower in highly expressed genes due to transcription-coupled repair mechanisms. Significantly mutated oncogenes include *TP53*, *KRAS*, *STK11* (also referred to as *LKB1*), *EGFR*, *BRAF*, *MET* and *NF1*, although the mutations seen in these genes vary depending on smoking status and gender. For instance, mutations in *EGFR* are more prevalent in women and in non-smokers. Additionally, *KRAS* mutations are frequently seen in smokers. Rearrangements are commonly found in *ALK*, *ROS1*, *RET* and *NTRK1*.

Copy Number Alterations (CNAs) include gains in chromosome 5p15, suggested to be targeting the Telomerase Reverse Transcriptase (*TERT*), and amplifications in 14q13.3 (*NKX2-1)* and *MYC* are also often observed. Other recurrent amplifications involve *EGFR*, *MET*, *KRAS*, *ERBB2*, and *MDM2* genes, and other deletions are in the genes encoding *LRP1B*, *PTPRD* and *CDKN2A*[41].

1.1.6.1.2 LUSCs

LUSCs begin in the top layer of the cells (squamous cells) that line the bronchi of the lung. LUSCs represent around 30% of all LCs. Recurrent mutations have been detected in genes associated with the cell cycle and apoptosis (*TP53*, *CDKN2A* and *RB1*), antioxidant pathways (*NFE2L2* and *KEAP1*), phosphatidylinositide 3-kinase signalling (*PIK3CA* and *PTEN*), epigenetic signalling (*MLL2*) and proliferation, apoptosis and squamous cell differentiation (*NOTCH1).* Additionally, inactivating mutations in the *HLA-A* locus have been suggested to predict response to immunotherapy. Furthermore, CNAs have been widely reported in LUSCs and include amplifications in *SOX2*, *NFE2L2*, *PDGFRA*, *FGFR1* and *CCND1* with deletions in *CDKN2A*[42,43] also observed. Other new amplifications have been recently found in *MYC*, *CDK6*, *MDM2*, *BCL2L1* and *EYS* with new deletions of *FOXP1*, *PTEN* and *NF1*[43] being found.

In LUSCs, 5'-C-phosphate-G-3' (CpG) transitions and transversions resulting in C>T and C>A mutation types are the most commonly observed type of mutation reported, whereas in non-CpG sites, C>A transversions have been found to be more common[43].

1.1.6.1.3 LNETs

Among neuroendocrine tumours of the lung, SCLC and LCNEC represent high-grade malignancies. Both present frequent aneuploidy as well as chromosomal alterations of greater than 10/Mb with means of 18.8 and 13.7 aberrations per tumour type respectively[44].

Compared with other neuroendocrine tumour types, LNETs are characterised by relatively few mutations and chromosomal aberrations. In pulmonary NETs, mutations have been found typically in genes encoding for proteins directly involved in the maintenance of epigenetic status, including members of the SWI/SNF complex, sister chromatid cohesion pathway related genes and histone covalent modifiers (Fig. 1.4).

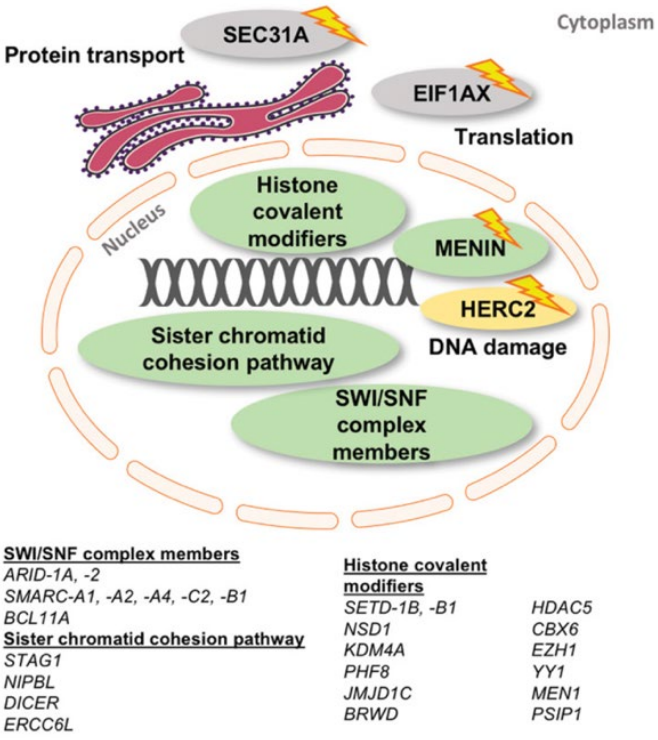


Figure 1. 4| Mutated genes and affected signalling pathways in LNETs. [Taken from: Di Domenico, A., *et al*. Genetic and epigenetic drivers of neuroendocrine tumours (NET). *Endocr. Relat. Cancer* 24, R315–R334 (2017)][45].

The top most frequently altered gene in LNETs is also *TP53*, followed by *LRP1B*[46]. *TP53* (90%) and *RB1* (65%) are hallmark mutations of SCLCs, although mutations in *RB1* have also been found in LCNECs[46,47]. Other significantly mutated genes that have been identified in SCLC are *KIAA1211* and *COL22A1* as well as *RGS7*and *FPR1* - both of which are involved in G-protein-coupled receptor signalling.

Furthermore, genomic losses within 3p pointing to focal events on 3p14.3–3p14.2 (harbouring *FHIT*) and 3p12.3–3p12.2 (harbouring *ROBO1)* have been reported. Loss of the *CDKN2A* locus, together with amplification of the *MYC* family genes (*MYCL1*, *MYCN* and *MYC)*, as well as of the tyrosine kinase gene *FGFR1*, and *IRS2* were recently identified[48].

On the other hand, LCNECs harbour very frequent *TP53* mutations (92%) followed by *RB1* mutations (42%). Other genes mutated in LCNECs include *STK11*, *KEAP1*, *ADAM*, *MYCL1*, *NKX2-1*, *RAS*, *BRAF* and *NFE2L2*. CNAs include *MYCL1, FGFR1, NKX2-1* and *MYC*. Statistically significant deletions have been found in *CDKN2A* (9p21, 8%) and a putative fragile site at *PTPRD* (9p24, 7%)[47]. Other genes identified mutated in both SCLCs and LCNECs include *FAT3*, *SMARCA4*, *NOTCH3*, *PIK3CG*, *PIK3CA*, and *KMT2D*[46].

1.1.6.1.4 L-CDs

Finally, molecular profiling of lung carcinoids (L-CDs) have previously shown chromatin remodelling genes, such as *MEN1*, *ARID1A*, *PSIP1*, *KMT2C* and *KMT2A* to be recurrently mutated, while *TP53*, *RB1* and *STK11* mutations have been found frequently altered in non-carcinoid LNETs[49].

Studies have emphasized the distinction between TCs and ACs as the most important prognostic factor. Molecular events distinguishing these subtypes include common *MEN1* and *TP53* mutations or deletions; *KMT2C* mutations; and *TERT*, *SDHA* and *RICTOR* amplifications in ACs; whereas TCs are not characterised by recurrent mutations but *RB1* deletions *MEN1* deletions at frequencies of maximum 15%.

1.1.6.1.5 Smoking and the Cancer Genomic Landscape

It is important to mention the different genomic landscapes of lung cancer between never smokers and smokers with the latter containing a significantly higher mutation frequency, predominantly cytosine to adenine (C>A) nucleotide transversions and non-actionable mutations such as those in *KRAS* and *TP53*. By contrast, never smokers usually have a predominant transition of cytosine to thymine (C>T, and a higher prevalence of actionable driving gene alterations including activating *EGFR* mutations, and *ROS1* and *ALK* translocations[14,17]. In addition to the differences at the molecular level, epidemiological and clinical differences have been observed between LCs arising in never smokers and smokers. For instance, never smokers that develop LC seem to do it at younger ages, although not in the United States and Europe where LC is diagnosed at the same or older age[50,51]. Moreover, women are more frequently affected than men and adenocarcinoma is the most common histological subtype in never smokers.

1.1.6.2 Epigenetic Alterations in Lung Cancer

Epigenetic mechanisms, mainly DNA methylation, histone modification, and noncoding RNAs (ncRNAs), are dynamic and reversible modifications that are involved in some important biological processes and together with genetic events affect cancer hallmarks.

DNA methylation is the most studied epigenetic modification in cancer responsible for the silencing of genes and chromatin structure. Promoter hypermethylation has been observed in several genes in LC as detailed in Table 1.2.

Gene	Mechanism	Epigenetic modification
RASSF1A	DNA repair; cell cycle	Hypermethylation
MGMT	DNA repair	Hypermethylation
CDKN2A/p16	Cell cycle	Hypermethylation
DAPK	Apoptosis; autophagy	Hypermethylation
P14	Proliferation; apoptosis	Hypermethylation
OTUD4	Cell cycle; apoptosis; DNA repair	Hypermethylation
CDH1/E-cadherin	EMT	Hypermethylation
RARβ	Metastasis	Hypermethylation
RUNX3	TGF-*β*/Wnt signaling pathway	Hypermethylation
APC	Wnt/*β*-catenin signaling pathway	Hypermethylation

Table 1. 2| Abnormally methylated genes in Lung Cancer. For each gene their functional role is given as well as the type of epigenetic modification that has been reported. Abbreviations: EMT, Epithelial-to-Mesenchymal Transition. [Taken from: Shi, Y. *et al*. Current Landscape of Epigenetics in Lung Cancer: Focus on the Mechanism and Application. *J. Oncol.* **2019,** 8107318, (2019)][52].

1.1.6.2.1 LUADs

In LUADs, *RUNX3* is involved in differentiation of the lung epithelial-lineage and also functions as a Tumour Suppressor Gene (TSG) that has been found inactivated often by DNA hypermethylation[53]. *CDKN2A*, the cyclin-dependent kinase inhibitor 2A gene encoding for the tumour suppressor proteins p16/p14-ARF with roles in regulating cell cycle, and the DNA repair gene *MGMT* have also both been found to be inactivated via promoter hypermethylation although the *MGMT* more frequently in never-smokers and advanced stages[54]. Inactivation of the TSG *RASSF1A* via promoter hypermethylation has been observed in 34% of NSCLCs, allowing tumour invasion and metastasis[55]. Finally, *DAPK* encoding for the death-associated protein kinase (a serine/threonine kinase) has been found to be methylated on average in 40.5% of NSCLCs[56].

1.1.6.2.2 LUSCs

LUSCs were recently profiled at the whole-genome level and revealed novel epigenetic signatures, with 44 genes identified for which DNA methylation level correlated with expression level. Aberrant methylation events were promoter hypermethylation of *SOX17* and *WIF1*, as well as other novel genes including *SFTA3*, *TCF21*, with hypomethylation of *AKR1B10*, the aldo-keto reductase family 1-member B10; B-assrestin-1 (ARRB1), the gap junction protein gene *GJB5* and *SEPINB5*, among others[57].

1.1.6.2.3 LNETs

As previously mentioned, (Section 1.1.6.1.3), LNETs display frequent mutations in genes encoding for epigenetic regulators influencing DNA methylation patterns. As a result, SCLC tumours for instance have been found to show significant hypomethylation compared to normal lung together with significant representation of DNA methylation peaks in neuroendocrine-specifying Transcription Factor (TF) genes including *BCL2*, *NEUROD1*, *ASCL1*, *HAND1*, *ZNF423*, *REST*, *TCF21* and *RB1*. Similarly the same observations have been made for several genes encoding for the polycomb repressive complex (PRC)[58,59], as well as

CASP8, *FAS* and *TRAIL-R1* gene hypermethylation-promoter silencing. Additionally, a molecular group of SCLC tumours showed high expression of *EZH2*, a member of the PRC2 complex that promotes trimethylation of histone H3 lysine 27 (H3K27me3) related to cell self-renewal and stemness, concomitant with increased methylation in CpG island-containing promoters.

Not many studies have focused on the DNA methylation profiles of LCNECs with large sample sizes but instead LCNECs have been studied together with NSCLCs and SCLCs. For instance, a study including 9 LCNECs together with LUADs, LUSCs and SCLCs found hypermethylated genes enriched in regulation of transcription, neural development and cell morphogenesis[57]. Unsupervised clustering performed in the study resulted in LCNECs grouped together with SCLCs suggesting similar genome-wide DNA methylation landscapes in neuroendocrine carcinomas. Furthermore, the TSG *RASSF1A* was found hypermethylated and downregulated in most pulmonary NETs including LCNECs, SCLCs, and L-CDs[60], and other cancer types[61], with its loss associated with cell proliferation.

Finally, given the observed role of epigenetic alterations in pulmonary tumours recent studies have integrated DNA methylation data with genetic and transcriptomic data. For instance in a study of L-CDs, three different clusters were found by Laddha *et al*.[62] which were characterised by the expression of *ASCL1*, *HNF1A* and *FOXA3*. Subsequently Alcala *et al*.[63] found, using a multi-omics integrative analysis, that *DLL3* and *SLIT1* also helped to differentiate L-CDs by. In the Laddha *et al.* study, *HNF1A* and *FOXA3* genes were hypermethylated and showed low expression in one group; *FEV*, *GATA2* and *PROCR* were hypomethylated and highly expressed in a second group; whilst in a third group *SOX1* was hypermethylated and highly expressed, and *SIX2*, *ONECUT2* and *IL1RL2* hypomethylated and highly expressed. The observations from these studies however need further investigation including mechanistic studies but this is not straightforward due to the current lack of appropriate L-CD cell lines and/or animal models. Additionally, most of the epigenetics studies are carried out in small data sets (L-CDs: $n<30$[62]; $n=63$[63]) due to the lower incidence of this type of cancer, thus larger collaborative as well as validation studies are still needed to confirm the findings of Laddha *et al.* and Alcala *et al*.

1.2 Cancer Genomic Landscapes

Carcinogenesis is a multistep process that involves four stages: initiation, promotion, progression, and malignant conversion. Exposure to carcinogens can lead to DNA damage, which if not repaired, results in mutation. In this context, an efficient DNA machinery may be sufficient to avoid cancer initiation. Otherwise a selective growth advantage acquired with a permanent mutation of a vital gene may be sufficient to initiate cancer in a cell if unrepaired. The promoters influence the efficiency of the carcinogenic process by allowing clonal expansion of initiated cells. These are normally non-mutagenic and non-carcinogenic, and generally this is an epigenetic event as the manifestation is seen at the gene expression level rather than in the genetic sequence. After that, the progression stage sees an increase in the number of defective cells with a permanent genetic growth advantage. This stage, which is irreversible, will eventually lead to malignant conversion. There may be between these four stages long gaps, and several cellular, genetic and epigenetic changes can take place to support the process.

During the process the cancer cells also exhibit different distinctive features, such as changes in the size and shape of the nucleus, and hampered tissue organization (Fig. 1.5). These new features are usually useful in the clinic for classification purposes, as different cell types and phenotypes can be used to recognise specific tumours.

Normal cells are controlled by growth suppressors. The latter mainly instruct cells to stop division, to undergo apoptosis or fix changes in damaged cells. Mutations can alter these control checks and alter the sequences of protein-coding genes of key genetic pathways. Genes, or the proteins transcribed by them, are organized in complex pathways and feedback mechanisms to control the levels of expression of genes that participate in cellular processes for the normal control of cells and tissues. Thus, some of these mutations may be "driver mutations" which drive the normal cell to be transformed into a cancerous one. Others may be passenger mutations having no impact on the cell.

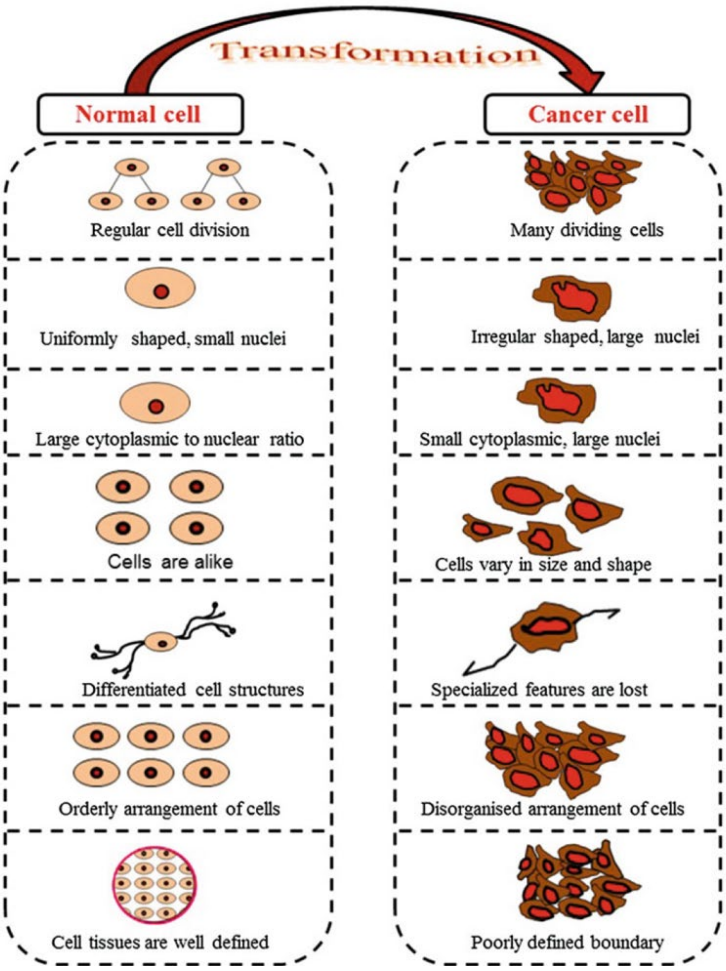


Figure 1. 5| Differences between normal and cancer cells. [Taken from: Epigenetics, Energy Balance, and Cancer. in *Energy Balance and Cancer* (ed. Berger, N. A.) 167–189 (Springer US, 2016). doi:10.1007/978-3-319-41610-6_7][64].

1.2.1 Driver and Passenger Mutations

Technological advances in molecular biology and Next-Generation Sequencing (NGS) platforms, along with microarray-based technologies, have enabled the obtention of large amounts of data. The latter, together with new bioinformatic tools, are enabling the molecular landscapes of human cancers to be deciphered.

Of particular interest is the identification of alterations in oncogenes, which generally encode proteins that regulate processes with potential to cause cancer. A "driver" mutation refers to somatic mutations that are able to improve the fitness of the cell, whereas "passenger" includes incidental mutations that do not confer any growth advantage and are neutral for the cell[65]. Nonetheless, single nucleotide changes within genes together with small

Insertions and Deletions (InDels) of <50 bases, are only a subset of the different abnormalities underlying cancer. Copy Number Alterations (CNAs), including amplifications and deletions; inactivation of genes through epigenetic silencing; and chromosomal translocations are also known to contribute to cancer development and progression.

As for driver genes they have been historically classified into two types of cancer genes based on whether they functioned to promote, termed as oncogenes (OCGs), or inhibit tumorigenesis, termed as Tumour Suppressor Genes (TSGs). Several studies have focused on cancer genes of these two categories. Nevertheless there is now increasing evidence that cancer genes may exhibit multiple and often contrasting functions, and the switch between OCG and TSG can be regulated at the DNA, RNA or protein level[66]. For instance, the dual role of RASSF1 in neuroendocrine tumours of the lung was already reported in 2021. In this case, two different promoters control the expression of different isoforms of the RAS-association domain family 1 (*RASSF1*) gene at 3p21.3[60,67]. Hypermethylation of one promoter was detected in tumour LNETs, as compared to the non-neoplastic lung and NSCLCs, and found responsible for the silencing of the RASSF1A/E isoform with TSG roles. The other promoter, responsible for the expression of *RASSF1C*, was never found hypermethylated in LNET, NSCLCs and paired lung tissue samples, and was detected at increased expression in all types of LCs. This was in line with its previously suggested OCG role of inducing cell proliferation and migration[68,69].

1.2.2 Important Pathways in Cancer

Different hallmarks have been widely associated with cancer and the different stages of carcinogenesis. Specifically, ten characteristic features were presented initially in the year 2000, updated by Hanahan and Weinberg in 2011[70], and include sustaining proliferative signalling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, activating invasion and metastasis, reprogramming energy metabolism, and evading the immune system. Two other traits that can also contribute to cancer development are genetic instability and mutations[71], and tumour-promoting inflammation[72,70]. To reach conclusions about the presence or absence of a hallmark process, these hallmarks have been associated with genes, biological pathways, or functional

properties (functional annotation). The use of well-annotated resources, such as Gene Ontology or biological pathways, allow data to be organised and classified for integrative approaches[73,74], as well as enabling the systematic association of hallmark concepts that are relevant for the description and interpretation of cancer research results.

Tumorigenesis is a result of the accumulation of many alterations in many genes, which are, in turn, under finely coordinated regulatory expression programmes. For example, overexpression or under-expression of a TF may result in overexpression or under-expression of other downstream genes. Several key genetic pathways have been implicated for progression of cancer and a brief review of these findings will be given next.

1.2.2.1 Genomic Instability

Genomic instability is a characteristic of almost all human cancers, however the stage at which it occurs during cancer development and its molecular basis are still unknown. Several pathways are involved in genomic stability maintenance, such as DNA damage check point, DNA repair pathways, mitotic checkpoints and telomere maintenance[75]. If these fail, shorter cell cycles and/or an advantage to bypass control systems can allow defective cells to keep proliferating resulting in cancerous tumour formation.

Genomic instability can be manifested through high frequencies of base pair mutations, Microsatellite Instability (MSI) and Chromosome Instability (CIN), all of which can happen alone or in combination in cancer:-

a. Increased frequencies of base pair mutations happen when DNA damage repair genes are lost and in combination with environmental or intrinsic factors. This can result in increased mutation rates and clonal evolution, referred to as "a mutator phenotype"[76].
b. Microsatellites are simple tandem nucleotide repeats scattered across the genome which as a consequence of their repetitive nature mismatches as well as insertion-deletion loops from replication slippage can occur naturally[77,78].
c. Chromosomal instability refers to incorrect number of chromosomes and/or abnormal chromosome structures resulting from the aberrant chromosome mis-segregation during cell division. Several linked but different mechanisms enable mitotic fidelity and have been intensely studied since aneuploidy was discovered in human tumour cells[79].

Additionally, many proteins have also been related to chromosomal instability examples being APC, BRCA1 and BRCA2, p53, Rb and Aurora.

These manifestations are normally overcome by sophisticated mechanisms in normal cells, mainly the Mismatch Repair (MMR) system, the transcription-coupled Nucleotide Excision Repair (NER) and the Base Excision Repair (BER) pathway[80]. Failure of repairing however can lead to mutational signatures, as marks made of specific mutational combinations and distributions resulting from the interplay between the damage created by mutagens, DNA repair systems, and the replication machinery (Fig. 1.6)[81]. Thus, the aetiologies can be inferred from the mutational signatures if the interplay between these errors and the repair machinery is known [82].

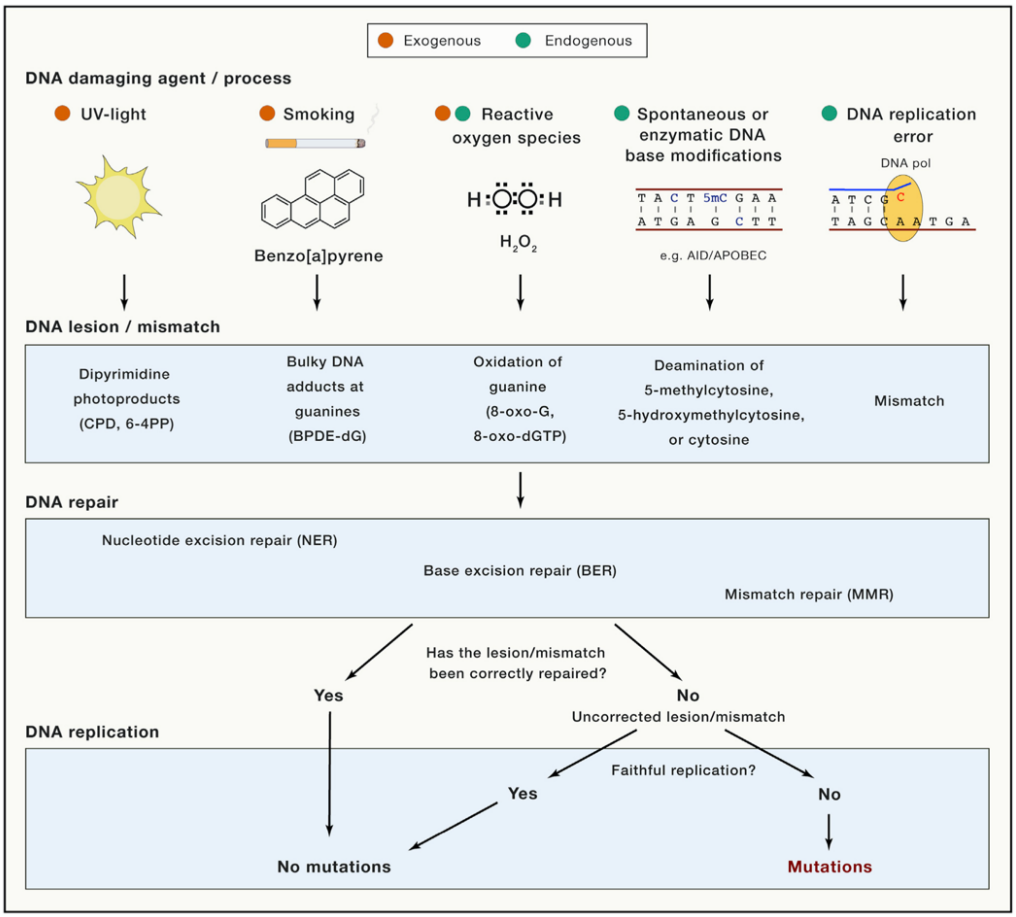


Figure 1. 6| Mutations result from the interplay between DNA lesions generated by damaging agents and DNA repair mechanisms. There are several DNA damaging agents/processes that lead to DNA lesions and mismatches. These lesions or mismatches are recognised and subsequently repaired by different machineries that can successfully repair and return to the original DNA sequence or leave uncorrected. In the latter case, lesions can lead to mutations after DNA replication if high-fidelity polymerases do not ultimately introduce the correct nucleotide. [Taken from: Gonzalez-Perez, A., Sabarinathan, R. & Lopez-Bigas, N. Local Determinants of the Mutational Landscape of the Human Genome. *Cell* 177, 101–114 (2019)][81].

For instance, bulky lesions of the DNA or adducts can be generated by tobacco carcinogens, normally at guanines. These adducts are subsequently recognised and repaired by several mechanisms, including the NER system. Such lesions, if unrepaired, can promote the collapse of the replication fork through DNA polymerase stalling. Factors such as the local DNA structures and the sequence context adjacent to an adducted nucleotide determine repair specificity and efficiency[83], as well as the resulting mutation, which primarily involve G>T (C>A) for tobacco carcinogens[84]. Mutational signature 4[82] is characterised by prevalent C>A mutations on transcribed strands, consistent with the propensity that many tobacco carcinogens have to form adducts on guanine. The causal relationship of this signature with tobacco smoking has been supported by a strong positive association between smoking history and the contribution of signature 4 to individual cancers, being mostly LC, head and neck squamous, and liver cancers[85]. As a result, signature 4 is widely recognised to be associated to tobacco smoking.

1.2.2.2 Apoptosis or Programmed Cell Death

Cancer cells overcome the mechanisms implicated in apoptosis to destroy homeostatic balance. Apoptosis can be triggered by both extrinsic and intrinsic pathways, both of which finally induce cell death by the activation of different procaspases in a cascading manner[86].

In the extrinsic pathway, receptors transcribed form the Tumour Necrosis Factor (TNF) superfamily of genes, and together with its ligands and adapter proteins activate a Death-Inducing Signalling Complex (DISC) by the binding of procaspase-8. Relevant proteins include FasL/FasR that pair with the FADD adapter, or TNF-α /TNFR1 that pair with TRADD/FADD/RIP adapters.

In the intrinsic pathway, mitochondrial events are controlled by the Bcl-2 family of proteins and cause the opening of the Mitochondrial Permeability Transition (MPT) pores resulting in the release of pro-apoptotic proteins into the cytosol. For example, cytochrome c among others ultimately activates the procaspase-9, and Smac/DIABLO abd HtrA2/Omi inhibits the “inhibitors of apoptosis” or IAP proteins. A second group of proteins includes

AIF, endonuclease G, CAD and caspase-3, which participate in the condensation of chromatin and its cleavage.

1.2.2.3 TFG-β Signalling Pathway

TFG-β is an extracellular signalling ligand that induces the expression of certain TFs that in turn affect transcription of hundreds of genes. The main agents in signalling the canonical TFG-β pathway include the Smad family of proteins. TFG-β acts as a tumour suppressive signalling mechanism in normal cells by controlling cell cycle for preventing proliferation, promote apoptosis or induce cell differentiation[87]. It however can switch to a OCG role once the TSG role of TFG-β signalling has been overcome by cancer cells and can then induce Epithelial to Mesenchymal Transition (EMT) and angiogenesis[88].

1.2.2.4 Nf-κB Signalling Pathway

Whilst extracellular signals activate the TFG-β pathway, the Nuclear Factor Kappa-light-chain-enhancer of activated B cells (Nf-κB) is rapidly induced intracellularly and controls several processes such as DNA transcription, cytokine production, inflammatory responses, cell adhesion, migration and survival[89,90]. Main signals include stress, heavy metals, ROS, ultraviolet radiation, and antigens coming from viruses and bacteria.

The nuclear TF represents a family composed of five structurally related members including NF-κB1 (also named p50), NF-κB2 (also named p52), RelA (also named p65), RelB and c-Rel. The NF-κB proteins are present in the cytoplasm in their inactive form sequestrated by the Inhibitor of κB (iκB) proteins. Different internal stimuli can then activate the degradation of these inhibitors by the iκB kinase, which will allow the release and the transcription of specific genes when the nuclear factor binds to regulatory DNA motifs. This way, the activation of Nf-κB pathway has been normally associated with oncogenic processes and overexpression of iκB showed to reduce tumour growth[91].

1.2.2.5 Wnt Signalling Pathway

Understanding of the Wnt pathway continues to evolve since the discovery of its first member, the proto-oncogene int-1 gene, which was then renamed as Wnt or "wingless/integrated". Recognition of the importance of this pathway in cancer started when the human tumour suppressor Adenomatous Polyposis Coli (APC) protein was found to downregulate β-catenin, with Wnt-1 found to upregulate it[92].

The canonical Wnt pathway involves the β-catenin protein as the transcription co-activator. When Wnt is released and binds the Frizzled (Fz) family of receptor proteins found in the cell surface, helped by other co-receptors such has LRP5/6, a signal is transmitted by direct contact to the Dishevelled (Dsh) protein which inhibits the elimination of β-catenin by the destruction complex formed by Axin, APC and PP2A. Hence, after Wnt binding, this complex is inactivated liberating the β-catenin which then translocates to the nucleus. There it acts as a TF, together with LEF and TCF TFs, enhancing the transcription of several target genes such as *MYC*, *VEGF*, *FGF4*, *FGF18*, the cyclin D1 gene, the E-cadherin gene and *NRCAM*[93], and many other genes of the same signalling pathways (https://web.stanford.edu/~rnusse/wntwindow.html).

1.2.2.6 MAPK/ERK Signalling Pathway

The MAP/ERK pathway is also initiated by external stimuli and its main function is to transmit this signal from outside the cell to its nucleus. This pathway is also called the RAS-RAF-MEK-ERK pathway due to the proteins involved (RTK and RAS), and the three protein kinases (RAF, MEK and ERK). The external molecule is often a mitogen, for example the Epidermal growth Factor (EGF), that signals cell division and proliferation which in turn activates the tyrosine kinase receptor (EGF Receptor [EGFR], or others such as Trk A/B and the Fibroblast Growth Factor Receptor [FGFR]). Then, the receptor is phosphorylated and a complex (GRB2 and SOS) binds and induces the binding of Ras to Guanosine Triphosphate (GTP) which becomes active and starts a cascade of activation by phosphorylation of the RAF kinase and followed by MEK, MAPK, and the target TFs. It is the latter that then regulate the transcription of genes that control cell cycle[39].

The RAS family of proteins, including HRAS, KRAS, and NRAS, and the RAF family protein BRAF, are often mutated in cancer. For instance, KRAS mutations at codons 12 and 13 are the most frequent hotspot mutations in NSCLC that cause the hyperactivation of many downstream effector pathways[94].

1.2.2.7 PI3K/AKT/mTOR Signalling Pathway

The PI3K/AKT/mTOR signalling pathway includes a family of intracellular lipid kinases that phosphorylate phosphatidylinositides or PIP2 that are activated intracellularly. The Phosphatase and Tensin homolog (PTEN) and PI3K inhibit or activate the activity of this pathway by dephosphorylation (PIP3 to PIP2) and phosphorylation (PIP2 to PIP3). PIP3 then can bind to the PDK1 protein that phosphorylates and activates AKT, also named as Protein Kinase B (PKB), which plays a role in multiple cellular processes such as protein synthesis and cell growth (via GSK3, mTORC1, mTROC2 and 4E-BP1), proliferation and inhibition of cell apoptosis (P21/Waf1/Cip1), motility, adhesion, neovascularization, and cell death[95].

1.2.2.8 Other Relevant Pathways in Cancer

Hippo, Myc and Notch and Nrf2 are other common pathways dysregulated in cancer. Hippo signalling controls organ size by regulating cell proliferation, apoptosis and stem cell renewal, and a kinase cascade involving Mst1/2, SAV1 and LATS1/2 kinases that end up with the inhibition of YAP and TAZ effectors[96].

The MYC signalling pathway lies at the crossroads of many growth promoting signalling pathways and is an intermediate molecule of many ligand-membrane receptor complexes[97]. For instance, MYC expression is transactivated upon nuclear translocation of B-catenin[98].

The NOTCH pathway involves the activation of the Notch receptor by its ligands, Delta and Serrate (known as Jagged in mammals), which function during diverse developmental and physiological processes including self-renewal and differentiation[99]. Its activation results in turn in the activation of the canonical Notch target genes: *Myc*, *p21* and the *HES*-family members., Its oncogenic or tumour-suppressor like activity however are highly context dependent[100].

Finally, Nrf2 is a cellular protector that is dissociated from its inhibitor KEAP-1 (when ROS levels rise) and translocates to the nucleus where it can control an estimated 250 genes the latter mainly involved in endogenous antioxidant protection and detoxification[101–103].

1.3 Epigenetics and Cancer

Signalling pathways involve cascades of signals that generally result in the activation or repression of downstream genes when specific TFs bind their regulatory sequences and the transcription machinery is recruited. The specific recognition of Response Elements (REs) by TFs at promoters and enhancer regions enables gene-specific transcription initiation. For gene transcription to occur, however, the gene regulatory regions need to be accessible to TFs and other regulatory units. In this context, epigenetics refers to heritable changes that do not affect the genetic sequence but control gene expression. Different epigenetic mechanisms are involved in gene regulation, including DNA methylation, histone modifications, nucleosome positioning and aberrant expression of non-coding RNAs, specifically microRNAs.

Altered DNA methylation patterns is considered a cancer hallmark and it is the most studied epigenetic modification and is the type therefore focused on within this thesis.

1.3.1 DNA Methylation

DNA methylation refers to the addition of a methyl group covalently to the base cytosine. In vertebrates, DNA methylation mainly occurs at cytosines in a CpG dinucleotide context. Most CpG dinucleotides in the human genome are methylated. CpGs, however, are not normally distributed as they have been severely depleted in the vertebrate genome to about 20% of the predicted frequency. The only exception of this so-called global CpG depletion is the specific category of GC- and CpG-rich sequences termed CpG Islands (CGIs) that are frequently located at the promoter regions of coding genes, where they are generally unmethylated. DNA methylation can directly prevent TF binding and lead to changes in chromatin structure that restrict access of TFs to the gene promoter[104]. As a result, aberrant DNA methylation can play a role in silencing of TSGs and activation of OCGs. In addition, methylation of CpG sites outside of islands is associated with transcriptional regulation; for example, methylation

within gene bodies is positively correlated with expression, while CpGs are often unmethylated at active enhancers[105,106].

According to the initial studies in the field, gene silencing seemed to be the principal function of DNA methylation., Accumulated evidence, however, has proven that it is also involved in multiple physiologic processes such as development and cell differentiation, genomic imprinting, X-chromosome inactivation, suppression of repetitive elements and genomic instability.

DNA methylation is a dynamic and reversible modification influenced by both environmental and intrinsic factors[107] as well as the DNA sequence and nongenetic *trans*-acting factors, all contributing to interindividual variability of individual landscapes[108,109]. Each cell type has the same genetic sequence but there can be different epigenomes that define cell fate and differentiation status[110].

DNA modifications are controlled by several epigenetic regulators, named "writers", "readers", "erasers" and "regulators". In mammals, writers include three canonical DNA Methyltransferases (DNMTs), DNMT1, DNMT3A and DNMT3B, that catalyse the addition of methylation marks to genomic DNA. DNAMT3A and DNMT3B are involved in the *de novo* methylation of DNA while DNMT1 is primarily responsible for the methylation of newly replicated DNA i.e. for the maintenance of specific patterns of methylation of a cell's genome through cell division[111].

The demethylation of 5mC can occur passively during the replication process or actively and involves the Ten-Eleven Translocation (TET) family of proteins or the Activation-Induced cytidine Deaminase (AID) system followed by BER that introduces an unmethylated cytosine.

Methylated CpG islands can also attract "readers" that bind to methylated DNA through a Methyl CpG binding Domain (MBD). These include MECP1 and MECP2, which can attract other histone deacetylases and chromatin remodelling subunits that normally reduce transcription of methylated gene promoters[112].

Other epigenetic modifiers include histone acetyltransferases (HATs), histone deacetylases (HDACs), histone methyltransferases (HMTs), histone demethylases (HDMs), and chromatin remodelling factors. In addition, remodelling complexes can alter the position of nucleosomes along the chromatin fibre and/or modify the association of histone octamers

with the DNA on major regulatory elements of transcription - promoters and enhancers nucleosomes - to activate transcription when the transcription machinery has been assembled[113]. Some examples of remodelling complexes include NURF, the SWI/SNF (BAF in mammals) and chaperone complexes.

1.3.2 Histone Modifications

In addition to DNA methylation, histone proteins play a critical role in the epigenetic regulation of gene expression. Histones control the packaging of DNA in the nuclei in structures called "nucleosomes" and, together with DNA methylation and other modifications, regulate the expression of genes. Cells have evolved elaborated mechanisms to dynamically modify the level of chromatin compaction to modulate gene function, as well as to allow the enormous DNA sequence to fit in the nucleus of cells facilitating distribution of the genetic material to daughter cells after replication and to regulate DNA accessibility for DNA damage repair. As such, chromatin can be highly compacted as *heterochromatin* or accessible as *euchromatin*. These regions contain specific combinations of epigenetic marks that play crucial roles for genetic stability. For instance, a key function of heterochromatin is to prevent the activity of transposable elements present in more than half of the genome.

Moreover, histones are subject to Post-Translational Modifications (PTMs) such as covalent modifications that are mostly concentrated towards more accessible N-terminal tails that protrude out of the nucleosome. This way histone modifications can exert several genomic functions by recruiting specific effectors, from altering the compaction degree of the chromatin fibber; serve as targets for other factors related with transcriptional activity; or mark the silencing and expression of target genes. Histones and their PTMs create a specific "histone code" that relate to particular chromatin states and functions. An example is H3K4me3, which marks active gene transcription and is normally enriched at the promoters of active genes[114,115].

1.3.3 Non-coding RNAs (ncRNAs)

Modern sequencing technologies have revealed that most of the genome is transcribed into non-coding transcripts that by far surpass the number of coding genes. These ncRNA species

comprise another layer of epigenetic regulation although their full functional repertoire has just started to be elucidated. Nevertheless they have been shown already to play crucial roles in nuclear functions such as transcription, RNA splicing, translation and chromatin remodelling, as well as being key regulators of proliferation, differentiation, apoptosis and cell development[116]. Historically they have been classified based on their length into microRNAs (miRNAs), PIWI-interacting RNAs (piRNAs), small nucleolar RNAs (snoRNAs) and long non-coding RNAs (lncRNAs).

Additionally, Transposable Elements (TEs) can be included in this category of ncRNAs. Their existence was first reported during the 1940s by the American scientist Barbara McClintock who discovered that transposition was causing maize to change colour from yellow to brown on individual kernels[117]. Genomic sequencing has revealed that the genomes of prokaryotes and eukaryotes contain a variety of TEs because of insertional events that occurred during evolution. In humans, these elements make up almost half of the nuclear DNA and can be classified into two major classes: DNA transposons which are flanked by terminal inverted repeats, encode a transposase, and mobilize by a 'cut and paste' mechanism; and retrotransposons which mobilize by replicative mechanisms that require the reverse transcription of an RNA intermediate and use the cell's RNAPII for their transcription. The integration of these sequences into new sites create target site duplications and double-strand breaks, leading to the activation of DNA repair mechanisms of the host cells to repair and fill gaps.

TE activity is well known to be under epigenetic control and wide-spread TE expression has been found in cancers with particularly extensive epigenetic dysregulation. A striking enrichment of demethylation at CpGs within TEs, as compared with the background demethylation level, was detected across ten TCGA cancer types suggesting that a greater loss of DNA methylation at TE regions may be a common tumour process. In addition, aberrant expression of TEs has been associated with the expression of host immune genes and activation of DDR pathways, rendering TE activation as a marker for DNA Methyltransferase 1 (DNMT1) inhibition[118].

1.3.4 The DNA Methylomes of Cancer

Until recently, most cancer studies have focused on DNA methylation gains at promoter regions which is a common mechanism for the inactivation of TSGs in all cancer types[119,120]. However, recent genome-wide DNA methylation studies, mostly in the context of the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA), have brought to light another major and common phenomenon in cancer which is a global decrease of CpG methylation. This observation was made already in 1973 by Feinberg and Vogelstein[121], and later by Gama-Sosa *et al*., who described a global reduction of 5mC content in tumour samples[122].

From these initial studies, further research has shown that DNA methylation changes are more complex than initially thought and are often related to genomic instability, particularly by demethylation of repetitive genomic elements and TEs and, less frequently, to activation of silenced OCGs[123]. Nevertheless, global hypomethylation has also been detected in cancers with rather stable genomes such as chronic lymphocytic leukemia[124]. Hence, extensive hypomethylation does not lead to genomic instability *per se*. Furthermore, the loss of DNA methylation resembles to what has been observed in aging processes where embryonic stem cells have the highest level of DNA methylation followed by primary cells in contrast to cancer cells, where a sharp decrease in DNA methylation was observed globally[125]. Secondly, although DNA methylation at CGIs seems to ensure a repressive chromatin environment its absence is not necessarily associated with activation of gene expression[126] and other mechanisms such as trimethylation of H3K27 (H3K27me3) by the polycomb-repressive complex seems sufficient to repress expression in the absence of DNA methylation[127]. Additionally, DNA methylation has been shown to play a role in the expression of alternative transcripts by methylating alternative promoter sequences[128]. Thirdly, other studies have revealed that DNA methylation frequently occurs at genes that were already silenced in normal cells[129,130], for example via H3K27me3[131–133]. Thus, it has been suggested that DNA methylation at CGIs could enable stable gene inactivation but more a consequence rather than a cause of gene repression.

Importantly DNA methylation at low CpG density regions outside promoters, specifically at gene bodies and intergenic regions, is another major finding in cancer. DNA

methylation at these regions has been associated with both activation and repression of transcription. For instance, at intragenic regions it can regulate the expression of both an alternative transcript and the regular transcript simultaneously with opposing (activating/repressive) effects[124]. Furthermore, DNA methylation has been associated with alternative splicing and polyadenylation[134], for which DNA methylation at specific sites can lead to exon skipping or incorporation, the usage of different polyadenylation sites and the production of transcripts of different lengths.

Nevertheless, many CpGs are not associated with alternative promoters, exons, polyadenylation sites, ncRNAs or TEs[135], suggesting that DNA methylation might influence transcription via other mechanisms. Several studies have highlighted the important role that DNA methylation may play at both intragenic and intergenic enhancer regions. For instance, a stronger correlation between expression was observed for enhancer regions as compared with promoters[124]. Regulatory elements outside promoters enriched for cell-type-specific TF binding sites are differentially methylated during development and between normal and tumour samples. Other regions outside promoters that are differentially methylated in cancer include polycomb-repressed regions, which normally affect already silenced genes but may favour malignant transformation by blocking the ability to reactivate genes. Heterochromatin is also prone to lose DNA methylation in cancer probably as a result of passive DNA methylation loss upon replication[136,137].

1.3.5 Somatic Cancer Mutations to Chromatin-related Proteins

A unified model of cancer has been proposed based on mechanisms involving both genetic mutations and epigenetic modifications, for which both can disrupt the function of genes involved in the regulation of the epigenome itself[138–142]. Supporting this idea, several epigenetic modifier genes have been found mutated in human cancers and provide a mechanism linking mutations to epigenetic alterations. The classes of genes include histone variants (direct substitution of a mutant histone isoform); DNMTs, HATs, HDACs, HDMs; and chromatin remodelling factors previously mentioned. A few examples of epigenetic genes that have been found altered in LC include *CREBBP*, which encodes the CREB Binding Protein (CBP) and was found mutated in 5.3% of the cases[143]; covalent histone modifiers and subunits

of the SWI/SNF complex that were found mutated in 40 and 22.2% in pulmonary carcinoids, recurrently affecting *MEN1*, *PSIP1*, and *ARID1A*, as well as *EIF1AZ*, *BREBBP*, *EP300*, *MLL*, *CHD7* and *MYCL* with varying prevalence among LNETs[47,48,144–148].

1.4 Translational Research in Cancer

1.4.1 Onco-omics Applications

The word "omics" refers to several molecular disciplines that use high-throughput technologies to characterise and quantify very large sets of biological molecules such as DNA (genomics), mRNAs (transcriptomics), proteins (proteomics), and metabolites (metabolomics). These different omics provide tools that have been rapidly exploited, especially in the field of oncology, and have revolutionized our understanding of human biological systems. Not only have they facilitated the study of multiple samples at once at high resolution, but they also have added layers of molecular complexity. Thus, cancers have evolved as highly heterogeneous diseases in which different layers of regulation are strongly interconnected. In addition, these omics technologies have opened a new era of "personalised medicine", in which patients are treated based on specific molecular alterations leading to a paradigm shift in patient care.

Importantly, omics have allowed the identification of biological markers or biomarkers as measurable and objective indicators of normal biological processes, pathogenic processes, or pharmacologic responses to therapeutic intervention at a given moment. These biomarkers can have molecular, histologic, radiographic, or physiological characteristics, and can be used for many applications in health care. For instance, biomarkers can be used for the detection, prevention, determination of individual disease risk, disease monitoring and therapeutic stratification, and thus have emerged as useful tools in clinical trials thereby enabling a more personalised healthcare system and improved disease outcomes.

1.4.2 Genomics and Epigenomics: On the Road of Translation into Clinical Practice

Since the completion of the Human Genome Project in 2003, much of the research in oncology has focused on the sequencing of cancer genomes with the aim to identify actionable alterations at the base of oncogenesis (Fig. 1.7). This is because of the feasibility of being able to compare tumour sequences against a reference human genome. Furthermore, the development of NGS techniques and its commercialization since 2006 have led to a "genomic era" of cancer research that has facilitated the detection of somatic and germline mutations, TMB and resistance mechanisms. Following these discoveries, cancer treatment has parallelly evolved from generic cytotoxic compounds, classically used to target every proliferating cell, to targeted therapies directed towards particular genetic alterations that have been found to be driving tumorigenesis. Importantly, LC patients' outcomes have been improved with the development of the EGFR inhibitors (erlotinib, gefitinib), the PI3K/AKT/mTOR inhibitors everolimus, and the NTRK/ROS1 inhibitors entrectinib. In addition several drugs are being investigated in clinical trials to overcome drug resistance, such as third-generation EGFR-TKIs (Osimertinib) targeting the frequent T790M resistant mutation[149].

Thanks to the development of sequencing technologies, new NGS techniques enable all the coding sequences of the genome (Whole Exome Sequencing, WES) and even the full genome (Whole Genome Sequencing, WGS) to be sequenced in a fairly quick and affordable way. Hence, the challenge now lies in the data analysis step, gradually up to an ever-challenging framework.

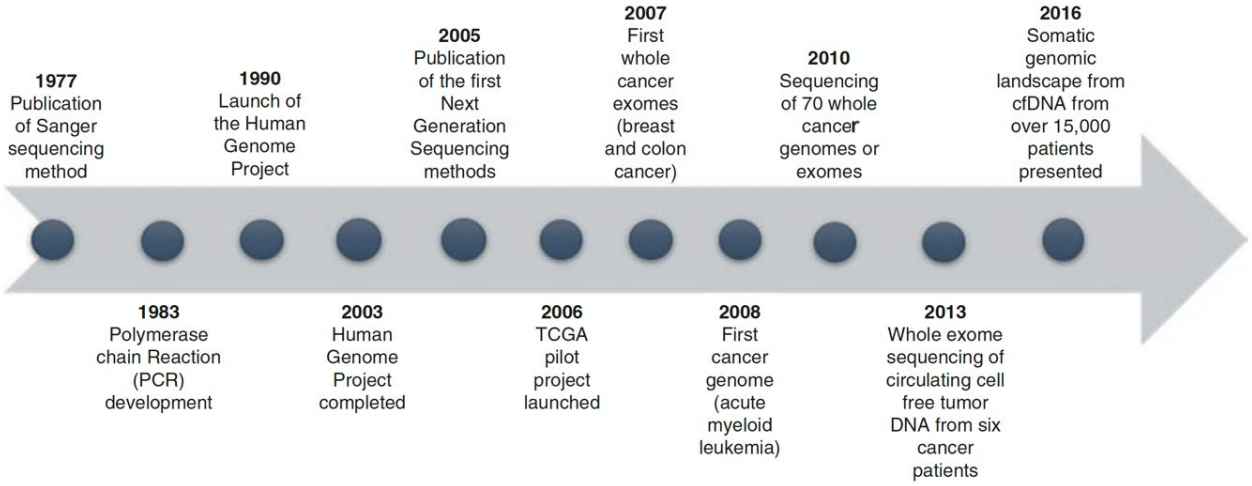


Figure 1. 7| Chronology of DNA sequencing development in relation to cancer. [Taken from: Flores-Pérez, J. A., De La, F., Oliva, R., Argenes, Y. & Meneses-Garcia, A. Translational Research and Onco-Omics Applications in the Era of Cancer Personal Genomics. *Advances in Experimental Medicine and Biology* vol. 1168 (2019)][150].

Adding another layer of complexity, biological pathways and cellular processes are under the impact of the epigenome status and recent findings have shown that drug responses could be largely impacted by the presence of aberrant epigenetic marks. Therefore, the discovery of several epigenetic alterations affecting key genes and cellular pathways involved in several tumour types has opened the "epigenomics era", in which novel therapeutic strategies are now also aimed at reversing epigenetic marks in cancer cells. This layer of gene regulation that directly influences phenotype provides an alternative therapeutic approach especially for individuals with "high risk" genotypes.

Epigenetic modifications are also attractive targets for the development of new therapies against cancer, mostly due to their reversible character. Current approved Food and Drug Administration (FDA) treatments involve inhibitors of epigenetic enzymes, and the most extensively studied group include DNMT inhibitors (DNMTi)[151,152]. In addition, small interfering RNAs (siRNAs) are considered as the new generation of biodrugs due to their specific and efficient response in RNA interference (RNAi)[153] and because they can be used to target specific epigenetic regulators.

1.5 Hypotheses

Further research is imperatively needed for a better understanding of molecular events of onset and progression of lung cancer as these are likely to be key and important for optimizing treatment strategies. To date there are only a few examples of studies for common and rarer types of lung cancer where an integrative analysis to investigate the relationship of genome to epigenome has been performed. The hypotheses of this thesis therefore are that:

- Genetic alterations and changes in DNA methylation will distinguish different LC subtypes and tumours from healthy tissue.
- The integration of genetic and DNA methylation data with clinical and gene expression data will enable improvement in the molecular classification and therapy selection for LC patients.

1.6 Thesis Objectives

This PhD project aimed to investigate the alterations in the genome and methylome in different lung cancer subtypes. To achieve this, the objectives of the study were:

(1) to identify the molecular alterations using WES, TCS and SNP genotyping

(2) to identify differentially methylated regions by using WGBS

(3) to relate the relevant genetic and methylation alterations with clinical parameters and gene expression data in different lung cancer subtypes.

The ultimate objective was to gain a more complete understanding of how genetic and epigenetic alterations may interact to drive tumorigenesis in different LC histotypes to improve patient stratification and identify novel potential targets.

Chapter 2: Methodology

2.1 Patients and Clinical Samples

Clinical research samples for the study were provided by the Biomedical Research Unit (BRU) biobank of the Royal Brompton and Harefield NHS Trust. All samples had been collected under the appropriate ethical approvals (RBH NIHR BRU Advanced Lung Disease Biobank [NRES reference 10/H0504/9] and Brompton and Harefield NHS Trust Diagnostic Tissue Bank [NRES reference 10/H0504/29]). Clinical samples consisted of fresh-frozen human lung tumour specimens and normal paired lung tissue obtained during lung resection. The tissue samples were collected prior to therapy.

Samples for genomics (DNA) were snap frozen and stored at -80^{o}C whilst those for transcriptomics (RNA) were stored at -80^{o}C in RNAlater within two hours of surgical excision. Nucleic acid extraction of the samples had been performed by Dr. E. Starren (former member of the Genomic Medicine Group, NHLI) prior to commencement of this present study. Briefly, genomic DNA was extracted using a phenol chloroform method plus a TissueRuptor to ensure tissue disruption. Tissue stored in RNAlater underwent extraction for total RNA using a Qiagen RNEasy Fibrous Midi Kit[154]. Tumour histology and cell abundance was determined through pathology review (Professor Andrew Nicholson) of haematoxylin and eosin staining. Clinical metadata associated with the research samples was collated by Dr. E. Starren.

2.2 Copy Number Aberration Peaks from SNP Genotyping

2.2.1 SNP Genotyping

Illumina Infinium OmniExpressExome (ver. 1.6) genotyping arrays (also referred to as single-nucleotide polymorphism [SNP] arrays) were used for genotyping 958,497 SNP markers in the genomic DNA extracted from tumour and normal matched lung samples. The Illumina SNP arrays (chips) combine DNA hybridization, fluorescence microscopy and a solid surface DNA capture (array) to allow comparison of thousands of genomic locations simultaneously. SNP data was generated by Eurofins.

The Illumina SNP array platform targets biallelic SNPs with two hybridization probes to specifically capture each of the two alleles, latter referred to as alleles A and B. DNA to be interrogated is taken, labelled using a fluorescent dye and subsequently hybridised to the probes arrayed on the chip. Only the target sequences created from the tested samples that are complementary to the probes bind. SNP genotypes are therefore determined based on the ratios of the hybridization intensities for A and B probes. Because SNP alleles only differ in one nucleotide and due to the technical challenge to achieve optimal hybridization conditions for all probes on the array, several redundant probes are used to interrogate each SNP to improve accuracy of genotype calls. After hybridisation, a scanner is used to measure fluorescence intensity of hybridized A and B probes for each SNP on the array, thus representing the signal strength for each allele. These output data are referred to as the raw intensities of the A and B alleles (R_A and R_B, respectively). Each SNP has an expected raw intensity (R) for a given cluster (sum of red and green signal intensities), thus SNP genotypes are determined by comparing A and B intensities to the reference. As a result, a heterozygous locus (AB) shows the same intensity for both alleles, whereas a homozygous locus (AA or BB) intensity is only seen for the one type of allele present (A or B).

The intensity data received from Eurofins was processed by downstream software to enable calling of Copy Number Alterations (CNAs), as explained in the following sections.

2.2.2 Summary of SNP Genotyping Data Analysis

In outline, SNP genotyping data analysis consisted of SNP clustering, SNP data Quality Control (QC), GC bias correction, Copy Number (CN) segmentation and calling of significant Copy Number Alterations (CNAs). Since SNP genotyping data was generated for both tumour and normal matched samples, calling of CNAs was first performed on the unaffected samples and significant copy number gains or losses were subsequently subtracted from the tumour segmented data before calling of somatic CNAs on the remaining tumour segments. A summary flowchart for the SNP genotyping data analysis is shown in Figure 2.1

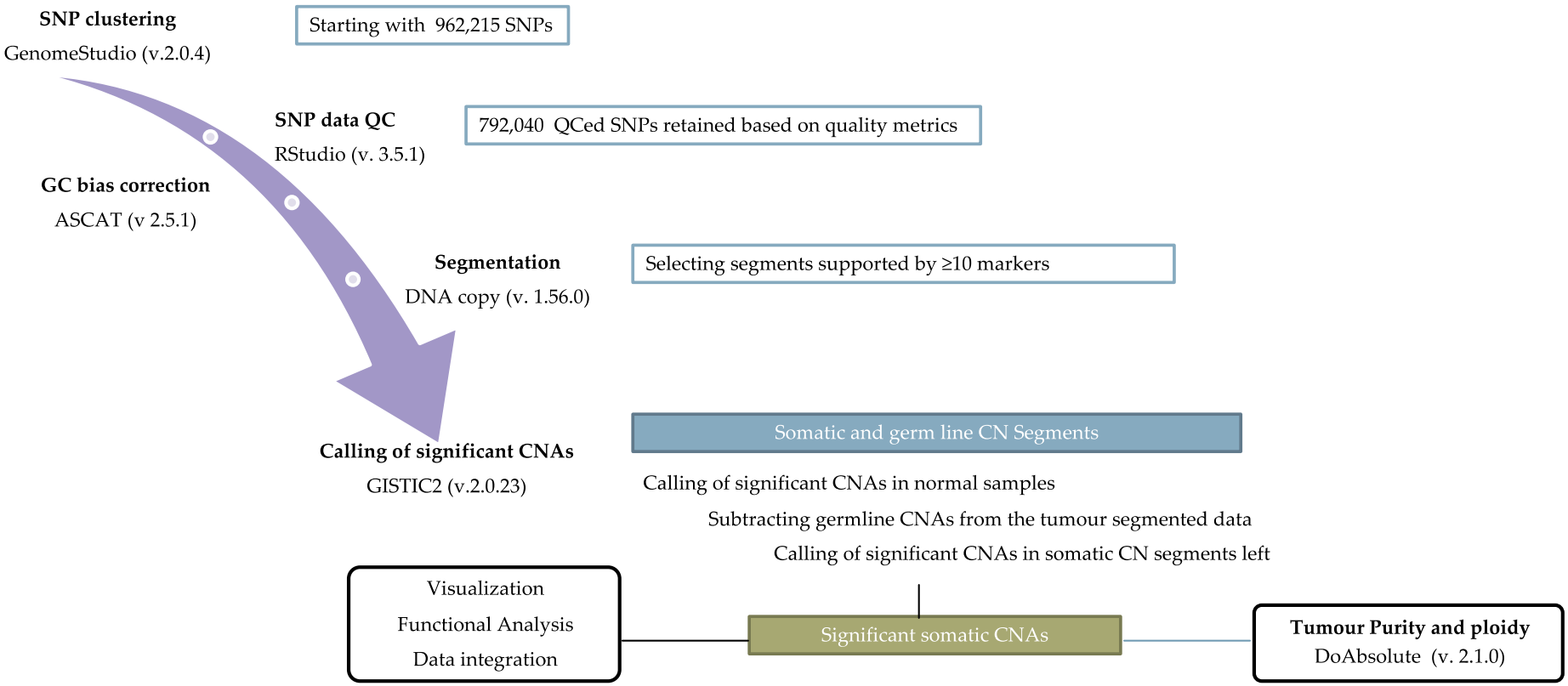


Figure 2. 1| Flowchart of SNP genotyping data analysis. SNP genotyping data analysis consisted of SNP clustering, SNP data Quality Control (QC), GC bias correction, Segmentation and selection of segments with> 10 markers and calling of significant Copy Number Alterations (CNAs). Since SNP genotyping data was generated for both tumour and normal matched samples, calling of CNAs was first performed on the unaffected samples and significant copy number gains or losses were subsequently subtracted from the tumour segmented data before calling of somatic CNAs on the remaining tumour segments.

2.2.3 GenomeStudio Data Pre-processing

The processing of Illumina genotyping arrays was performed with the GenomeStudio software (ver. 2.0.4). This software enabled conversion of raw intensities into genotype calls, as well as optimizing call rates, generating SNP statistics and confirming gender of the samples.

Loading of sample intensities was performed by selecting directories with intensity files. In addition, two additional files for annotation purposes were downloaded from the support Illumina website (https://support.illumina.com), specifically the *Infinium OmniExpressExome-8 v1.6 Cluster File* and the manifest file in Bitmap Image File (BMP) format (GRCh37/ hg19). Next clustering of intensities for all SNPs was performed by applying a cluster algorithm to the fluorescent levels to form clusters that distinguished samples into AA, AB and AB clusters. After this sample and SNP quality was evaluated.

Assessment of sample quality was performed using the GenomeStudio software and was based on three different scores:

- Call rate, as the percentage of SNPs with genotype calls for a sample. A call rate of 95-98% was used, and any sample below the call rate was excluded from further analysis.
- GenCall (GC) score cut-off or no-call threshold, a quality metric calculated for each genotype ranging from 0 to 1 to filter poor quality SNPs or samples. Genotypes with GC score less than a given threshold were declared as missing due to being considered too far from the cluster to make reliable genotype calls.

No-call rate, indicating the proportion of missing values, or 'no calls' in each sample. A 0.15 no-call threshold as the standard for Infinium data.

Two different files were obtained from GenomeStudio that were then used for further analysis:

- SNP table: contained chromosome, position, AB frequency, SNP name, number of no calls, minor frequency, call frequency, GenTrain and cluster separation scores. The call frequency or call rate, was calculated as *Calls/(No calls + Calls)*; the GenTrain score, as the quality score for samples clustered for a locus, and the Cluster separation score, measuring how well-separated a cluster was from other clusters.

- Final Report: containing sample ID, SNP name, chromosome, position, allele1 forward, allele2 forward, and two derived measures that allowed estimation of copy number status: (1) the *Log R ratio (LRR)* is the $\log_2$-transformed value of the intensity of the SNP, $(R_A+R_B)/R_{expected}$, where $R_{expected}$ is an interpolation generated by GenomeStudio. Hence the LRR was the metric used to normalize signal intensities. (2) The *B allele frequency* (BAF), reflecting the proportion of hybridized sample that carried the B allele. This metric is an adjusted value generated by GenomeStudio that normalized genotyping calls assuming three canonical clusters (A/A: 0.0, A/B: 0.5, B/B: 1).

2.2.4 QC of SNP Array Data

Quality Control[155] of SNP genotyping data was performed in R Studio (ver. 3.6.1) using tidyverse[156] R package (ver. 1.3.0). SNPs were filtered out if they met any of the following criteria:

- Were classified as InDel markers [I/D]
- Chromosome *XY* SNPs or SNPs in pseudoautosomic regions
- Obtained a GenTrain score of <0.8
- Obtained a Cluster Separation of <0.8
- Obtained a Call Frequency of <0.8
- Obtained -0.5 < Heterozygosity Excess> 0.5

After that, a Perl script was used to keep only one SNP for those that shared genomic coordinates. As a result, only SNPs with unique genomic locations were retained for further processing.

2.2.5 GC Correction

Several high-density SNP genotyping arrays exhibit a spatial correlation or waviness in the intensity signal detected that can prevent the accurate detection of copy number variation. This waviness in the signal has been recently confirmed to be associated with the amount of input DNA and GC content of the probes. For instance, probes with high GC content bind better to their target sequence hence they will show higher signal intensity and LRR metrics,

leading to biased copy number detection. Moreover, this waviness has been detected independently of the platform used to determine DNA copy number status[157].

Therefore, SNP genotyping data was adjusted for genomic waves with the *ascat.GCcorrect* function implemented by the Allele-Specific Copy number Analysis of Tumours (ASCAT)[158] R package (ver. 2.5.1). The function requires a GC content file specific to the platform used for genotyping. For GC wave correction of the lung SNP dataset, a custom GC content file was therefore created based on the location of the genomic probes and chromosome size based on the hg19 human reference genome.

LRR values were thus corrected for biases in GC percentage around each SNP marker across tumour and non-tumour samples separately. In addition, ASCAT plots for LRR and BAF values were generated and were used to visualize noisy samples that needed to be excluded from further analysis.

2.2.6 Data Segmentation

The next step involved partition of the genome into genomic regions of equal copy number with the DNAcopy[159] R package (ver. 1.56.0). The segmentation process employed the Circular Binary Segmentation (CBS) algorithm to segment DNA copy number data and involved the following steps:

I. The creation of a CNA object from the GC corrected LRR values with the *CNA* function.
II. Smoothing outlier LRR values with the *smooth.cna* function.
III. Segmentation of the smoothed GC corrected LRR data with the *segment* function.

Two arguments were used to refine the segmentation process. First, a minimum number of 2 and 3 SNP markers were chosen for a segment to be a valid segment for the tumour and normal CN data respectively. Since tumour samples were expected to present more genomic instability, a lesser number of markers would represent a valid segment whereas normal samples are expected a more stable genome hence, a higher number of markers was chosen. Taking into account the size of the hg19 genome of 3,095,677,412 base pairs[160,161] (including X and Y chromosomes) and that the input CN data at this stage contained

751,546 markers, it was estimated that an average of ~4.1Kb of the genome was represented by each SNP. In addition, the average size of a human gene is 10-15Kb[162]; therefore a threshold in tumour samples of 2 markers minimum would cover a stretch of ~8.2Kb and account for sub-genic/ exon-level CN variation expected in tumour genomes. While in non-tumour samples, where less CN variation is expected, a threshold of 3 markers minimum (~12.3Kb; approximated gene size) would be suitable.

Furthermore, assuming that CN data showed a normal distribution, segments were selected by establishing standard deviation thresholds (specified by the "sdundo" argument in the *undo.splits* option). In this way the difference in signal intensity between a segment and the neighbouring region was restricted by specifying the number of Standard Deviations (SD) to be qualified as a valid segment for the tumour and normal CN data respectively. Or in other words, how different a segment's signal intensity had to be from the mean signal intensity from neighbouring regions in each sample. Following the quantile function, a segment LRR had to be different by 80% (or 1.28 SD) of the segment LRR values in tumours because more segments with variable LRR values were expected in tumours. In the case of non-tumour samples, a segment LRR had to be different by 95% (or 1.96 SD) of the LRR values. The parameters used for CN segmentation for the tumour and normal samples therefore were:

- Tumour CN segmentation: undo.splits = "sdundo", undo.SD = 1.28, min.width = 2
- Normal CN segmentation: undo.splits = "sdundo", undo.SD = 1.96, min.width = 3

Finally, only segments that obtained a minimum support of 10 markers were selected for calling CNAs (see Section 2.2.7 below).

DoAbsolute (ver.2.1.0) was used to infer tumour purity and ploidy with default parameters[163,164]. Mutational data was used as recommended for estimation of tumour purity by serving as an alternative point of reference regarding tumour progression (sub-clonal events) that allow a more comprehensive modelling of tumour heterogeneity. In addition, computation of tumour purity and ploidy was performed separately using each histotype as primary disease for specific tumour karyotype matching.

2.2.7 Focal Copy-Number Alteration Calling with Gistic2

Calling of Copy Number Alterations was performed with the Genomic Identification of Significant Targets In Cancer (Gistic2)[165] software (ver.2.0). The Gistic2 module identified genomic regions that were recurrently amplified or deleted in a group of samples. Similar to the segmentation step (Section 2.2.6 above), calling of CNAs was run separately for tumour and normal samples due to biological reasons. Gistic2 was run with a Copy Number Variant (CNV) file specifying germ line CNVs for these to be excluded from the analysis. The CNV file specified the genomic location on the reference hg19 genome. The Gistic2 software was run with the below options:

- geneGistic2 1 \ Flag indicating that the gene Gistic2 algorithm should be used to calculate the significance of deletions at a gene level instead of a marker level.
- broad 1 \ Flag indicating that an additional broad-level analysis should be performed.
- brlen 0.5 \ Threshold used to distinguish broad from focal events, given in units of fraction of chromosome arm.
- conf 0.90 \ Confidence level used to calculate the region containing a driver.
- armpeel 1 \ Flag to assign all events in the same chromosome arm of the same sample to a single peak when peaks are split by noise of chromothripsis.
- savegene 1 \ Flag indicating to save gene tables.
- rx 0 \ Flag to not remove X and Y chromosomes.
- gcm extreme # Method for reducing marker-level copy number data to whichever of min or max is furthest from diploid.

The output from Gistic2 consists of a list of aberrant regions at the CN level or "peaks" with an assigned G-score that considers the amplitude and the frequency of occurrence across samples. In addition, False Discovery Rate (FDR) q-values were calculated for each peak. Regions with q-values below 0.25 were considered significant. After running Gistic2, the genomic regions with CN variation, a genomic descriptor of each significant region, q-values and residual q-values were obtained. Concretely, three different genomic regions were obtained:

- Wide peak limits: consist in boundaries most likely to contain the target genes.

- Peak limits: delimited part of the aberrant region with greatest amplitude and frequency of alteration.
- Region limits: boundaries of the entire significant region harbouring CNA.

Two significance measures were also obtained:

- Q-values: the q-value of the peak region.
- Residual q-values: adjusted q-value of the peak region after removing adjacent amplifications or deletions that overlap other significant peak regions in the same chromosome.

2.2.8 Subtraction of Recurrent CN Segments from Germline Samples

Next segments with a sharp signal in control samples were subtracted prior to running Gistic2 on the tumour data to select putative somatic CNAs. For this, the wide peak limits that obtained a residual q-value of $\leq 10^{-5}$ from the normal samples were subtracted from the tumour segmented data employing the *coverageBed* function from bedtools (ver.2.29.0). Tumour segments that overlapped by 50% with the normal wide peak limits were excluded and were used for calling of CNAs as detailed in Section 2.2.9.

2.2.9 Calling of Somatic Focal Copy-Number Alteration (CNA) with Gistic2

After germline CN subtraction, the remaining segments considered "somatic" segments were subsequently used to identify significant CNAs. Gistic2 software (ver.2.0.23) was run with the beforementioned options used to identify recurrent CNAs in non-tumour samples.

2.2.10 Maftools and Downstream Interpretation of Somatic CNAs

Output files generated by the Gistic2 programme were used as input for Maftools (ver. 2.2.10) for summarising CN data and for visualization purposes. For instance, Maftools allowed the type of CNA (amplification or deletion) per sample to be retrieved as well as the cytoband,

wide peak limit and associated q-value. The genes contained or located nearby the significant peaks are also detected.

2.2.11 Copy Number Burden (CNB) Calculation

Copy Number Burden was calculated per sample based on the segment size of amplifications and deletions per autosome size. Calculation was performed with RStudio software.

2.3 Targeted Gene Panel Sequencing

2.3.1 Agilent Gene Panel Design

For NSCLC, an Agilent Gene Panel was designed based on published literature[166–168] and findings from prior in-house whole exome sequencing of a set of 70 paired tumour and normal tissue NSCLC samples. This set included cases of adenocarcinoma, squamous as well as carcinoids. The panel was designed using the Agilent software SureSelect DNA Advanced Design Wizard based on the Human Genome version from February 2009 assembly (GRCh37/ hg19). The gene panel focused on the exonic regions of 52 genes (Table 2.1) that have been found recurrently mutated in NSCLC, such as *TP53*, *EGFR*, *PIK3CA*.

AKT1	*EP300*	*MTAP*	*RIT1*
ALK	*ERBB2*	*MYC*	*ROBO1*
ARID1A	*ERBB4*	*MYCL*	*ROS1*
ARID1B	*FBBXW7*	*NF1*	*SETD2*
ARID2	*FGFR3*	*NFE2L2*	*SF3B1*
BRAF	*FHIT*	*NOTCH1*	*SMARCA4*
CCND1	*FOXP1*	*NRAS*	*SOX2*
CCND3	*HRAS*	*NTRK1*	*STK11*
CDK4	*KEAP1*	*PIK3CA*	*TERT*
CDKN2A	*KRAS*	*PTEN*	*TP53*
CREBBP	*MAP2K1*	*RB1*	*TSC1*
CUL3	*MDM2*	*RBM10*	*TSC2*
EGFR	*MET*	*RET*	*U2AF1*

Table 2. 1| Genes contained in the Agilent Gene Panel for targeted capture sequencing. The gene panel consists of 12,129 probes with a total size of 266,937 Kb.

2.3.2 SureSelect NGS Target Enrichment for Illumina Multiplexed Sequencing

Library preparation was done according to the Sure Select QXT target enrichment system (Agilent Technologies) for the Illumina Multiplexed Sequencing platform (Illumina) following manufacturer's instructions (Fig. 2.2).

Genomic DNA samples were adjusted to 25 ng/µl for enzymatic fragmentation and adaptor-tagging (process called tagmentation). The adaptor-tagged DNA libraries were PCR-amplified, AMPure beads purified and adjusted to 750 ng in 12 µl using nuclease-free water. Prepared libraries were hybridized to the Capture Library and selected using streptavidin-coated beads.

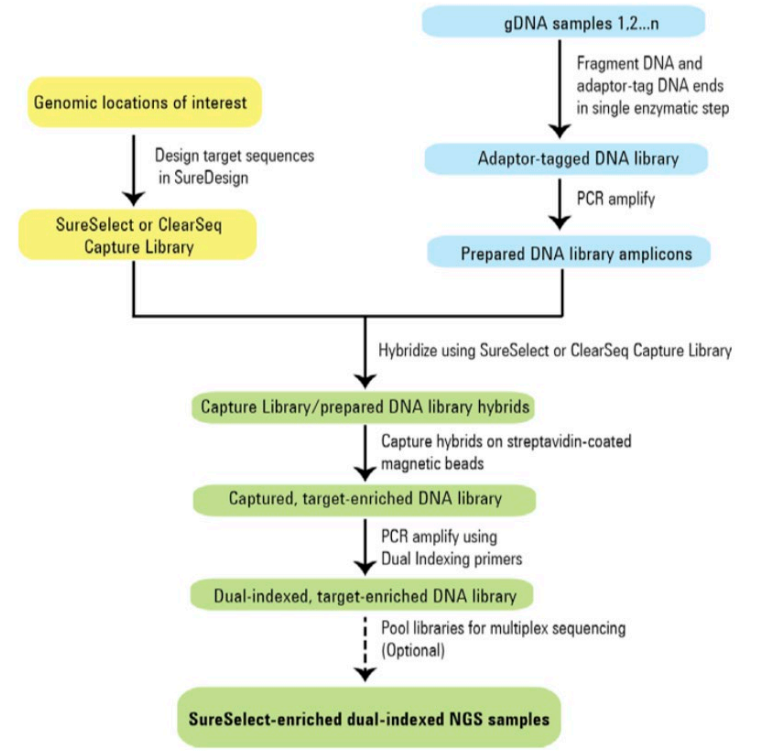


Figure 2. 2| SureSelect NGS Target Enrichment sample preparation workflow for the Illumina Multiplexed Sequencing platform protocol. [Taken from: Version, D. SureSelectXT Target Enrichment System for the Illumina Platform Protocol. 1–102 https://www.agilent.com/cs/library/usermanuals/Public/G7530-90000.pdf (2021)][169].

The targeted captured libraries were PCR amplified using the appropriate distinct pair of dual indexing primers to allow further multiplexing. Amplified captured libraries were

purified with AMPure XP beads (Agilent) and combined in equimolar amounts. Target-enriched libraries were diluted to the optimal seeding concentration of 1.8 pM together with PhiX control that was used as a sequencing quality control.

Both pre and post cleaned-up libraries were analysed with Agilent 2100 Bioanalyzer and High Sensitivity DNA assays, following the manufacturer's instructions, to determine the DNA fragment size and concentration of each prepared library. Pre-capture libraries had a DNA fragment size between 245 to 325 bp, while post-capture libraries had a DNA fragment size between 325-450 bp.

2.3.3 Illumina NextSeq Sequencing

Libraries were sequenced with paired-end chemistry on the Illumina NextSeq 550 (Illumina) Next Generation Sequencing (NGS) automated sequencer at the Clinical Genetics & Genomics Laboratory, Royal Brompton Hospital. The Illumina sequencing instrument uses a sequencing by synthesis technology that allows to measure the intensity of each base in each sequencing run. Binary Base Call (BCL) files are the raw data generated by the Illumina sequences that store the base and the quality score of each called base. These BCL files were converted to FASTQ files and simultaneously demultiplexed with the Bcl2fastq2 Conversion Software (v2.19). Fastq data thus obtained were quality controlled and processed for alignment, mutation calling and variant annotation, according to an established pipeline within the Genomic Medicine group detailed in the sections below.

2.3.4 Summary of DNA Sequencing Data Analysis

Analysis of DNA sequencing data generated from Targeted Capture Sequencing (TCS) involved QC of raw reads, trimming of low-quality ends, alignment or mapping against the human reference genome, recalibration/refinement of the reads, somatic mutation calling. Calling of somatic SNPs and Indels was performed by the analysis of matched tumour-normal samples, and high confidence calls were subsequently used for variant annotation (Fig. 2.3).

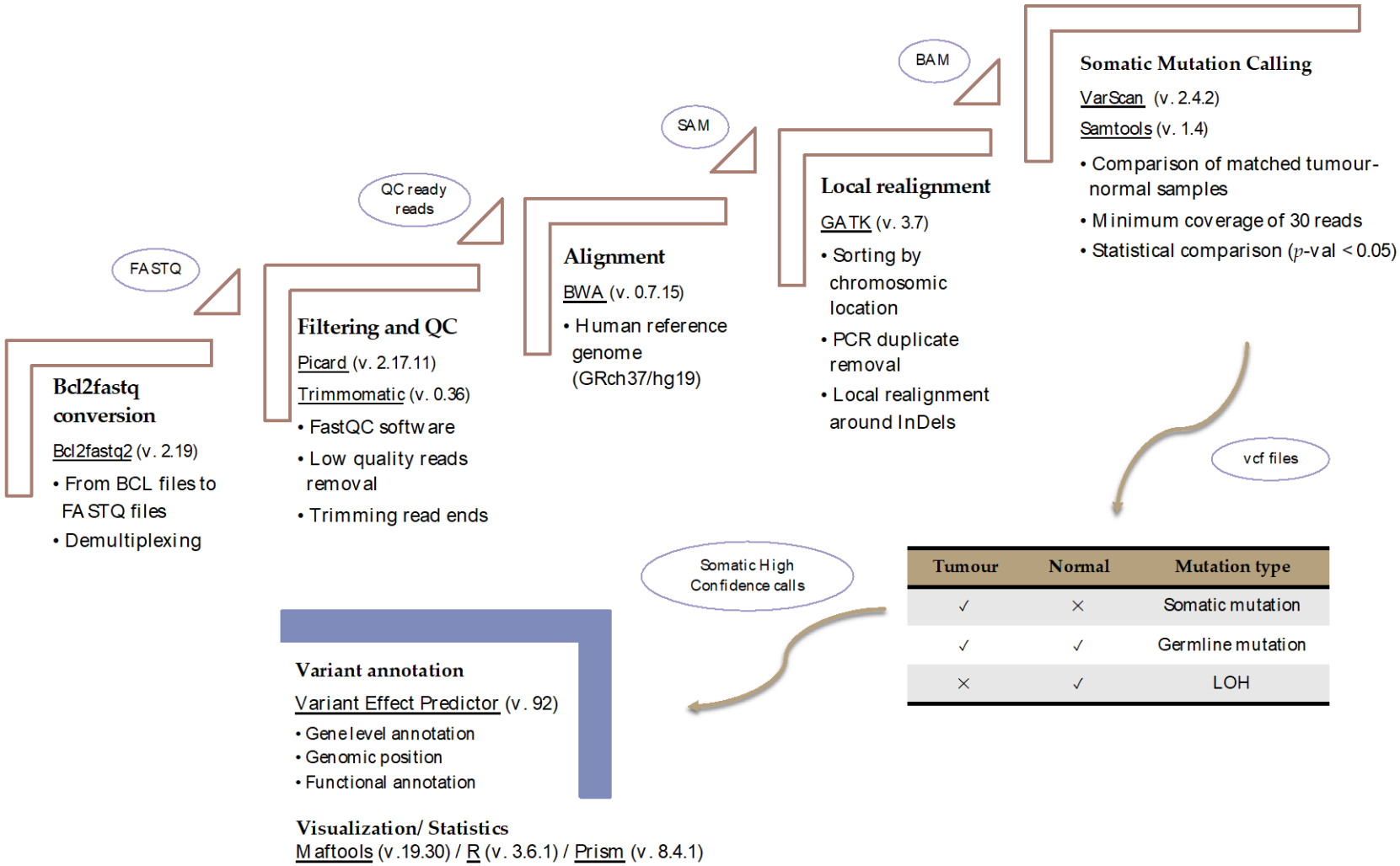


Figure 2. 3| Flowchart of the mutational analysis for the Lung Cancer cohort. Mutational data analysis involved Bcl2fastq conversions, QC of raw reads, trimming of low-quality ends, alignment against the human reference genome, local realignment and recalibration/refinement of the reads and somatic mutation calling by analysis of paired tumour and normal samples. Finally somatic high confidence calls were used for variant annotation and downstream analyses.

2.3.5 QC of Fastq Reads

Paired-end Illumina NextSeq data consists of the sequencing of both ends of the same fragment. Therefore, the sequences (or reads) come in two data files corresponding to the forward paired-end sequence (R1) and the reverse paired-end sequence (R2). The FastQC software provides assessment of the quality of the sequenced bases for both reads in each sample. FastQC was therefore run before and after adapter trimming of paired-end reads with Trimmomatic (ver. 0.36). FastQC reports were examined and Phred scores were used to exclude low quality reads.

2.3.6 Mapping and Somatic Variant Calling

The Human Genome December 2013 assembly (GRCh37/hg19) was downloaded from the University of California Santa Cruz (UCSC) genome database. Alignment of the trimmed Fastq sequences against the human reference genome was performed using BWA mem (ver. 0.7.15) to obtain Sequence Alignment Mapping (SAM) files. SAM files thus obtained per sample were sorted by chromosomal location and read alignments deemed to be PCR duplicates were removed with Picard (ver. 2.17.11) to obtain BAM files (the binary version of a SAM file). Then, Genome Analysis Tool Kit (GATK) (ver. 3.7) allowed local realignment around known Insertions and Deletions (InDels), from 1000 Genomes Project (https://www.internationalgenome.org), to correct mapping errors that could have been generated by the genome aligner. Further, DepthOfCoverage from GATK was used to obtain main coverage across the total regions covered by the Gene Panel.

Next, somatic mutation calling was performed, employing VarScan (ver. 2.4.2), through comparison of matched tumour and normal samples. The VarsCan tool expects both a normal and tumour file in SAMtools pileup format and position sorted. Samtools (ver. 1.4) *mpileup* command was used for summarising the base calls of the BAM aligned reads to the reference sequence. Then, the output mpileup files for the tumour and normal samples were used to run VarScan Somatic. Specifically, a statistical comparison for changes in variant allele frequency is made at every genomic position where both normal and tumour meet the following criteria:

- a minimum of 10 supporting reads [--min-reads2]

- a minimum base quality of 20 for Phred-scaled base qualities [--min-avg-qual]
- a minimum depth of coverage of 30X [--min-coverage]
- a minimum allele frequency of 0.01 [--min-var-freq]
- Fisher's Exact Test *P*-value of <0.05 [--p-value]

Furthermore, if a difference in variant allele frequency between tumour and normal was significant and the base observed in the tumour did not match the normal, a somatic variant was called. In contrast, if the difference was significant but the base observed in the normal pairing did not match the reference, a germline variant was called. Variant Calling with VarScan allowed the Variant Call Format (VCF) files for both SNPs and InDels to be obtained that were next used as input for gene annotation.

2.3.7 Variant Annotation

Gene annotation aims to gather data around the raw DNA sequence in order to determine the functional effect of the called variants based on different types of genetic and genomic information. Gene annotation was performed with the Ensembl Variant Effect Predictor (VEP) (ver. 92) which provided additional information on genes affected by each variant (gene level annotation); location information at genomic, mRNA and protein level; and the consequence of the variant on the protein sequence (functional annotation). Moreover, custom Plugins were used to add functionalities and retrieve additional annotation data, for example pathogenicity and splicing predictions, or population frequency data. The plugins were the VEP plugin that reports Clinical Sequencing Nomenclature (CSN) for variants, the ExAC plugin to obtain allele frequencies, the database of splice-site consensus Single-Nucleotide Variants dbscSNV plugin for reporting splicing variants, and the Combined Annotation Dependent Depletion (CADD) plugin to obtain CADD scores for both Single-Nucleotide Variants (SNVs) and InDels.

For functional annotation, known impact predictions were obtained from sequence ontology analysis. This analysis classified variants into High (when leading to gain/ loss of stop-codon, frameshift variant or alteration at the splice acceptor/ donor site), Moderate (when leading to missense variant), Low (when leading to silent variant, or variant affecting

un-translated region of the mRNA, or those affecting non-coding genes) or Modifier (when they are intragenic or intronic) depending on the consequence of the variant on the protein sequence. Moreover, protein impact predictions from SIFT, polyPhen and CADD algorithms, and dbscSNV scores for functional predictions of SNVs within splicing consensus regions (scSNVs) were obtained. These scores indicated whether a variant was deleterious or benign. CADD scores compile information from 63 different annotation databases, including PhastCons, GERP, PhyloP, SIFT and PolyPhen, and range from 1 to 99, with a higher score indicating greater deleteriousness.

VEP also searched the Ensembl Variation database to retrieve known variants that overlapped with the input variants and their allele frequency (AF) from dbSNP, 1000 Genomes, NHBLI ESP, ExAC (Exome Aggregation Consortium) and gnomAD (Genome Aggregation Database) databases. Finally, the Catalogue of Somatic Mutations in Cancer (COSMIC) identifiers (IDs) from variants reported in various cancer genomes were also obtained with VEP.

2.3.8 Filtering Based on Impact and Population Frequency

After variant annotation, variants were filtered and prioritised to identify variants most likely to impact function. Filtering of the VEP annotated variants was performed taking into account population-level frequency, clinical impact and functional impact. Specifically, variants were selected through the following filtering process:

- Variant allele observed at ≥1% frequency in tumour
- Variant had no associated frequencies in dbSNP, 1000 Genomes, NHBLI ESP, ExAC and gnomAD databases or when, if the variant was known already in population-level databases, its observed incidence was <0.001
- At the functional impact level, only high and moderate impact variants were selected, or, the dbscSNV predicted score was >0.6 for variants in splicing regions.

In addition to the above, CADD score>=15 was used and correlated with SIFT and polyPhen prediction scores to predict potential protein-damaging effects of missense variants. Variants suspected to be artefacts were manually examined with the Integrated Genome Viewer (IGV) to discard potential false positives. Variants were filtered out if strand-bias was observed in

the distribution of the supporting reads for the variant allele, or if additional mismatches were observed, with respect to the reference in close proximity, to distinguish from sequencing noise. In addition, any SNV present +/- 3 bases from an Indel was presumed to be a mis-alignment.

2.3.9 Downstream Interpretation

Finally, all the variants (including high-allele fraction variants) were checked manually on Mutation Taster, COSMIC 3D, cBioPortal and in the literature to discard polymorphic or known benign variants. The COSMIC data portal was used to visualize if a detected variant had been previously reported in other cancers or mapped near cancer hotspots or in important functional protein domain following the rationale that variants with no additional reports in COSMIC or mapping into unknown domains have a lesser chance of being biologically relevant. Literature was also checked to identify mistakenly annotated genes i.e., a loss of function variant that was reported in the literature with an opposite oncogenic role. Lastly, expression data in tissue panels at NCBI Entrez and Gene Cards websites were reviewed in order to discard variants that were not known to be expressed in the lungs, as indicative that such variant may not be passing the RNA-protein bridge.

2.4 Somatic Mutations from WES

2.4.1 WES Sequencing

WES and mutation calling was performed in collaboration with Prof. Mark Lathrop, Dr. Markus Munter and the Canadian Centre for Computational Genomics (C3G) team at the McGill Genome Centre, Montreal, Canada. Sequencing libraries were prepared with the SureSelect^{XT} Target Enrichment System (Agilent SureSelect Human All Exon V4) and sequenced with 100 bp Paired-End Illumina HiSeq2000 Sequencer.

2.4.2 Callset Generation from McGill WES Data

This analysis was carried out by Dr. R. Eveleigh (McGill Genome Centre, Montreal, Canada) as follows:

Briefly, Fastq files were trimmed and aligned against the Human Genome December 2013 assembly (GRCh37/hg19) with BWA mem. SAM files thus obtained per sample were sorted by chromosomal location with GATK (ver.3.7) and read alignments deemed to be PCR duplicates were removed with Picard (ver. 2.9.0) to obtain BAM files. Samtools *Fixmate* was used to ensure that paired-end reads contained the correct information about the mate read, and the resulting BAM files were further processed to remove biases in the data through InDel realignment and Base Quality Score Recalibration (BQSR) from GATK. Finally, somatic mutations and InDels were identified with MuTect (ver. 1.16) and Scalpel (ver. 0.4.1) software, respectively. Somatic calls were then combined, and further steps involved decomposing multiallelic variants from VFC files, genetic variant annotation and functional effect prediction with SnpEff (ver. 4.3), and addition of metadata with Genome MINIng (GEMINI) (ver. 0.14-0.20) software.

2.4.3 Filtering of the Gemini Annotated Variants

Taking the annotated variants provided by McGill filtering based on known impact predictions and CADD scores was performed. Only variants with predicted high or medium impact, with CADD score>=15, and not flagged as polymorphic were selected for integration with TCS data.

2.5 Downstream Analysis of Merged Somatic Mutational Data from WES and TCS

2.5.1 Integration of WES and TCS Data

The selected somatic variants from WES and TCS experiments were merged using the command line and RStudio and were then used to generate Mutation Annotation Format

(MAF) files as input for Maftools. Maftools allowed the generation of several plots to summarise mutational data:

- Oncoplots or waterfall plots to summarise annotated variants per sample.
- MAF summary plots to display the number of variants and variant types as stacked bar plots and boxplot respectively.
- Boxplots showing overall distribution of substitution types, as well as stacked bar plots showing fraction of conversions in each sample.
- Lollipop plots for visual exploration and localisation of variants in different amino acid motifs and amino acid changes.
- Rainfall plots for visualisation of inter variant distance on a linear genomic scale to enable genomic co-localisation of variants to be detected.
- Comparison of Tumour Mutational Burden (TMB) against TCGA cohorts.
- Boxplots of Variant Allele Frequency (VAF) to examine clonal status of genes of interest.

Additionally, Maftools allowed several analyses to be conducted including:

- Pair-wise Fisher's Exact test to detect significant pairs of genes with mutual exclusivity or co-occurrence.
- Calculate log odds ratio for genes of interest and generation of forest plots.
- Identification of enriched mutations in categories or groups of interest.
- Identification of mutated genes that are druggable.
- Identification of genes enriched in main oncogenic signalling pathways.
- Identification of *de novo* Mutational Signatures.
- Estimation of APOBEC enriched mutations and samples.

2.5.2 Tumour Mutational Burden (TMB) Calculation

TMB was defined as the total number of somatic, coding, base substitution and InDel mutations detected per Mega base (Mb) of genome examined. *CallableLoci* from GATK (ver. 3.7) was used to obtain callable bases applying the following criteria: a minimum read depth of 10 before a locus was considered callable, a minimum base quality of 20 (based on Phred

scores) and a minimum mapping quality of 30 reads to count towards depth. TMB was calculated as:

$$TMB\ (mut/Mb)\ = \frac{Total\ SNV\ count}{Callable\ Loci} x\ 10^6$$

2.5.3 Mutational Signature Analysis

2.5.3.1 Identification of COSMIC Catalogued Mutational Signatures with deconstructSigs

The DeconstructSigs R package (ver. 1.8.0) was used to determine the contribution of known mutational processes in lung samples. Input data consisted of a data frame containing total mutational data for each sample. Then, the number of times a mutation was observed in each trinucleotide context was calculated for each sample. To determine the signatures characterising lung cancer tumours, signature matrices were calculated based on the fraction of times a mutation was seen in each of the 96-trinucleotide context for each COSMIC catalogued signature (provided by the package). The software then computed weights for each signature using an iterative approach after selecting an initial signature that most closely reflected the mutational profile of a sample by minimising the Sum-Squared Error (SSE) between a tumour sample and the signature. The "exome2genome" normalisation method was used as recommended for exome data to reflect the absolute frequency of each nucleotide context as it would across the whole genome. As a result, a reconstructed mutational profile was obtained based on the final weights. Signature visualisation employed the *plotsignatures* and *makePie* commands.

2.5.3.2 Identification of *de novo* Mutational Signatures with Maftools

To gain insights into the biological mechanisms involved in tumorigenesis, the most frequent combinations of somatic mutations were identified and related to catalogued COSMIC signatures. The workflow of the functions implemented in the Maftools package is shown in Figure 2.4.

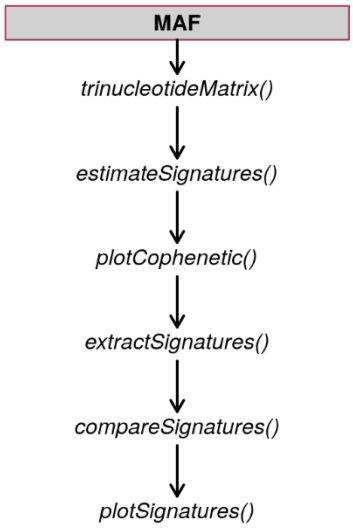


Figure 2. 4| Maftools R workflow used for *de novo* Mutational Signature identification. A trinucleotide matrix was extracted by scanning immediate 5′ and 3′ bases flanking the mutated sites, then number of signatures (n) was decided based on the Cophenetic correlation plot and through non-negative matrix factorization decomposed the matrix into n signatures that were in the end compared to known signatures from the Catalogue Of Somatic Mutations In Cancer (COSMIC) signature (ver.2) database. [Taken from: Mayakonda, A., Lin, D., Assenov, Y., Plass, C. & Koeffler, H. P. Maftools: efficient and comprehensive analysis of somatic variants in cancer. 1–10 (2018) doi:10.1101/gr.239244.118][170].

Specifically, signatures were first estimated using the Non-negative Matrix Factorization (NMF) method[82] by inputting the selected putative pathogenic variants. Then, the number of signatures to be extracted was decided upon based on the Cophenetic correlation value, and finally the identified signatures were related to the Catalogue Of Somatic Mutations In Cancer (COSMIC)[171,172] Legacy signature (ver. 2) database based on cosine similarity values.

2.6 Analysis of DNA Methylation Data

2.6.1 Preparation of libraries for Whole-Genome Bisulfite Sequencing (WGBS)

WGBS libraries were constructed by the McGill Genome Centre (Montreal, Canada) using the KAPA High Throughput Library Preparation Kit (Roche/KAPA Biosystems) with template input being 1 μg of genomic DNA spiked with 0.1% (w/w) unmethylated lambda and pUC19

DNA (Promega). DNA was sonicated (Covaris) and fragments sizes of 300-400 bp were confirmed on a Bioanalyzer DNA 1000 Chip (Agilent). Following fragmentation, DNA end repair of double stranded DNA breaks, 3'-end adenylation, adaptor ligation and clean-up steps were conducted according to the KAPA Biosystems' protocols. The sample was then bisulfite converted using the Epitect Fast DNA bisulfte kit (Qiagen) following the manufacturer's protocol. The resulting bisulfite DNA was quantified with OliGreen (Life Technology) and amplified with 9-12 PCR cycles using the KAPA HiFi HotStart Uracil+ DNA Polymerase (Roche/KAPA Biosystems) according to the recommended protocols. The final WGBS libraries were purified using Ampure Beads, validated on Bioanalyzer High Sensitivity DNA Chips (Agilent) and quantified by PicoGreen (ThermoFisher).

2.6.2 Whole-Genome Bisulfite Sequencing

Sequencing of WGBS prepared libraries was carried out with the paired-end Illumina HiSeq X Next Generation Sequencing at the McGill Genome Centre, Montreal, Canada.

2.6.3 Summary of WGBS-seq Data Analysis

Pre-processing of WGBS-seq data was performed with GenPipes[173], a python-based framework created at McGill Genome Centre. Specifically, the standard Methyl-Seq pipeline adapted from Bismark[174] was used for read alignment and processing of WGBS-seq data. Data analysis was performed in UNIX programming language. After pre-processing of the data, the next steps involved assessment of sequencing performance and estimation of CpG DNA methylation parameters for calling of Differentially Methylated Regions (DMRs). Downstream analyses included motif and pathway analyses, as well as analysis of cis-regulatory regions. A summary flowchart of the steps carried out for WGBS-Seq data analysis is given in Figure 2.5, and details for the individual steps are explained further in this chapter.

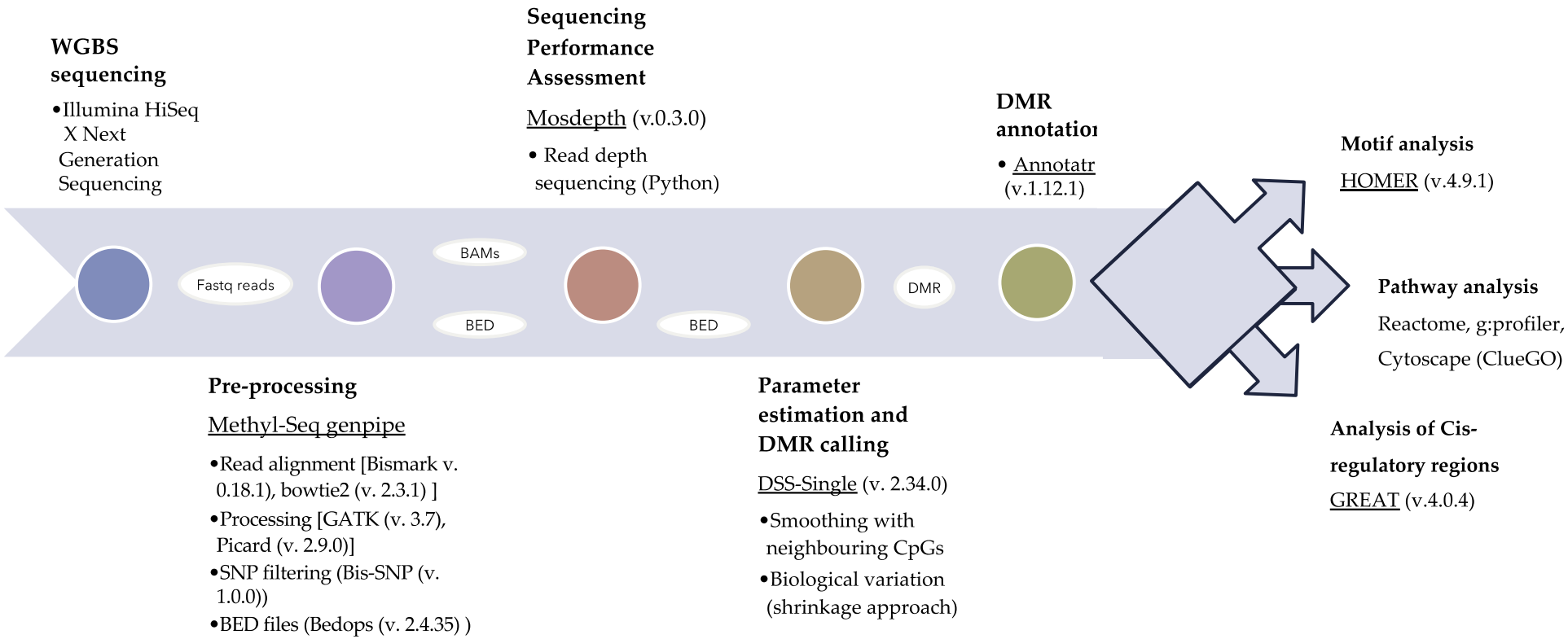


Figure 2. 5| Flowchart of the WGBS-Seq data analysis. Sequencing of WGBS was carried out with the paired-end Illumina HiSeq X Next Generation Sequencing at the McGill Genome Centre (Canada). Then, pre-processing of WGBS-seq data was performed with GenPipes, and involved alignment against the bisulfite converted Human Genome and generation of CpG DNA methylation profiles. Assessment of sequencing performance ensured that methylation differences were not caused by samples with different depths of coverage. Finally, CpG DNA methylation parameters were estimated for calling of Differentially Methylated Regions (DMRs). Downstream analyses included motif and pathway analyses with the annotated DMRs, as well as analysis of cis-regulatory regions.

2.6.4 QC of Fastq Reads

Similar to the DNA sequencing analysis (Section 2.4), the first step of analysis of the WGBS sequencing data involved assessment of the quality of the sequenced bases for both reads in each sample with the FastQC software. FastQC reports were examined before and after adapter trimming and Phred scores were used to exclude low quality reads.

2.6.5 Read Alignment

Alignment of paired-end reads to bisulfite converted Human Genome (GRCh37/hg19) was performed with bismark (v 0.18.1) and bowtie2 (ver. 2.3.1) according to the bismark user guide manual with default options to obtain SAM files (Sequence Alignment Mapping) files. Non-directional option instructs Bismark to use all four strands (OT, CTOT, CTOB and OB) for alignment. SAM files per sample obtained were sorted by chromosomal location with GATK (ver.3.7) and read alignments deemed to be PCR duplicates were removed with Picard (ver.2.9.0) to obtain BAM files (binary form of SAM).

DepthOfCoverage from GATK was used to obtain coverage tracks per sample together with other metrics generated with Samtools (ver.1.4). SAM files were also processed with the bismark methylation extractor command from bismark with default options to extract methylation in CpG, CHG and CHH contexts. Output of bismark methylation extractor was processed to generate a CpG methylation profile by combining cytosines from both forward and reverse strands.

Identification of SNPs is important for accurate quantification of methylation levels, especially given the fact that C>T is the most common substitution in the human population (65% of all SNPs in dbSNP) and these usually occur in the CpG context. Bis-SNP[175] allowed the identification of SNPs and InDels to produce the standard SNP and InDel variants files in Variant Call Format (VCF). The output VCF files were used to filter SNPs and Bedops (ver. 2.4.35) applied to obtain the resulting Browser Extensible Data (BED) files.

2.6.6 Assessment of WGBS Data Read Coverage

To prove that any detected methylation difference between samples (i.e., tumour vs normal) was not influenced by sequencing performance, genome-wide sequencing coverage was assessed with Mosdepth[176]. Mosdepth is a command-line tool that measures sequencing depth from BAM files across the genome. Specifically, the cumulative distribution indicating the proportion of total bases covered for at least a given coverage value was obtained and subsequently used to visualize whole-genome sequencing performance for all the samples. This served as a QC before using methylKit which uses data from individual cytosines in CpG context to obtain coverage and percent methylation by ensuring that methylation differences are not caused by samples with different depths of coverage.

2.6.7 Pre-processing of WGBS Data

Output BED files were the input into the methylKit R package allowing descriptive statistics on samples to be obtained, and filtering performed based on read coverage and hierarchical clustering, using correlation distance and Principal Component Analysis based on samples' methylation profiles.

Basic statistics about the methylation data generated included sample coverage and percent methylation per sample, as well as histogram plots for CpG coverage and percent methylation distribution. Furthermore, *filterByCoverage* command allowed filtering samples based on sequencing coverage. Bases that obtained a coverage below 10X were discarded, to increase statistical power, as were bases with more than 99.9th percentile of coverage for samples suffering from PCR bias. The *normalizeCoverage* command was then used to normalize coverage between samples using a scaling factor derived from differences between median of coverage distributions.

Finally, methylKit allowed the merging of base-pair locations that are covered in all samples for comparative analysis with the *unite* command. The output objects from methylKit were then used for the following analyses:

i. Exploratory data analysis with unsupervised methods:
    a. Hierarchical Clustering Analysis (HCA)
    b. Principal Component Analysis (PCA)

ii. Genomic-context based binning
iii. Differential Methylation (DM) Analysis

2.6.8 Exploratory Data Analysis with Unsupervised Methods for Hypothesis Generation

Both HCA and PCA analysis are methods that allow graphical representations of high-dimensional biological data. WGBS was used to explore the similarity of the DNA methylation profiles of all the samples for hypothesis generation.

Since tumour and non-tumour samples from different lung cancer subtypes (L-CD, LUAD and LUSC) were investigated in this study, this exploratory analysis allowed preliminary information regarding the similarity of their methylomes, such as to what extent DNA methylation was different between cancer subtypes, between tumour and normal samples to be obtained.

i. Agglomerative HCA built a tree-like structure, named a dendrogram, in which the leaves represent individual objects which are successively allocated together for those showing a high degree of similarity. These objects are then collapsed into a higher object or a cluster and processed as a single object in subsequent steps. For WGBS data, each of the samples was represented by the leaves, and enabled the identification of groups of samples with similar DNA methylomes.
ii. Similarly, PCA also enabled dimensionality reduction by projecting data from the two variables that do not correlate and that explain most of the variance. Thus, the complexity of high dimensionality data, such as in this case WGBS data, is reduced into Principal Components (PCs) that concentrate much of the information.

Therefore, HCA and PCA allowed the allocation of samples into homogeneous groups based on their DNA methylation profiles to establish groups for comparison for Differential Methylation analysis.

2.6.9 Genomic Context Based Binning

WGBS is the most comprehensive technique to study DNA methylation covering more than 90% of the cytosines in the human genome. Besides the importance of the coding genome and its regulation, another goal of this study was to investigate to what extent the non-coding genome was aberrantly methylated. To do so, the genome was first binned into genic and non-genic categories to explore and find whether, and to what extent, the different genomic regions contributed to distinguish tumour from non-tumour samples and between tumour subtypes. For this purpose, the Table Browser from the UCSC Genome Browser website was used to retrieve the genomic coordinates of different genomic regions, including promoter regions (by considering 1Kb upstream genes), vista enhancers, repeat regions (considering only four main classes of repeat elements: SINE, LINE, LTR and DNA), intronic regions and exons.

Additionally, regions mapping to different histone marks in lung-tissue were downloaded from the Encyclopedia of DNA Elements (ENCODE). BED narrow peaks for histone marks, obtained from ChiP-Seq experiments on normal left lung tissue of a 54-year-old male adult, were retrieved from the ENCODE database[177].

The *makeGRangesFromDataFrame* from the GenomicRanges (ver.1.38.0) was used to convert input data frames to "GRanges" objects which were then used as input into for *regionCounts* of the methylKit package to extract methylation levels on the regions of interest, as well as perform PCA analysis with the DNA methylation data from these regions.

2.6.10 Calling of Differentially Methylated Regions (DMRs)

The final step in the DNA methylation analysis involved calling of DMRs between groups with the Dispersion Shrinkage for Sequencing data with single replicates (DSS-single) R package (ver.2.34.0). DSS-single is a statistical method for analysing WGBS data that accounts for spatial correlation of methylation levels, sequence depth and biological variation.
The differential methylation analysis involved several steps as follows:

i. Text files containing DNA methylation data per CpG were read and converted to "BSseq" objects.

ii. Mean DNA methylation levels were then estimated for all CpG sites considering information from neighbouring CpG sites in windows of 500 bp.

iii. Sample variance was estimated within defined groups of samples through an empirical Bayes procedure[178] . This is termed the "shrinkage approach" and involves using groups of samples with similar biology (test vs control) as a surrogate for technical replicates, in order to provide more accurate estimates of dispersion.

iv. A statistical hypothesis test for the equality of mean DNA methylation levels (μ) at each CpG site (i) between the two groups was performed based on the following hypotheses:

- Null-hypothesis $H_0: \mu_{I1} = \mu_{I2}$
- Alternative-hypothesis $H_1: \mu_{I1} \neq \mu_{I2}$

Only CpGs that showed a mean methylation difference greater than *delta* between the two groups were used for statistical analysis. Statistical analysis was performed using a Wald test and $P<1x10^{-6}$ was considered statistically significant, and if achieved a Differentially Methylated Cytosine (DMC) was called. Then, a DMR was called when 3 DMCs where found in a region of at least 50 bp.

The above steps were executed using the following commands implemented by the DSS-single R package:

- makeBSseqData
- *DMLtest* with a smoothing span of 500 bp
- callDMR

Finally, DMRs were manually curated and only DMRs with a methylation difference of > 20% between the two comparison groups were considered for annotation.

2.6.11 Annotation of DMRs

DMRs consisted of genomic coordinates which showed higher or lower DNA methylation levels as compared to another set of samples. DMRs were annotated into different categories with the annotatr R package (ver. 1.12.1). Specifically, annotatr allowed retrieval of information on (I) how far from a CpG site the DMR was located and (II) where in a gene or outside a gene a DMR was located. The different genomic categories are explained below:

I. CpG annotations (Fig. 2.6):

- CpG island (CGI): regions rich in CpG sites frequently located at the promoter regions of coding genes where they modulate gene transcription.
- CpG shore: regions immediately flanking and up to 2Kb away from CGIs. They appear variably methylated in cancer and development.
- CpG shelf: regions flanking CpG shores 2Kb upstream/downstream.
- InterCGI: remaining genomic regions that are not considered CGI, CpG shore or CpG shelf.

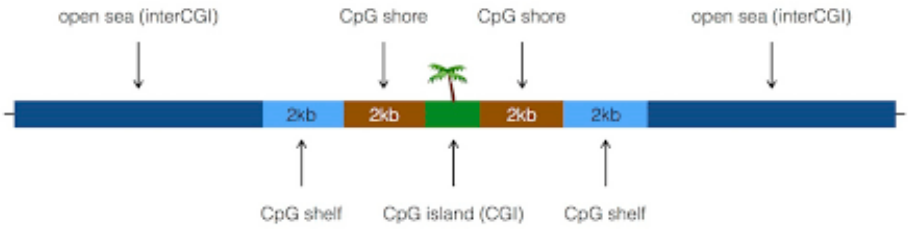


Figure 2. 6| CpG annotations that were retrieved with the annotatr R package based on how far from a CpG site the DMR was located.

II. Genic annotations (Fig. 2.7):

- 1-5Kb: 1 to 5Kb upstream of the TSS.
- Promoter: <1Kb upstream of the TSS.
- The five prime Untranslated Region (5'UTR): region immediately upstream of the coding sequence that is transcribed into mRNA and recognised by the ribosome to initiate translation. However, these regions are usually not translated into proteins.
- Exon: part of a gene sequence that is transcribed and translated into protein.
- Intron: part of a gene sequence that is transcribed but not translated into protein.
- The three prime Untranslated Region (3'UTR): region downstream of the coding sequence that is transcribed into mRNA and follows the termination codon and precedes the poly(A) tail. Similar, to the 5'UTR, this region is neither translated into protein.
- Intergenic regions: exclude the previous annotations.

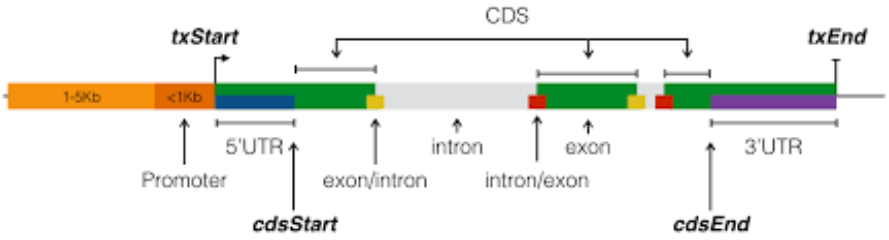


Figure 2. 7| Genic annotations that were retrieved with the annotatr R package based on where in a gene or outside a gene a DMR was located.

The annotatr R package also allowed the number of regions in each annotation type, as well as allowed visualization of CpG annotation counts based on user specified classes - such as hypomethylated and hypomethylated DMRs to be obtained.

2.6.12 Assessment of DNA Methylation in Transposable Elements

Transposable elements (TEs) are patterns of nucleic acids dispersed throughout the genome that account for more than half of the human genome. Most of these do not encode for RNA or proteins but have been historically argued to play important functional and structural roles and have been found aberrantly methylated in many diseases, including cancer.

To assess the level of DNA methylation in these TEs, the four main classes of TEs were focused upon:

- Short interspersed nuclear elements (SINE), which include ALU elements.
- Long interspersed nuclear elements (LINE).
- Long terminal repeat elements (LTR), which include retrotransposons.
- DNA repeat elements (DNA).

Since TEs make up almost half of the nuclear DNA, further assessment was conducted to establish whether and to what extent TEs overlapped with the DMRs identified. To do so, *bedtools coverage* was used to obtain the DMRs that overlapped with TEs at least on a 30% fraction of 30% of the TE size. DMRs were then classified into two groups based on their repeat content: repeat-rich (rrDMR), for DMRs with a repeat content of ≥ 30%, or repeat-free (rfDMR) DMRs, for those with a repeat content of <30%.

2.6.13 Identification of DNA-binding Motifs in DMRs

To identify enriched Transcription Factor (TF) binding motifs in the genomic regions of interest, the HOMER *findMotifsGenome* command from the HOMER (ver. 4.9.1) tool was employed. CpG normalization was used to normalize CpG% content and repeat masked sequences were not considered for this analysis.

2.6.14 Analysis of Cis-regulatory Regions

Cis functions of non-coding genomic regions were studied with the Genomic Regions for Enrichment of Annotations Tool (GREAT) (ver. 4.0.4). Gene regulatory domains were defined as 5Kb upstream and 1Kb downstream of the nearest genes, up to 1Mb extension in each direction, as specified by the "basal plus extension" association rule. The *t*-test metric was used for ranking genes and Gene Ontology (GO) results and only those that obtained a normalised enrichment score of >2, and multiple hypothesis testing corrected *P* values of <0.01 for both the binomial and the hypergeometric distribution-based tests, were considered significant.

2.7 RNA sequencing

2.7.1 Preparation of RNA Sequencing Libraries

RNA sequencing libraries were prepared using the extracted total RNA by Dr. S. Dwyer within the Genomic Medicine Group. Briefly quality and concentrations of samples were analysed with the 2100 Bioanalyser and total RNA Nano Kit (Agilent Technologies, California, United States) following the manufacturer's instructions. The Illumina TruSeq stranded total RNA Gold sample preparation protocol (RS-122-2301) was used to prepare libraries for each of the samples. Samples were pooled and then RNA sequencing was performed at the Imperial BRC Genomics Facility using the Illumina HiSeq4000 system with Sequencing by Synthesis (SBS) chemistry for 75 and 100 paired ends.

2.7.2 Summary of RNA Sequencing Data Analysis

In outline, analysis of RNA sequencing data involved pre-processing and QC of the raw reads, read alignment, processing of the aligned reads to filter out low expressed genes and normalization. Downstream analyses carried out with the normalised RNA-seq data included differential expression and gene set enrichment analyses, as well as data integration with other types of data (Fig. 2.8).

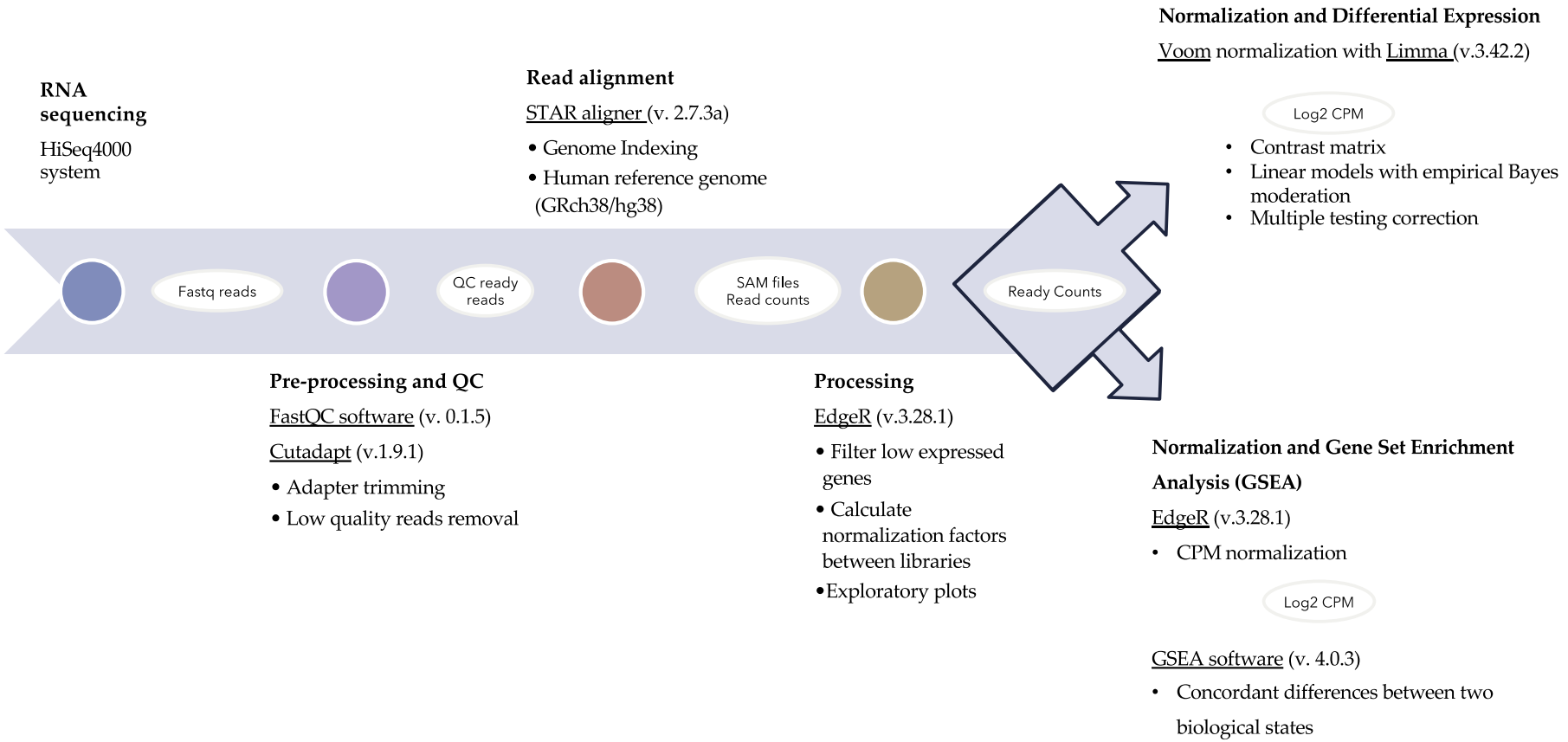


Figure 2. 8| Flowchart of the RNA Sequencing data analysis. RNA sequencing data was generated on a HiSeq4000 system. Analysis of RNA sequencing data involved pre-processing and QC of the raw reads, read alignment, processing of the aligned reads to filter out low expressed genes and normalization. Downstream analyses carried out with the normalised RNA-seq data included differential expression and gene set enrichment analyses, as well as data integration with other types of data.

2.7.3 Pre-processing and QC of Fastq Reads

The FastQC software was used to provide an overview of the quality of the raw RNA sequencing data. Following FastQC, reads were adapter trimmed using cutadapt (ver.1.9.1) in Trim Galore with at least 5 bases match between adapter and read, discarding reads shorter than 50 bp. A second FastQC was run to ensure that adaptors were trimmed and that a low percentage of reads (<10%) were filtered out.

2.7.4 Mapping and Gene Expression Quantification

Read alignment or mapping was performed with the Spliced Transcripts Alignment to a Reference (STAR) (ver.2.7.3a) against the Human Genome December 2013 assembly (GRCh38/ hg38). STAR alignment uses a two-step process to achieve efficient mapping of the reads by accounting for spliced alignments. This means that STAR first finds the longest matching sequence for a read ("Maximal Mappable Prefixes") and the different parts of the read that are left unmapped are subsequently matched against the next longest sequence matching the reference genome as separate alignments. The separate aligned reads that are not multi-mapping are first stitched together to create a complete read, followed by those that obtained the best alignment based on gaps produced by any mismatches and/or InDels.

Aligning reads using STAR involved the creation of a genome index and mapping reads to the genome or read alignment. Both steps were performed in the UNIX programming language. The creation of a genome index was performed with the *genomeGenerate* option using the Ensembl GTF (General Transfer Format) and Ensemble DNA fasta files for the hg38 human genome. The *–sjdbOverhang* option was used to define the number of bases to concatenate from donor and acceptor sides of the junctions, taking into account the length of the trimmed reads. Ideally this would be the length of the donor/acceptor sequence on each side of the junctions (mate_length – 1). Since reads were 75 and 100 bp, separated STAR indexes were generated for each insert size RNA sequencing libraries.

Next STAR alignment of the trimmed Fastq sequences was performed with the *alignReads* command against the indexed genomes. Aligned SAM files were thus obtained as output containing alignment, summary mapping statistics and detailed information about the

run. In addition, the read counts per gene were obtained as tab delimited files by specifying the *--quantMode GeneCounts* option. Additional parameters for STAR alignment included a minimum mapped length of 15 bp for the chimeric segments *(--chimSegmentMin*), length of the donor/acceptor sequence on each side of the junctions (*--sjdbOverhang)*, a maximum gap in the read sequence between chimeric segments of 3 bp (-- *chimSegmentReadGapMax*) and --*BySHout* was used to reduce the number of spurious junctions. Finally, read counts for all the samples were combined in a matrix of *m* genes and *n* samples to be used for differential expression (DE) analysis.

2.7.5 Transposable Element (TE) Expression Analysis

The majority of RNA-seq data analysis software are not designed to map reads coming from Transposable Elements. Measurement of TE expression was therefore performed with *TEcounts* (ver. 2.2.1) using input BAM files generated with Picard *SortSam* command (ver. 2.25.0). TEcounts was specified to be run on input BAM files sorted by chromosome position and for first-strand cDNA library.

Since reads originating from TE could align to multiple loci with equal quality in the genome, TEcounts was run as *multi* mode ensuring counting of all reads. TEcounts first perform read assignment by using an iterative algorithm to optimally distribute ambiguously mapped reads. For TE-associated reads, the multi-mode counted all available alignments of a read, where every alignment was assigned a weight of 1/n, n being the number of alignments. Consequently, the total contribution of a multi-read to the library size was the same as a unique-read, allowing normalisation not to be influenced by library size of the samples. After read assignment, TE transcript estimation using an Expectation-Maximization (EM) algorithm to determine the maximum-likelihood estimates for all multi-read transcripts was performed. Then, the estimated relative abundance of TE transcripts from multi-reads was integrated with unique-read counts to compute the total relative abundance. Similar to gene expression analysis, the output was merged into a single count matrix of *m* TEs and *n* samples to be used for differential expression (DE) analysis.

2.7.6 Differential Expression (DE) Analysis

Differential Expression analysis involved two main steps:

1. Data pre-processing and normalisation (with *edgeR* and *limma*).
2. Differential expression analysis with linear modelling.

Processing of the gene and TE count data involved first filtering and removal of lowly expressed genes. For the detection of DE genes, genes that were not expressed at a meaningful level should be discarded for several reasons. First, transcripts with a read count within the 0-10 range could be considered as artefacts or not biologically meaningful. Secondly, these can increase the number of DE genes after multiple testing correction. And finally, from a statistical point of view, removing low count genes allows a more reliable estimation of the mean-variance relationship. Thus, genes with less than ~5 counts in 20% of the samples were assumed as lowly expressed genes and were filtered out prior to downstream analysis. Finally, raw counts were transformed onto Counts Per Million (CPM), taking into account library size of samples as they can vary due to the sample quality or the sequencing. Transformation from the raw-scale and filtering of lowly expressed genes, as well as unsupervised clustering of samples in an unsupervised manner was carried out with *edgeR* R package (ver. 3.28.1).

Normalisation was then carried out to account for external factors or variables that were not of biological interest but could have an effect on gene expression. This variability is reflected in the variance or statistical dispersion of the data which is expected to not be homogeneous for RNA-seq experiments, where each sample and sequencing run results in individual blocks of data which have been aggregated for comparison purposes. Thus, this heteroscedasticity in gene expression data must be adjusted before fitting it onto a linear model. For this, the *Voom* function in the limma package (ver. 3.42.2) was implemented to fit a linear model for each gene and taking into account the sequencing depths (library sizes) by applying weights to each log-count observation based on its predicted variance. Additionally, the voom plot provided a visual representation of the level of filtering based on graphical representation of the variance as well as ensuring effective removal of lowly expressed genes. Counts were transformed to log 2 counts per million reads (log2 CPM), where "per million reads" were defined based on the normalization factors.

After voom normalisation, differential analysis was performed using linear models with limma R package (ver. 3.42.2). Specifically, *lmFit* and *contrasts.fit* functions were used to perform linear modelling where the first function fits a linear model using weighted least squares for each gene. Subsequently comparison between groups were obtained as contrasts of these fitted linear models with the *contrasts.fit* command. Next, empirical Bayes moderation was carried out to smooth standard errors that were larger or smaller to the average standard error from other genes. Finally, *topTable* was used to extract a table of the top-ranked genes from the linear model fit with 95% confidence interval. Finally *P*-values were adjusted for multiple testing by using the Benjamini, Hochberg, and Yekutieli method[179] using a FDR threshold of ≤ 0.01.

2.7.7 Gene Set Enrichment Analysis (GSEA)

GSEA allows biological insights from gene expression data based on groups of genes that shared common biological functions or signalling pathways to be extracted. GSEA is a computational method that determines whether an a priori defined set of genes shows concordant statistically significant differences between two biological states or phenotypes.

In order to perform this analysis, three files were provided to the GSEA software, including a normalised expression dataset in Gene Cluster Text (GCT) format, a phenotype annotation file and a gene set from the Molecular Signatures Database (MSigDB)[180,181].

The GSEA algorithm provides an Enrichment Score (ES) reflecting how over or under expressed a gene is from the provided expression GCT file with respect to a ranked list of genes, and a ranking metric measuring a gene's correlation with a phenotype. GSEA calculates the ES as the maximum deviation from zero encountered by walking down the ranked list of genes, where a positive ES indicates gene set enrichment at the top of the ranked list, and a negative ES indicates that a gene set is enriched at the bottom of the ranked list. Similarly, a positive ranking metric value indicates correlation with the first phenotype, and a negative value indicates correlation with the second phenotype.

GSEA was run with 1,000 permutations. In the setting of exploratory discovery, a FDR of ≤0.25 and a nominal *P*-value of <0.01 was used in order to identify candidate hypothesis and not overlook potentially significant results.

2.7.8 Cluster Validation Analysis for RNA-seq Data

Clustering is an unsupervised technique used to group together objects which are close to each other to identify inherent structure within the data. For this study, gene expression data from RNA-seq experiments was used for clustering with the aim to find groups of samples that shared similar expression patterns, and thus shared signalling pathways.

Since analysis of RNA-seq data from 15 carcinoids revealed two differentiated groups of samples (Chapter 6), the clValid R package was used for validation of the results from the clustering analysis. The *clValid* command was used to retrieve statistical measures for different number of clusters. The statistical measures fall into three categories:

- Internal validation, for measures that rely on information in the data only, such as compactness and separation between data objects in the clusters. Internal validation can be assessed based on three measures:
    - Connectivity: degree of connectedness of the clusters. Ranges between zero and infinite and should be minimised.
    - Silhouette Width: is the average of each observation's Silhouette value, which defines the degree of confidence in the clustering assignment of a particular observation. Silhouette values range from -1 to 1, where poor clustering will show values near -1 and well-clustered observations will have values closer to +1.
    - Dunn Index: is the ratio of the smallest distance between observations not in the same cluster to the largest intra-cluster distance. Ranges between zero and infinite and should be maximised.
- Stability validation, referring to measurement of the sensitivity of the clustering algorithm by removing each column one at a time and rerunning the clustering. It is represented in several measures such as *average proportion of non-overlap* (APN), the *average distance* (AD), the *average distance between means* (ADM), and the *figure of merit* (FOM), all of which should be minimised.
- Biological measures: evaluates the ability of the clustering to produce biologically meaningful clusters. It can be only applied to biological data such RNA-seq data. This is implemented by providing a biological class for each gene imputed.

Thus, clValid enabled the optimal number of clusters for a particular algorithm used to be determined. The *summary* statement provided the validation measures in a table for the different clustering methods as well as plots for these measures.

Chapter 3: Genomic Landscape of Lung Tumours

3.1 Introduction

Lung cancer (LC) is a world-wide challenge due to its high incidence and associated low survival rate. The latter has not improved, in contrast to other tumour types, with 5-year survival remaining below 20%. Major advancements in sequencing technologies and screening efforts are however insufficient as these are not available for most countries and tumours evolve with new alterations driving their heterogeneous nature. As a result, drug resistance is very frequent among advanced and metastatic LC patients and remains a focus of many preclinical and clinical studies.

In depth understanding of the molecular events during disease treatment, as well as detection of the alterations that determine LC onset, are major challenges. Nonetheless, molecular defined non-small cell lung cancer (NSCLC) subgroups have already shown responses to targeted therapies resulting in improved clinical outcomes. This highlights the importance of maintaining the continuation of investigations of the genetic and epigenetic alterations to identify effective genetically defined subtypes of lung tumours optimizing them for biologically informed patient stratification for personalized therapeutic approaches.

To date there have been very few studies that have systematically explored the genetic and epigenetic alterations in common and rare lung cancers integrating expression and genome-wide DNA methylation data. Thus, the key aim of this chapter is to identify and describe the relevant molecular alterations in different lung cancer subtypes.

3.1 Research Samples Summary

A total of 322 samples, lung cancer tumour and matched normal samples, from 159 patients were used for the genomic study. Eighty-nine patients had lung adenocarcinoma (LUAD), 35 had lung squamous carcinoma (LUSC), 22 had lung carcinoids (L-CD) and 13 had lung neuroendocrine tumours (LNET) (Table 3.1). The patients' ages ranged between 50 and 82 years old.

Specifically, DNA sequencing data was generated for a total of 86 paired tumour-normal samples by using Targeted Capture Sequencing (TCS) of the exonic regions of 52 genes

(Chapter 2, Table 2.1). Furthermore, a total of 73 pairs of tumour and normal samples underwent Whole Exome Sequencing (WES) (Chapter 2, Section 2.4.1). A summary of the number of samples that underwent TCS and WES is provided in Table 3.1.

	LUAD	LUSC	L-CD	LNET	ALL
WES	26/26	13/13	22/22	12/12	73/73
TCS	59/63	20/22		0/1	77/86
Total	85/89 (+ 1 unpaired)	33/35	22	0/13	150/159

Table 3. 1| Number of Lung Cancer samples that were analysed by either WES or TCS across the different histological subtypes. Fractions are given to indicate out of the total number of patients per histology type that underwent the analysis how many in which at least one somatic mutation and/or InDel was detected. Abbreviations: WES, Whole Exome Sequencing; TCS, Targeted Capture Sequencing; LUAD, Lung Adenocarcinoma; LUSC, Lung Squamous Carcinoma; L-CD, Lung Carcinoids; LNET, Lung Neuroendocrine Tumours.

3.2 Targeted NGS of Lung Cancer

3.2.1 Sequencing Library Preparation and Quality Control

Prior to next generation sequencing, prepared sequencing libraries (Chapter 2, Section 2.3.2) were checked (pre and post capture) with an Agilent 2100 Bioanalyzer to confirm that an appropriate DNA fragment size between 325-450 bp was achieved. Figures 3.1 a and b show representative bioanalyzer results that were obtained for all samples sequenced in this project.

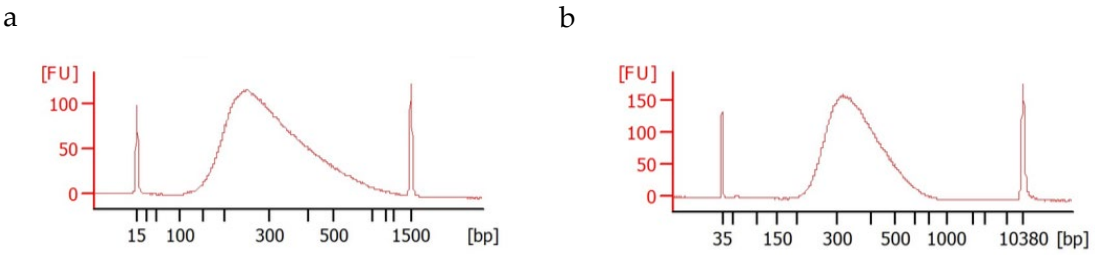


Figure 3. 1| DNA bioanalyzer traces pre and post capture for the paired tumour and normal NSCLC samples obtained using the Agilent 2100 Bioanalyzer and High Sensitivity DNA assays. DNA fragment size of the pre-capture (a) and the post-capture library (b) was verified, DNA fragment size between 245-325 bp and 325-450 bp respectively.

3.2.2 Quality Control of Illumina NextSeq Raw Data

The Illumina NextSeq generated per-cycle base call (BCL) files. The FastQC software provided assessment of the quality of the sequenced bases based on the Phred quality score, which is–

$$-10\ log_{10}\ P$$

where P is the probability of error of the nucleobases generated by automated DNA sequencing. A score of 20 (meaning 1 in 100 chance of error) is generally accepted as the minimum acceptable score. The FastQC report for all the FastQ files obtained a high Phred score and only required trimming of low-quality bases at the 3'end of the reads (Fig. 3.2).

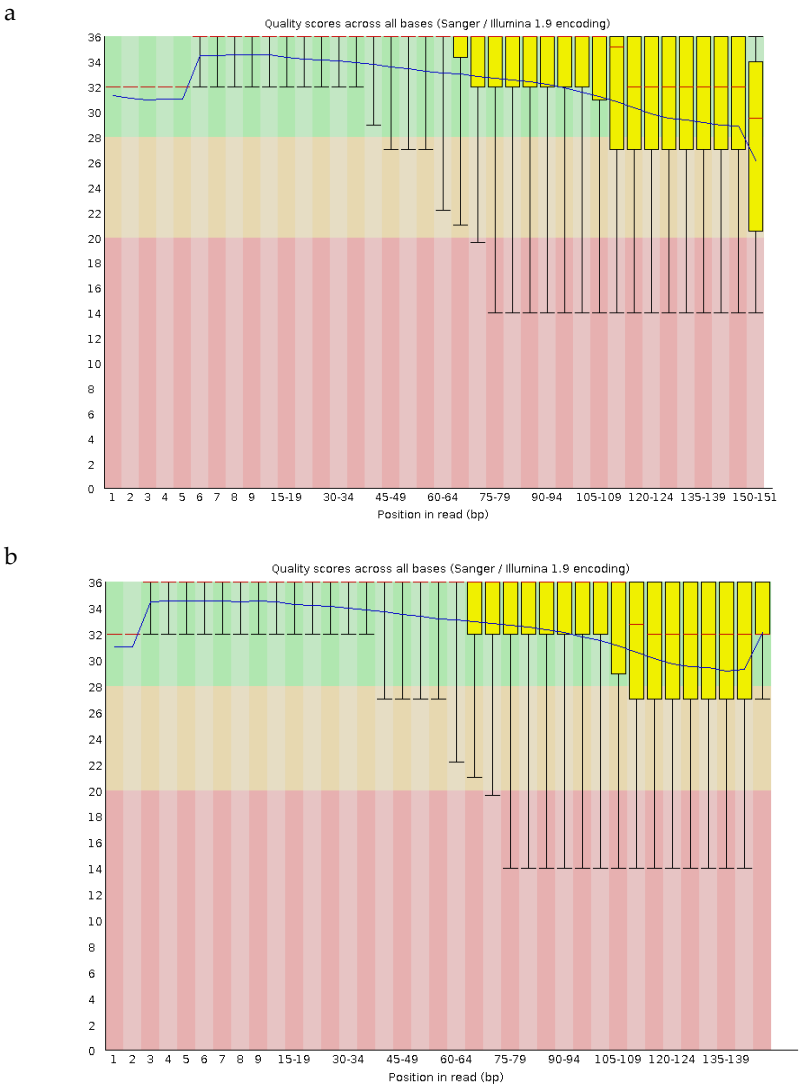


Figure 3. 2| Representative Phred quality scores of the raw data before (a) and after (b) trimming. Distribution of Phred quality scores across the length of fastq reads. *X*-axis is the base position and *Y*-axis is the Phred score. Except for the last few bases, where the sequencing platform's sensitivity drops, the distribution is well within high-quality range. This is remedied by trimming of low-quality bases towards the 3' end of the reads.

Furthermore, pre-processing of the output raw files from the NextSeq prevented low-quality reads from entering the variant evaluation. In addition, local InDel realignments were performed to correct potential mapping errors around regions that are known to contain InDels. Properly paired reads (*i.e.*, those alignments where the R1 and R2 are in opposite orientation to each other and the genomic gap between them is within the accepted range) for each sample were at 90% or above, meaning that sequences were correctly aligned. PCR duplication percentage was below 1%, meaning that when the duplicates were removed there was minimal data loss.

3.2.3 Validation of NGS of Exome Sequenced NSCLC Samples

To validate the NGS panel, it was initially run on a set of 10 patient samples - paired NSCLC tumour and normal tissue - that were already whole exome sequenced. Four pairings were from the LUAD subtype and one from the LUSC subtype. Mean on-target coverage was obtained from Illumina next generation data analysis as previously described (Chapter 2, Section 2.3.6). Coverages for all the samples ranged between 500-970X (Table 3.2). The results confirmed successful gene panel design and sequence alignment of the samples.

Annotated variants for NSCLC patients were obtained from somatic mutation calling and subsequent manual curation within the Integrative Genome Viewer (IGV). The ten patients were known to collectively have somatic variants in 24 of the genes within the 52 gene panel. Mutation in the *TP53* gene was the most frequent somatic alteration being present in 4 out of the 5 patients. The tumour suppressor gene *CREBBP* and *SF3B1* gene involved in RNA splicing were the second most frequent somatic variants with each being present in 40% of the patients (or 2/5 patients). In addition, alterations in oncogenes *PIK3CA*, *EGFR*, *NRAS* and *KRAS* were found, among others. Importantly, WES missed two deletions in *TP53* in sample 345 and *CDKN2A* in sample S2 which were confidently identified by using the targeted capture sequencing probably due to low read depth associated with WES compared with high read depth associated with the TCS. In addition, a germline deletion in *NTRK1* affecting gene splicing was identified for patient S1, as 276 reads matched the alternative allele in the normal sample out of 568 total reads (49% associated variant allele frequency).

						Normal sample		Tumour sample		
Sample	Genomic position (GRCh37/ hg19)	Gene	Variant class	Consequence	Impact	Ref. allele	Alt. allele	Ref. allele	Alt. allele	VAF
S1	17:7579511	*TP53*	deletion	frameshift variant	High	502	0	499	65	11.52%
	7:55259515	*EGFR*	SNV	missense variant	Moderate	618	1	604	80	11.70%
	3:78666910	*ROBO1*	SNV	missense variant	Moderate	913	0	852	26	2.96%
	1:156848908-156848910	*NTRK1*	deletion	splice region variant	Low	292	276	342	291	45.97%
S2	9:21971084	*CDKN2A*	deletion	frameshift variant	High	808	2	651	142	17.91%
	1:115256529	*NRAS*	SNV	missense variant	Moderate	518	1	411	80	16.29%
	2:198281581	*SF3B1*	SNV	missense variant	Moderate	765	0	695	82	10.55%
	17:7577081	*TP53*	SNV	missense variant	Moderate	1009	0	844	156	15.60%
S3	2:198267468	*SF3B1*	SNV	missense variant	Moderate	705	0	580	20	3.33%
	16:3827617	*CREBBP*	SNV	stop gained	High	438	0	510	14	2.67%
	1:27057793	*ARID1A*	SNV	stop gained	High	1263	0	1054	35	3.21%
	2:29498331	*ALK*	SNV	missense variant	Moderate	1333	1	773	235	23.31%
	6:117679114	*ROS1*	SNV	missense variant	Moderate	668	3	549	125	18.55%
	9:139399245	*NOTCH1*	SNV	missense variant	Moderate	1324	0	1154	51	4.23%
	12:25380275	*KRAS*	SNV	missense variant	Moderate	806	1	636	177	21.77%
S4	17:7578384-7578427	*TP53*	deletion	frameshift variant	High	695	0	283	112	28.35%
	3:178936082	*PIK3CA*	SNV	missense variant	Moderate	322	0	153	52	25.37%
S5	17:7577548	*TP53*	SNV	missense variant	Moderate	606	2	198	39	16.46%
	12:46246527	*ARID2*	SNV	missense variant	Moderate	726	0	305	37	10.82%
	2:212248538	*ERBB4*	SNV	missense variant	Moderate	774	1	321	14	4.18%
	16:3827662	*CREBBP*	SNV	splice region variant	Low	317	0	42	2	4.55%

Table 3. 2| Annotated variants from VEP and manual curation from IGV. Two deletions in *TP53* and *CDKN2A* (in red) were not identified with WES. Manual curation from IGV of the validation set from host group (Amit Mandal, NHLI, Personal communication).

3.2.4 Targeted NGS of NSCLC – Additional Samples (that were not whole exome sequenced)

After validation of the TCS panel, 172 paired fresh frozen human NSCLC normal and tumour samples, and one unpaired sample, were successfully sequenced using the Illumina NextSeq 550 platform (Chapter 2, Section 2.3.3). The analysed samples had a tumour content varying from 10 to 100%. After processing and verification of appropriate DNA fragment size as

previously described (Chapter 2, Section 2.3.2), samples were 24-plexed on each NextSeq run. On average for each sample the yield of the sequencing was 2.32 Mb with a coverage of 860.7X after removing PCR duplicates.

Calling of somatic SNPs and Indels were performed by the analysis of matched tumour-normal samples using VarScan (Chapter 2, Section 2.3.6). The LUADs analysed in this study displayed a large number of DNA alterations, with an average of 2.57 somatic SNPs and 0.54 InDel per tumour analysed. LUSCs displayed a total of 3.27 somatic SNPs and 0.5 InDel per tumour. TCS data was not available for LNET and L-CD histologies (Table 3.1). To refine variant calls, high-confidence somatic alterations were obtained by further filtering on the basis of coverage, Variant Allele Frequency (VAF), strand presentation and associated *P*-value. High-confidence calls were used for gene annotation with VEP. Variants were selected as previously explained (Chapter 2, Section 2.3.8) based on VAF (Appendix 2) and clinical impact.

3.3 Whole Exome Sequencing

A total of 73 paired fresh frozen human LC normal and tumour samples were successfully sequenced using the Illumina HiSeq2000 platform.

A total of 19,197 Single Nucleotide Variants (SNVs) and 942 InDels were detected in the 73 LC paired samples. After filtering based on known impact predictions and CADD scores (Chapter 2, Section 2.4.3), a total of 7,681 somatic genetic alterations (7,362 mutations and 319 InDels) in 4,995 different genes were retained and underwent further analysis.

Somatic mutations were detected in all 73 tumour samples, whilst somatic InDels were detected in only 60 samples (82%). Of the 13 tumours with no detected InDels, 85% were of the L-CD histology.

C>A and C>T substitutions were the most frequent class of base substitution across the 73 tumours, with a median of 76 variants detected per sample (Fig.3.3). The top recurrently mutated genes were *TP53* (30%), *CSMD3* (27%), *RYR2* (25%), *TTN* (23%), *LRP1B* (22%), *USH2A* (15%), *COL11A1* (15%), *ANK2 (*14%), *NAV3* (12%) and *C1orf173* (12%).

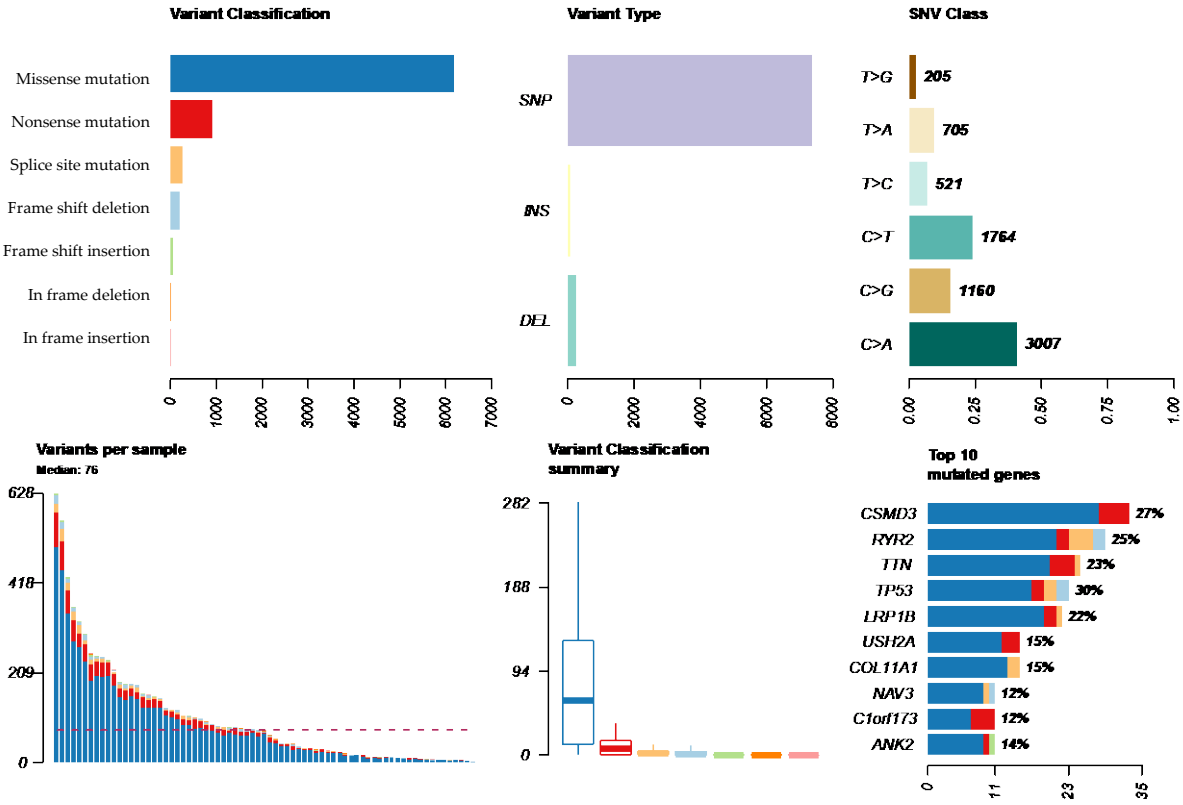


Figure 3. 3| Summary of the somatic variants detected by WES in 73 Lung Cancer tumours. Bar plots show proportion of variant classifications of base substitution and InDels; variant types; Single Nucleotide Variant (SNV) class; number variants per sample; box plot of variant classification summary; and top 10 mutated genes.

A summary of the types of the SNV substitution class, number of variants per sample, TMB and top mutated genes at the whole-exome level and amongst the genes that were included in the gene panel (*n*=52) is given in Table 3.3 below.

	Counts per SNV class		Variants per Sample	TMB (mean, SD)	Top 10 mutated genes	Top 10 mutated genes amongst the 52 of the gene panel
LUAD (*n* =26)	T>G	99	81	16.97 (15.28)	*LRP1B* (35%), *CSMD3* (31%), *RYR2* (23%), *COL11A1* (23%), *TTN* (23%), *TP53* (23%), *ANK2* (23%), *OBSCN* (19%), *NAV3* (19%), *CSMD1* (19%)	*TP53* (23%), *KRAS* (19%), *STK11* (19%), *CDKN2A* (15%), *KEAP1* (15%), *EGFR* (12%), *ARID1A* (8%), *ERBB4* (8%), *ROBO1* (8%), *SF3B1* (8%)
	T>A	335				
	T>C	220				
	C>T	740				
	C>G	485				
	C>A	1492				
LUSC (*n* =13)	T>G	48	109	21.76 (21.04)	*TP53* (62%), *TTN* (62%), *CSMD3* (38%), *LRP1B* (38%), *RYR2* (38%), *USH2A* (31%), *ANK2* (23%), *COL11A1* (23%), *KMT2D* (23%), *OTOF* (23%)	*TP53* (62%), *NFE2L2* (15%), *PTEN* (15%), *ARID1A* (8%), *CCND1* (8%), *CDKN2A* (8%), *CREBBP* (8%), *KEAP1* (8%), *KRAS* (8%), *MET* (8%)
	T>A	174				
	T>C	133				
	C>T	461				
	C>G	390				
	C>A	775				
LNET (*n* =12)	T>G	44	127	25.08 (11.15)	*TP53* (67%), *CSMD3* (58%), *RYR2* (42%), *MYH4* (42%), *USH2A* (33%), *SHPRH* (33%), *ENPP2* (33%), *C5orf42* (25%), *APOB* (25%)	*TP53* (67%), *RB1* (25%), *ERBB4* (17%), *STK11* (17%), *ALK* (8%), *ARID1A* (8%), *ARID1B* (8%), *CDKN2A* (8%), *ERBB2* (8%), *KRAS* (8%)
	T>A	174				
	T>C	125				
	C>T	454				
	C>G	265				
	C>A	657				
L-CD (*n* =22)	T>G	11	12	6.37 (3.65)	*ARID1A* (19%), *VARS* (10%), *RYR2* (10%), *PCDH8* (10%), *PABPC1* (10%)	*ARID1A* (19%), *BRAF* (5%), *SF3B1* (5%)
	T>A	19				
	T>C	32				
	C>T	70				
	C>G	15				
	C>A	74				

Table 3. 3| Summary of the mutational data obtained through WES in four different Lung Cancer subtypes. Specifically, types of the SNV substitution class, number of variants per sample, TMB and top mutated genes at the whole-exome level and amongst the genes (*n*=52) that were included in the gene panel is detailed. Abbreviations: LUAD, Lung Adenocarcinoma; LUSC, Lung Squamous Carcinoma; LNET, Lung Neuroendocrine Tumour; L-CD, Lung Carcinoid; SNV, Single Nucleotide Variant; TMB, Tumour Mutational Burden; SD, standard deviation.

3.3.1 Known COSMIC Mutational Signatures from WES Data

Mutational signatures define specific profiles of trinucleotide changes that are likely to be related to a mutational process active in a tumour sample. To gain insights into the biological mechanisms involved in LC carcinogenesis, mutational signatures were identified and related to those collated in the COSMIC database (https://cancer.sanger.ac.uk/signatures/) using deconstructSigs (Chapter 2, Section 2.5.3.2). Only SNVs were used for this analysis.

COSMIC mutational signature (CMS) 4 was the most prevalent mutational signature and was detected in nearly all LC histotypes (57.7% in LUADs, 69.2% in LUSCs and 83.3% in LNETs), the exception being L-CDs where prevalence was low (4.5%). CMS 4 is characterised by C>A and T>A mutations, consistent with the most frequent types of base substitutions identified in the tumours (Table 3.3). CMS 4 exhibits a strong transcriptional bias compatible with purine nucleotides being repaired by the transcription-coupled nucleotide excision repair machinery. This signature has also been attributed to environmental mutagens and has been found especially enriched in lung cancer genomes of tobacco smokers[82].

In LUADs, CMS 2 and CMS 29 were both identified in 57.7% of tumours. CMS 2 has been associated with the activity of AID/APOBEC family of cytidine deaminases. CMS 29 exhibits transcriptional bias for C>A mutations associated to guanine damage (similar to CMS 4) and has been associated with CC>>AA double nucleotide substitutions. The CMS 29 signature has been found exclusively in gingiva-buccal oral squamous cell carcinoma of tobacco chewers[182,183] and is related to smokeless tobacco. This however was not always the case and a need to re-examine this signature has been recently suggested[184].

CMS 3 and 13 were each identified in 53.8% of LUSC tumours. CMS 3 has been found associated with failure of DNA Double-Strand Break (DSB) repair by Homologous Recombination (HR), whilst CMS 13 was found associated with AID/APOBEC activity causing predominantly C>G mutations[185–189].

For LNETs, in addition to the very frequent CMS 4 (83.3%), CMS 16 and CMS 29 were identified in 66.7% and 50% of tumours respectively. CMS 16 has been found in liver cancers although its aetiology remains unknown, but has been recently found significantly associated with male gender and alcohol consumption[190]. Consistently, 62.5% of LNETs and 61.5% of LC tumours harbouring CMS 16 were males.

The signatures more frequently identified in L-CDs were CMS 3, present in 45.5% of the tumours, associated with failure to repair DSBs. Also, CMS 29 associated with tobacco chewing habit; and CMS 16 and 8 both of unknown aetiologies (each in 36.4% of the tumour samples). Two other recurrent mutational signatures (CMS 20 and CMS 24), associated with defective DNA mismatch repair (MMR) and exposure to aflatoxin, were also identified in L-CDs (Supplementary Table 3.1).

In addition to the percentage of samples showing specific CMSs, the contribution of known mutational processes to an individual tumour mutational profile was determined based on the weights assigned by deconstructSigs. The normalized weights of each signature across the different LC tumours is shown in Figure 3.4.

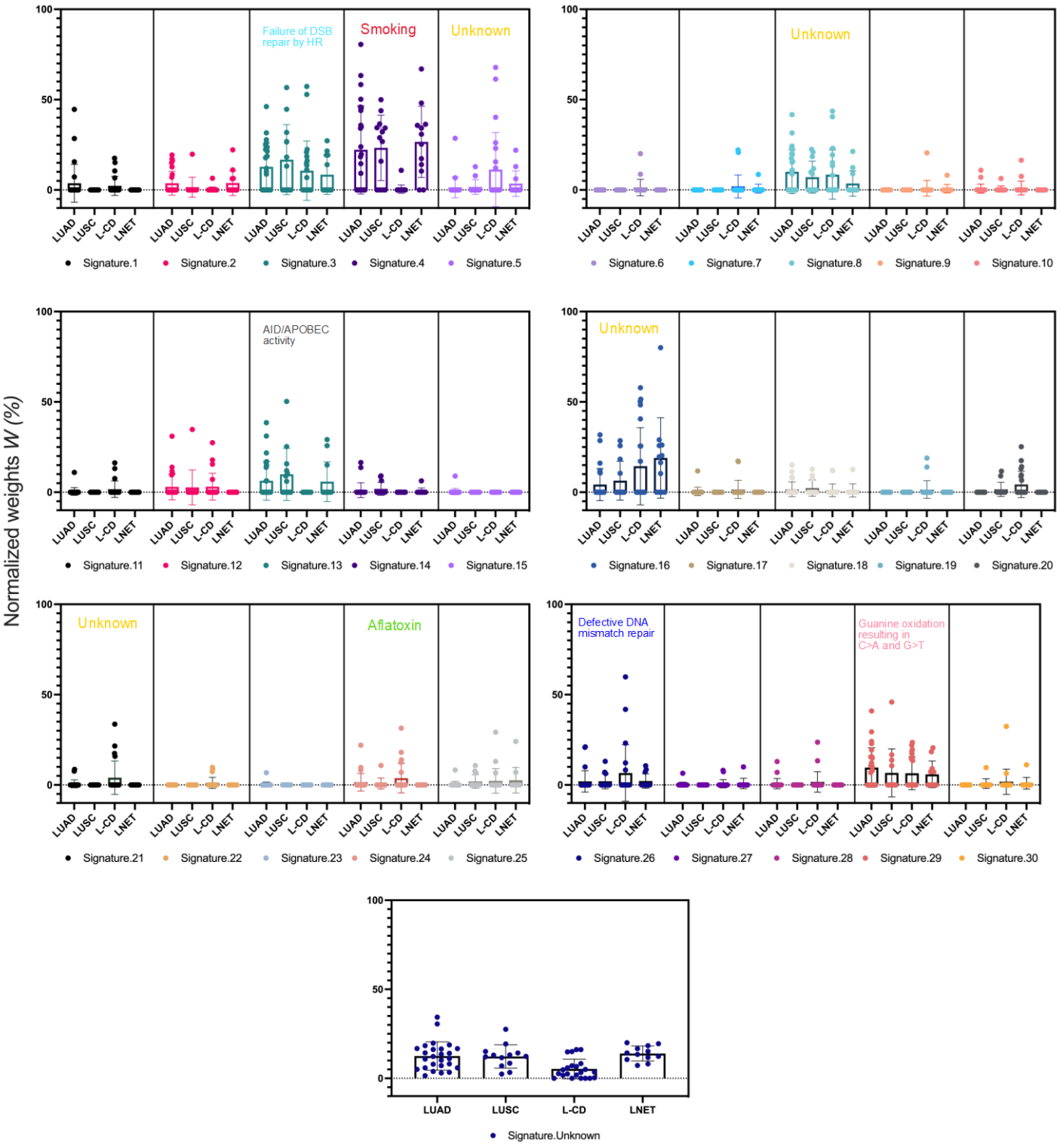


Figure 3. 4| COSMIC Mutational signatures identified with deconstructSigs from total SNVs detected by WES in the four LC histotypes. Box plots and whiskers show the median, first quartile and third quartiles. Individual values representing individual tumours of each histotype are shown superimposed on the graph. Abbreviations: LUAD (Lung Adenocarcinoma); LUSC (Lung Squamous Carcinoma); LNET (Lung Neuroendocrine Tumour); L-CD (Lung Carcinoid).

CMS 4 was not only the most frequent mutational signature across the four different LC subtypes but also contributed the highest to the mutational profiles of LUADs (inter-sample range 9.18%-80.55%), LUSCs (inter-sample range 15.89%-49.96%) and LNETs (inter-sample range 10.41%-66.95%), but only showed a 10.86% contribution in one L-CD tumour sample.

Other signatures also contributed highly to the mutational profiles across all four LC histotypes. These signatures were CMS 3 (inter-sample range LUAD: 8.78%-46.17%; inter-sample range LUSC: 13.17%-56.69%; inter-sample range: 13.98%-27.25%; L-CD: 6.85%-52.28%); CMS 8 (inter-sample range LUAD: 8.77%-41.73%; inter-sample range LUSC: 8.97%-22.51%; inter-sample range LNET: 8.99%-21.31%; inter-sample range L-CD: 8.06%-43.63%) and CMS 29 (inter-sample range LUAD: 6.84%-40.96%; inter-sample range LUSC: 10.59%-45.92%; inter-sample range LNET: 7.14%-20.59%; inter-sample range L-CD: 13.47%-23.54%).

Interestingly, unknown mutational signatures were found in all LUAD, LUSC and LNET tumours, and 81.8% of L-CD tumours. Thus, the next step of the analysis was the identification of *de novo* mutational signatures (dnCMS) (Chapter 2, Section 2.5.3.2).

3.3.2 *De Novo* Mutational Signatures

A total of 3 dnCMSs were extracted for LUADs and were related to exposure to tobacco smoking (CMS 13; cosine similarity of 0.791); spontaneous deamination of 5-methylcytosine (CMS 1; cosine similarity of 0.844) and exposure to tobacco (smoking) mutagens (CMS 4; cosine similarity of 0.933) (Fig. 3.5).

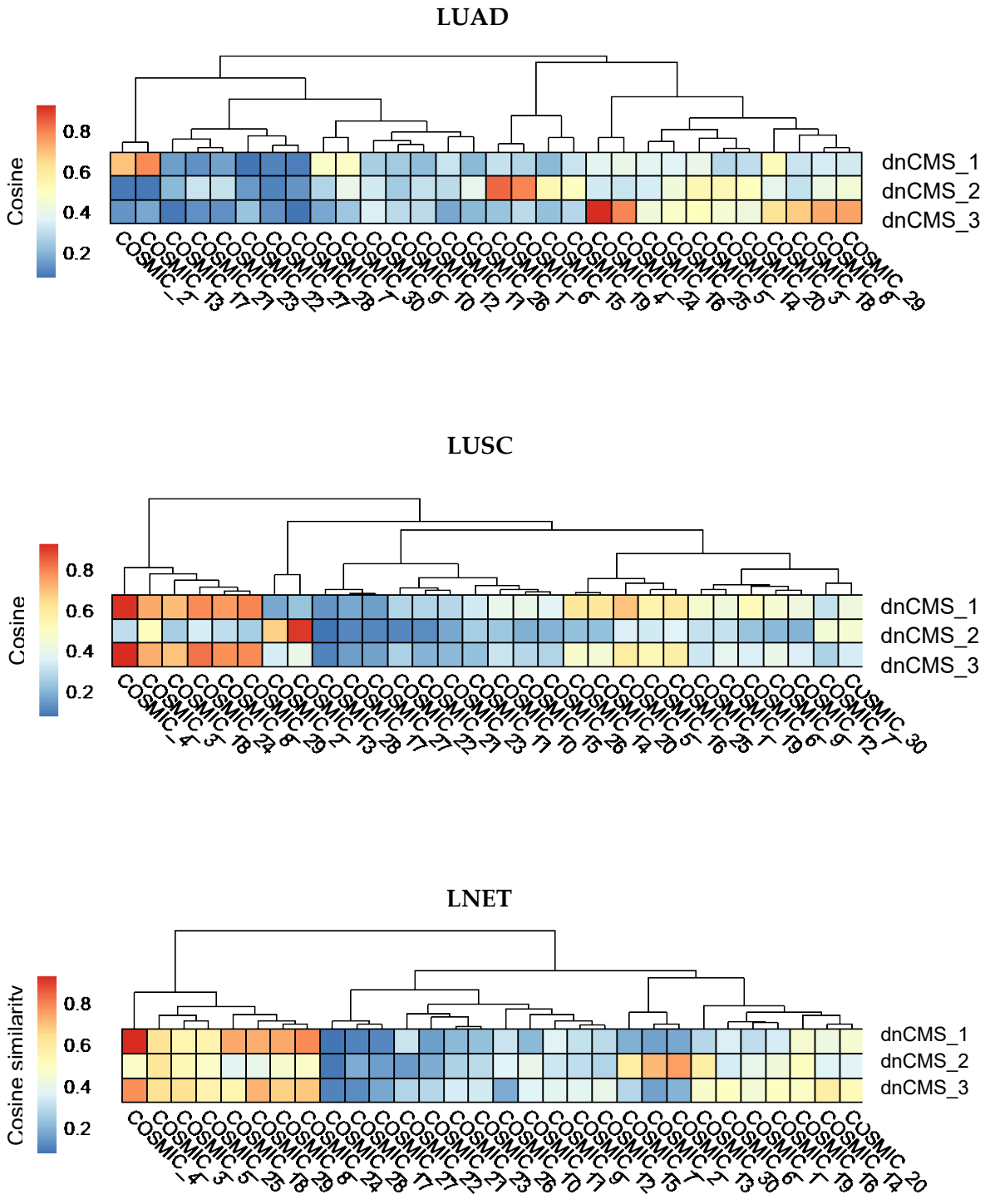
LUSCs showed two dnCMSs, both associated with exposure to tobacco smoking mutagens (CMS 4; cosine similarities of 0.892 and 0.883) and to APOBEC cytidine deaminase (C>G) (CMS 13; cosine similarity of 0.868).

Similarly, LNETs showed the same spectrum of mutational signatures as compared to LUSCs (CMS 4; cosine similarities of 0.895 and 0.753 and CMS 13; cosine similarity of 0.726).

For L-CDs 6 dnCMSs were detected, one related to spontaneous deamination of 5-methylcytosine (CMS 1; cosine similarity of 0.801), two associated with exposure to aflatoxin (CMS 24; cosine similarities of 0.499 and 0.577), one to UV exposure (CMS 7; cosine similarity of 0.6); one to defective DNA mismatch repair (CMS 6; cosine similarity of 0.625) and one

equally related to exposure to tobacco chewing mutagens and unknown aetiology (CMS 18 and CMS 29, respectively; cosine similarities of 0.535 for both).

Heatmaps of the cosine similarities of dnCMS extracted to known COSMIC signatures are shown below in Figure 3.5 for each LC histotype.



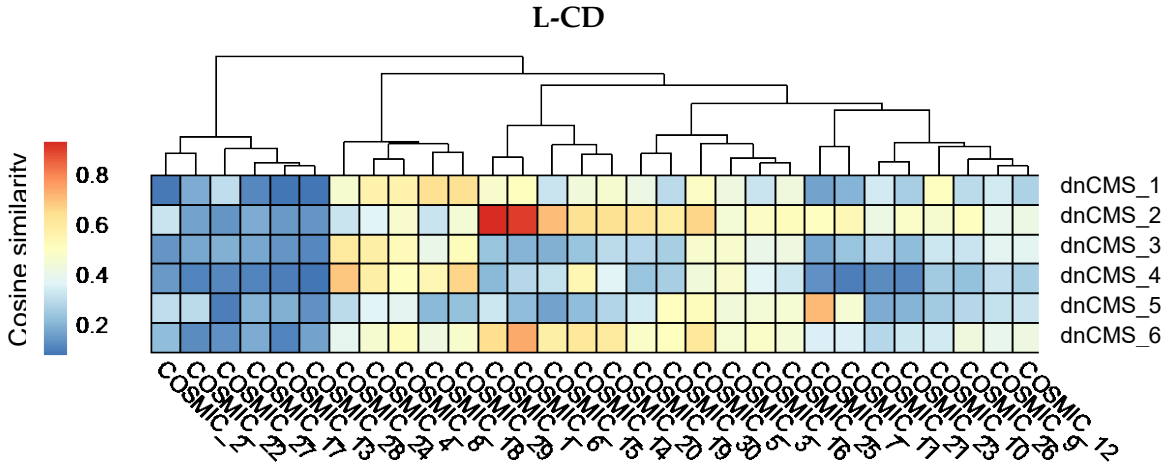


Figure 3. 5| *De novo* Mutational signatures identified in lung cancer histotypes using matrix factorization and compared to known mutagenic processes identified by Alexandrov and colleagues. Abbreviations: dnCMS, *de novo* (COSMIC) Mutational signatures; Lung Adenocarcinoma (LUAD); Lung Squamous Carcinoma (LUSC); Lung Neuroendocrine Tumours (LNETs) and Lung Carcinoids (L-CDs).

3.4 Integration of TCS and WES DNA Sequencing Data

To get further insights into the common and different genetic alterations between the four different LC subtypes, mutational data obtained by WES and TCS were merged, with a focus on the 52 genes that were included in the gene panel. A summary of the most frequent somatic mutations and InDels in the 52 genes across all lung tumours from the four different histologies is shown in Figure 3.6.

TP53 was the topmost frequently altered gene across nearly all subtypes (LUADs 45.88%; LUSCs 87.88%, LNETs 66.67%) with the exception of the L-CD subtype that showed no mutations. Other recurrently mutated genes in LUADs were *KRAS* (30.58%), *STK11* (22.35%), *EGFR* (15.29%) and *RBM10* (14.11%). In LUSCs frequent mutations were identified in *PTEN* (27.27%), *CDKN2A* (21.21%), *KEAP1* (18.18%), *NF*1 and *RB1* (15.15% each). LNETs top mutated genes included *RB1* (25%), *STK11* and *ERBB4* (16.67% each). Finally, L-CDs harboured *ARID1A* mutations (18.18%), *SF3B1* and *BRAF* mutations (4.55% each) and distinctively had no other genetic alterations within the most frequently altered genes seen for NSCLCs and LNETs. Out of the 159 tumours, only 4 LUADs, 2 LUSCs and 1 LNET of LCNEC sub-histology were found to have no mutations, representing ~4% of the whole LC data set.

Co-occurrence or mutual exclusivity of mutations across the 52 genes of the panel was investigated next. In LUADs, mutations in *EGFR* and *KRAS* appeared as mutually exclusive ($P = 0.007$), while gene pairs *KRAS-SMARCA4* ($P = 0.035$) and *SETD2-CREBBP* ($P = 0.046$) were co-mutated in LUADs.

Overall, median Variant Allele Frequencies (VAFs) of >20% were obtained for the top 10 mutated genes across LC subtypes, with *RB1* showing the highest median VAF of 27.78%. In LUADs, *STK11* (27.65%), *ROS1* (27.27%) and *CDKNA* (25.51%) were the genes with highest VAFs. In LUSCs, higher median VAFs were detected for *MET*, *EGFR* and *NF1*, with VAFs of 34.92%, 32.15% and 31.46%, respectively. Although clonal genes usually show VAFs of around ~50% assuming pure samples, these findings suggest potential distinct genes driving tumorigenesis within NSCLC tumours. Genes mutated in LNETs showed the highest median VAF across the four different histotypes with VAFs of 77.27% for *CDKN2A*, 66.04% for *PTEN*, 62.69% for *ARID1A*, 50% for *TP53*, 49.18% for *ALK* and 46.49% for *KRAS*. In contrast, one L-CD showed a median VAF of 36.59% for *SF3B1*.

Figure 3. 6| Oncoplot of the genomic alterations identified by WES and TCS in 159 Lung Cancer Patients. Columns represent each patient and in rows are listed the 52 genes from the gene panel. Genomic alterations are coloured by type of mutation or InDel. Left bar plot represents median Variant Allele Frequencies for each gene. Bars on the right represent percentage of patients with genetic alteration for the four different histologies. Bottom bar indicates clinical feature membership. Abbreviations: VAF, Variant Allele Frequency; TNM, tumour, nodes and metastasis; TCS, Target Capture Sequencing; WES, Whole-Exome Sequencing; PYH, Pack-year history. Pack-year history categories: heavy smoker (>=80 cigarretes); smoker (<20,<80); light smoker (=<20); never smoker; LUAD, Lung Adenocarcinoma; LUSC, Lung Squamous Carcinoma; LNET, Lung Neuroendocrine Tumour; L-CD, Lung Carcinoid; SCLC, Small Cell Lung Cancer; LCNEC, Large Cell Neuroendocrine Carcinoma; LNET-Combined, combined Small Cell Lung Cancer and Large Cell Neuroendocrine Carcinoma.

3.4.1 Association of Genetic Alterations with Clinical Parameters

Several genes detected altered by WES and TCS were found to be associated with particular LC histotypes (Fig. 3.7). *TP53* appeared associated with LUSC histology ($P = 4.26 \times 10^{-7}$; FDR = 1.1×10^{-4}) followed by *PTEN* ($P = 5.52 \times 10^{-5}$; FDR= 7.29×10^{-3}); whilst *KRAS* ($P = 8.87 \times 10^{-5}$; FDR= 7.29×10^{-3}), *STK11* ($P = 1.32 \times 10^{-3}$; FDR= 5.97×10^{-2}), *EGFR* ($P = 2.45 \times 10^{-3}$; FDR= 9.52×10^{-2}), *RBM10* ($P = 1.58 \times 10^{-2}$; FDR= 0.289) and *SETD2* ($P = 3.88 \times 10^{-2}$; FDR= 0.542) were associated with LUADs. In LNETs, *CSMD3* ($P = 1.26 \times 10^{-4}$; FDR= 7.29×10^{-3}), *MYH4* ($P = 1.33 \times 10^{-4}$; FDR= 7.29×10^{-3}), *USH2A* ($P = 5.59 \times 10^{-3}$; FDR= 0.181), *RYR2* ($P = 6.24 \times 10^{-3}$; FDR= 0.181), as well as *PKHD1*, *C5orf42*, *OTOF*, *NEB*, *PTPRD*, *C1orf173* and *RB1* ($P < 0.04$; FDR= 0.181 for each of these seven genes) showed significant associations. L-CDs did not show any association with genetic alterations because of the lack of recurrent mutations (Section 3.1 above).

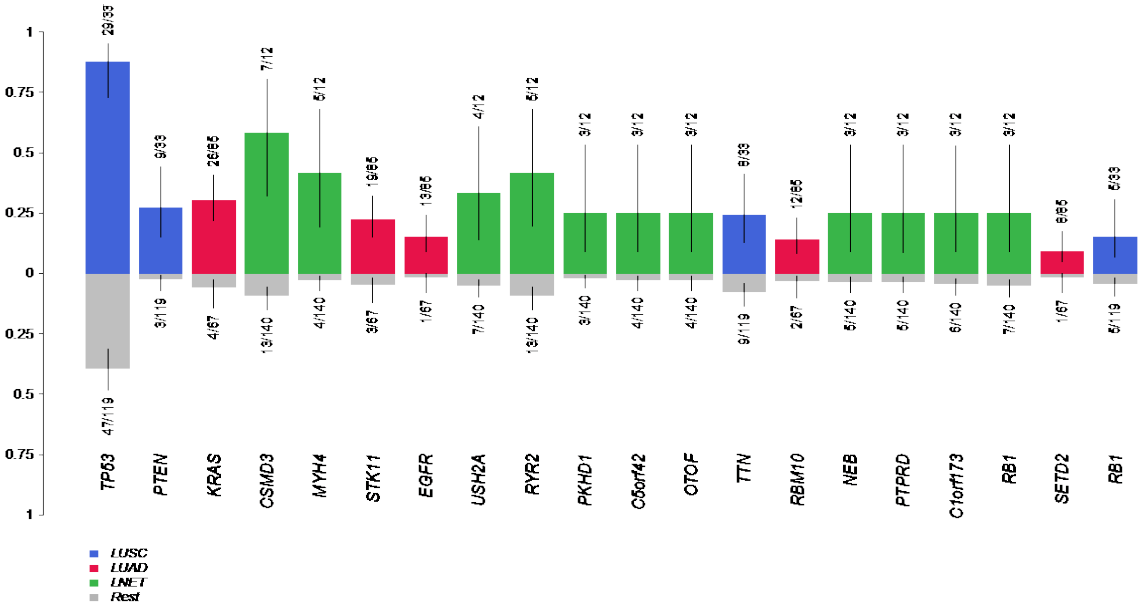


Figure 3. 7| Association of the genetic alterations detected by both WES and TCS with Lung Cancer histological subtypes. Bar plots show the fraction of samples with a genetic alteration in a gene out of the total of samples with detected mutations on the same gene in a LC histotype (top bar) compared with the rest of LC histotypes (bottom grey bar). Abbreviations: LUAD (Lung Adenocarcinoma); LUSC (Lung Squamous Carcinoma); LNET (Lung Neuroendocrine Tumour).

The different number of exonic regions scanned with WES and TCS could however have biased these results. Associations between genetic alterations and histological subtypes were therefore assessed separately for samples sequenced by WES and TCS. In tumours that

underwent WES, *TP53* was enriched in both LNETs ($P = 4.95x10^{-3}$; FDR= 0.37) and LUSCs ($P = 1.03x10^{-2}$; FDR= 0.58). *TTN* ($P = 1.19x10^{-3}$; FDR= 0.26) was enriched in LUSCs; whilst *MYH4* ($P = 4.62x10^{-3}$; FDR= 0.37) and *CSMD3* ($P = 1.43x10^{-2}$; FDR= 0.64) were enriched in LNETs, and *EYS* ($P = 1.95x10^{-2}$; FDR= 0.72) in LUADs.

Similarly, for tumours that underwent TCS, *TP53* ($P = 5.15x10^{-5}$; FDR= $1.23X10^{-3}$) but also *PTEN* ($P = 9.11x10^{-4}$); FDR= $1.11X10^{-2}$) were enriched in LUSCs, with *KRAS* ($P = 1.89x10^{-2}$; FDR= 0.15) and *STK11* ($P = 4.81x10^{-2}$; FDR= 0.28) were enriched in LUADs.

In the case of biological sex, *TP53* (two-sided Fisher's exact test: odds ratio 2.48 [95% CI 1.23, 5.06], $P = 9.06x10^{-3}$) and *TTN* (two-sided Fisher's exact test: odds ratio 7.75 [95% CI 1.7, 72.35], $P = 1.89x10^{-3}$) mutations appeared significantly enriched in men consistent with recent findings showing X-linked genes engaged in p53 networks[191], whilst no mutations showed any enrichment in females. In the case of tobacco smoking, assessed by the number of cigarettes smoked per year (PYH: pack-year history), mutations in *SF3B1* ($P = 3.96x10^{-3}$; FDR= 0.54), *DST* ($P = 2.25x10^{-2}$; FDR= 1), *DSPP* ($P = 2.25x10^{-2}$; FDR= 1), *NF1* ($P = 2.51x10^{-2}$; FDR= 1) and *SMARCA4* ($P = 3.64\ x10^{-2}$; FDR= 1) were associated with heavy smokers whilst *STK11* ($P = 1.16x10^{-3}$; FDR= 0.31) and *TP53* ($P = 2.51x10^{-2}$; FDR= 1) mutations were significantly associated with medium smokers.

Considering survival, LUAD tumours harbouring *EGFR* mutations and small deletions in exons 19 and 20 correlated with longer survival time compared to patients with wild-type *EGFR* although this was not statistically significant ($P = 0.0871$).

3.4.2 APOBEC enrichment

The activity of several cytidine deaminases, which convert cytosine bases (C) to uracil (U), has been associated with DNA damage and the cause for mutations and genetic alterations in many cancers. In addition to their mRNA editing function, cytidine deaminases can also bind to ssDNA substrates, thereby affecting several cellular functions (positively or negatively). The latter include immune editing, retroelement restriction, DNA damage responses, gene expression and DNA demethylation amongst others. These enzymes, when mis-regulated, can become a major source of genetic alterations. Specifically, clusters of mutations have been found in tumour genomes associated to an increased APOBEC expression.

The enzymes target limited areas of ssDNA in order to deaminate C, mainly TCW motifs, where "W" corresponds to either adenine (A) or thymine (T). Therefore, APOBEC-signature mutations have been defined as C>T and C>G substitutions in TCW context, as well as the complements (WGA with G>A and G>C substitutions). So in this chapter the base substitutions were scanned for in the +/-20 nucleotide context surrounding the specific motifs by using maftools (Chapter 2, Section 2.5.1).

For LC samples with mutations in these specific trinucleotide contexts, 24.8% were found to be associated to APOBEC activity. Looking at individual histotypes, LUSC tumours were the subtype with the highest APOBEC enrichment with 28.6% of tumours being associated to its activity, followed by LUADs (26.15%), LNETs (25%) and L-CDs (15%).

Furthermore, ssDNA substrates can be produced by the activity of the DNA mismatch-repair (MMR) pathway. MMR genes obtained from the Human DNA repair genes[192] table were therefore examined for mutations and found to be mutated in 6.45% of APOBEC-enriched tumours and similarly in 5.32% of non APOBEC-enriched tumours, suggesting other potential sources contributing to APOBEC activity.

3.5 SNP Genotyping of Lung Cancer Tumours

SNP genotyping data (Chapter 2, Section 2.2.1) was available for 155 patients out of the total 157 patients for which DNA sequencing data (WES or TCS) was available, with only two LUADs that were DNA sequenced but not SNP genotyped.

3.5.1 GenomeStudio Data Pre-processing

The first step of analysing Illumina SNP genotyping data involved uploading the raw intensity data into GenomeStudio together with two additional files (Chapter 2, Section 2.2.3) that enable better clustering to be performed. Since the level of intensity represents the strength for one of the two alleles (A or B) of the targeted sequences to bind one of the two probes for a particular SNP, the SNP cluster algorithm allows formation of clusters that distinguish samples into AA, AB and BB and allows identification of problematic samples.

The best parameter to measure sample quality is the call rate, which measures the percentage of SNPs with genotype calls for a sample. The call rate standard generally used is 95-98%.

Most of the samples (84%) obtained a call rate of >98% with the lowest call rate value being 79.74%. Among the samples that obtained a relatively low call rate, 83% were tumour samples (Supplementary Fig. 3.1) that were of good DNA quality and anticipated to have large amounts of copy number alterations. In addition, the GenCall score quality metric, that indicates the reliability of the genotypes called, was assessed and found to positively correlate with call rate (Supplementary Fig. 3.2).

Both metrics allowed evaluation of the quality and performance of DNA samples in the experiment, and no sample was discarded or considered for pre-processing due to biological reasons at this stage.

3.5.2 QC of SNP Array Data

As mentioned in Chapter 2, Section 2.2.4, QC of SNP array data involved filtering SNPs that were classified as InDel markers, that mapped into pseudoautosomal regions or chromosome "0", as well as considering important scores representing quality metrics such as GenTrain score, cluster separation score and call frequency. From a total number of 962,215 initial SNPs, a final set of 777,193 QCed SNPs was retained and used for downstream analysis.

3.5.3 GC Bias Correction

Next the QCed SNP genotyping data was adjusted for genomic waves with ASCAT (Chapter 2, Section 2.2.5). A total of 12 samples were discarded at this stage due to noisy ASCAT plots (Fig. 3.8) based on Log R Ratio (LRR) and B-allele frequency (BAF) metrics that were used for estimation of copy number status.

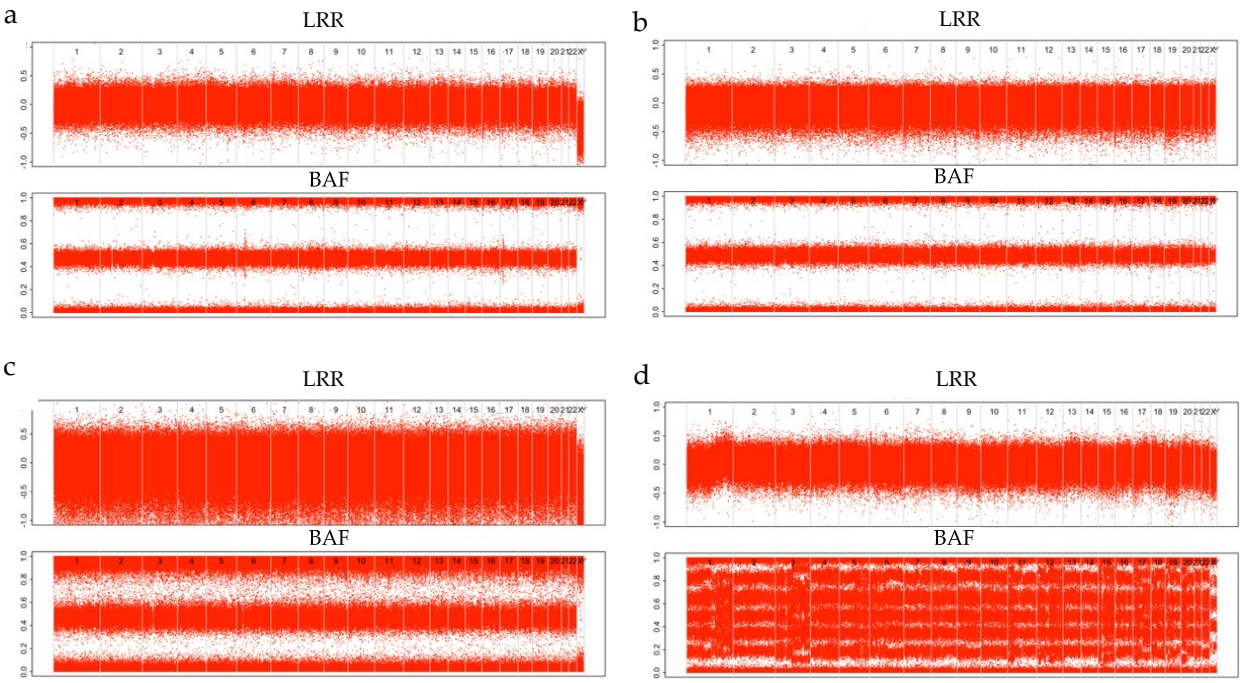


Figure 3. 8| Log R ratio (LRR) and B allele frequency (BAF) plots of four different LC samples. On the top (a-b), "good" ASCAT plots and on the bottom (c-d), two noisy samples.

3.5.4 Purity and Ploidy

Tumour purity refers to the proportion of cancer cells in the sample biopsy, whilst tumour ploidy refers to the number of sets of chromosomes in a cancer cell. There are many new computational methods that can be used to infer tumour purity. Purity variation has been observed between cancer types and can depend on both intrinsic (clinical variability) and extrinsic factors, such as how samples have been collected. In this present study ABSOLUTE was used to infer LC tumour purity and ploidy *in silico* (Chapter 2, Section 2.2.6, Fig. 3.9).

Samples had a minimum of 16% and maximum of 69% tumour purity overall across LC histotypes, with a median tumour purity of 23%. More specifically, L-CDs showed the lowest level of tumour purity with an average of 21.50%, followed by LUADs (23%), LUSCs (24.50%), and LNETs (25%). This is consistent with different patterns observed from a pan-cancer genomic study[193] in which cancers resulting from mutagenic exposures, such as LUADs and LUSCs, showed low purity levels probably due to the role of the microenvironment or the difficulty to distinguish cancer cells from the microenvironment in relation to tumour spread.

The median ploidy of the whole set of LC samples was 2.2, with L-CDs showing the lowest ploidy (2.12) followed by LUADs (2.16), LUSCs (2,47) and LNETs (3.14). Both SCLC and LCNEC have been characterised by frequent aneuploidy[44]. Indeed, the ploidy level was not significantly different between the subtypes studied here ($P = 0.3743$). Ploidy levels could also reflect a higher number of CNAs in LNETs due to an increased cell tolerance to segregation errors[194,195].

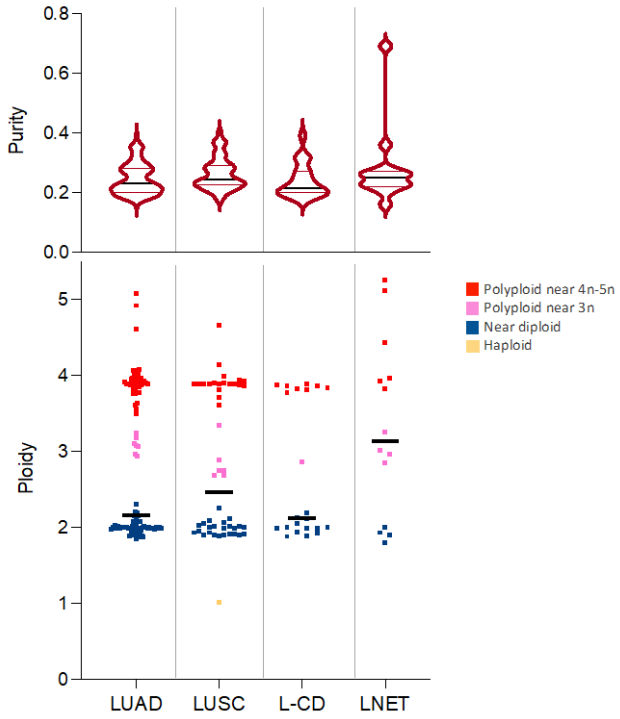


Figure 3. 9| Sample purity (top) and ploidy (bottom) across LC histological subtypes. Violin plots show the median, first quartile and third quartile of purity in each subtype. Scatter plot shows the median and ploidy category for each tumour sample across subtypes. Abbreviations: LUAD (Lung Adenocarcinoma); LUSC (Lung Squamous Carcinoma); LNET (Lung Neuroendocrine Tumour); L-CD (Lung Carcinoid).

The inferred tumour purity was tested for correlation with tumour content as determined through pathology review (Prof. A.G. Nicholson) of haematoxylin and eosin staining in samples for which data was available ($n = 87$) (Supplementary Fig. 3.3). A positive trend was observed for the L-CD histological subtype (R=0.4, $P = 0.16$), whereas a linear or negative trend was observed for LUSCs (R= -0.28, $P = 0.34$) and LNETs (R= -0.13, $P = 0.76$); with no detectable relationship observed for LUADs (R= -0.0057, $P = 0.97$). None of these trends, however, achieved significance.

3.5.5 Data Segmentation

After the GC bias correction step, data was subsequently segmented into regions of estimated equal copy number with DNAcopy (Chapter 2, Section 2.2.6) that allowed genomic regions with abnormal copy number levels to be discriminated. A total of 31,148 non-tumour segments were generated and only segments with marker support of >10 segments were considered for Copy Number calling, leaving 20,800 segments for further analysis with Gistic2. For tumour data, from a starting number of 43,553 initial segments, 30,956 segments remained after marker support-based filtering.

After obtaining segments supported by 10 markers in both tumour and normal data, focal CNA calling with Gistic2 (Chapter 2, Section 2.2.9) was performed first to identify significant CNAs in the normal samples. A total of 21 peaks were identified as significantly altered at the copy number level in the normal data (Supplementary Table 3.2). These peaks were subsequently filtered out from the tumour data to keep somatic events. After this filtering a total of 30,223 tumour segments were left for somatic calling of CNAs, with many mapping to autosomal chromosomes (92.4%).

Exploring these segments those with negative LRR values were deemed as deletions whilst those with positive LRR values were potential amplifications. Overall, more deletions (*n*=6,817) than amplifications (*n*=2,462) were identified across all subtypes (Fig. 3.10), although median sizes were larger for amplification events (Supplementary Fig. 3.4).

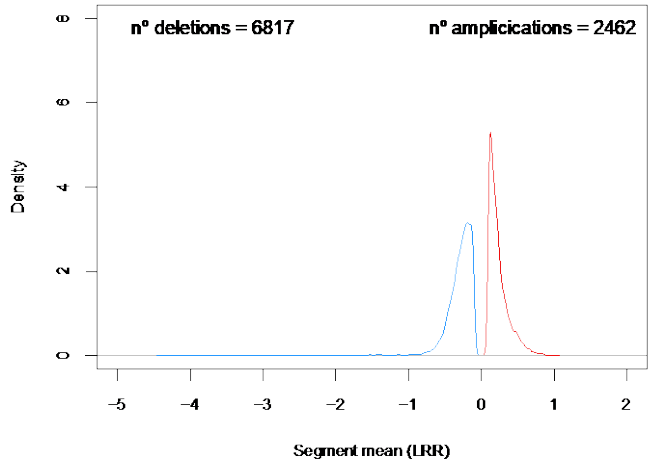


Figure 3. 10| Distribution of segment mean LRR values in putative deletions (LRR<0; in blue) and amplifications (LRR>0; in red). Segments were obtained after GC wave correction, segmentation and subtraction of significant peaks from LC normal samples. Abbreviation: Log R Ratio (LRR).

3.5.6 Copy Number Alterations (CNAs)

Known and novel somatic CNAs at genome-wide significance were identified in 96.8% (*n*=152/157) of lung tumours (Fig. 3.11). Significant *EGFR* amplifications (cytoband 7p11.2) were found in 19% of LC patients (FDR = 9.71×10^{-5}) and regions encompassing the *CDKN2A*/*CDKN2B*/*MTAP* locus were deleted in ~20% of LC patients. In addition, recurrent significant deletions were identified in *ZNF180* (34%, FDR = 1.66×10^{-40}) and *ZNF404* (29%, FDR = 5.03×10^{-7}) both located in the same cytoband (19q13.31). Furthermore, *KIF20B* located at cytoband 10q23.31 was deleted in 30% of lung tumours (FDR = 1.51×10^{-36}) and *KIAA1456* located in cytoband 8p22 in 25% (FDR = 9.17×10^{-14}). Significant amplifications were identified in cytoband 8q24.2 encompassing *MYC* (amongst other genes) in 28% of LC patients (FDR = 3.52×10^{-8}) and in cytoband 20q13.33 encompassing solely *SYCP2* (16%). A listing of the significantly altered genes at the copy number level can be found in Supplementary Data 3.1 in the Appendix.

Intergroup differences were examined at the CN level. LUADs showed frequent deletions in *ZNF180* (47.13%), *ZNF404* (41.38%), *KIF20B* (37.93%) and in a genomic stretch containing *INSL4*, *INSL6* and *JAK2* (37.78%). LUSCs showed a high number of *BCHE* amplifications (60%), *ZNF180* deletions (47.13%), *MYC* amplifications (42.85%) and *RAD50* deletions. In contrast, *EGFR* amplifications were more frequent in LUADs (30%) than in LUSCs (11.42%). For instance, *BCHE* amplifications (Fisher's exact test: $P = 2.029 \times 10^{-10}$) and *RAD50* deletions (Fisher's exact test: $P = 8.175 \times 10^{-4}$) were significantly enriched in LUSCs, whereas *EGFR* amplifications were highly frequent in LUADs although this did not achieve statistical significance (Fisher's exact test: $P = 0.091$).

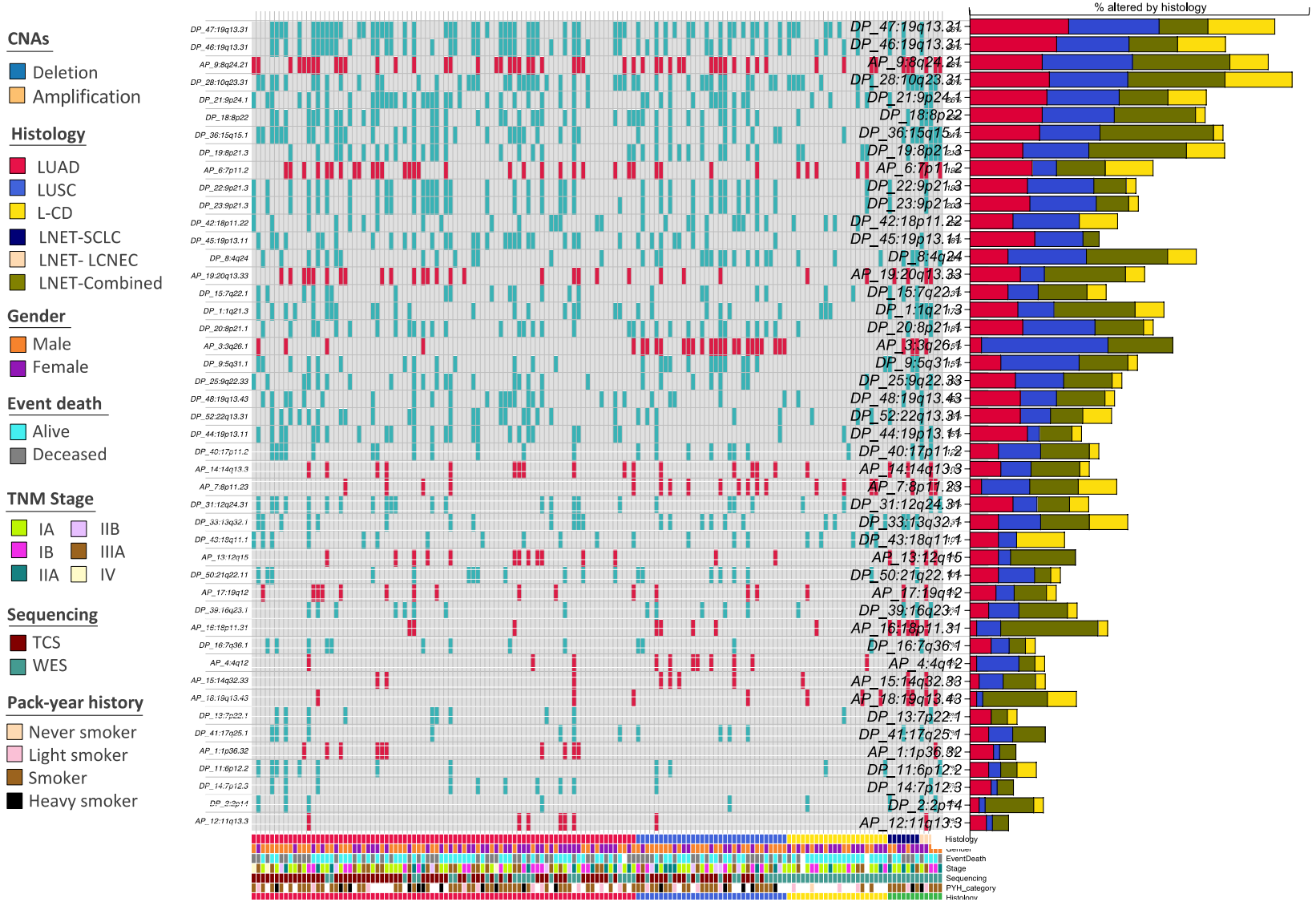


Figure 3. 11| Amplifications and deletions in significant cytobands identified with Gistic22 in 157 LC tumours. Columns represent each lung cancer (LC) patient following the same order as in Figure 3.1 and in rows are listed the 46 significant cytobands. Significance was considered when residual q-values after removing segments shared with higher peaks were <0.05. Only cytobands in autosomal chromosomes are shown. Bottom bar indicates clinical feature membership. Abbreviations: TNM, tumour, nodes and metastasis; TCS, Target Capture Sequencing; WES, Whole-Exome Sequencing; PYH, Pack-year history. Pack-year history categories: heavy smoker (>=80 cigarretes); smoker (<20,<80); light smoker (=<20); never smoker; LUAD, Lung Adenocarcinoma; LUSC, Lung Squamous Carcinoma; LNET, Lung Neuroendocrine Tumour; L-CD, Lung Carcinoid; SCLC, Small Cell Lung Cancer.

Common events shared between LUADs and LUSCs included *KIF20B* deletions, with frequencies of 37.93% in LUADs and 37.14% in LUSCs.

LNETS were the histotype with the highest number of CNAs. *MIR4310* and *SPTBN5* located in the same genomic cytoband were detected as being deleted in 53.85% of LNET tumours, followed by *MYC* amplifications (46.15%), *KIF20B* deletions (46.15%), *PDLIM2* deletions (46.15%) and *TGIF* amplifications (46.15%). Opposite to NSCLCs, *ANKRD12* was not deleted in any LNET patient.

Finally, L-CDs showed recurrent *ZNF180* and *KIFPB* deletions (31.82% for each), followed by *ZNF404* and *ROCK1* deletions (22.73% each) and *EGFR* amplifications (22.73%). Despite the very low number of somatic mutations and InDels, L-CDs harboured a high number of CNAs.

3.6 Cancer Genome Burden and Mutational Signatures

3.6.1 Tumour Mutational Burden and Copy Number Burden

Tumour Mutational Burden (TMB) and Copy Number Burden (CNB) was calculated (Chapter 2, Section 2.5.2 and 2.2.11, respectively) separately for TCS and WES samples. LNETs showed the highest TMB (with a median of 25.34 Mutations/Mb [30.54-18.82]) and CNB (18.32% of the genome under CN aberrations [35.58-2.091]). Conversely L-CD showed the lowest TMB (with 4.08 Mutations/Mb [10.18-3.67]) and CNB (4.1% of the genome under CNB [7.49-0.94]).

For NSCLCs, LUSCs show higher TMB as compared to LUADs, with 22.58 [29.80-11.97] and 17.37 Muts/Mb [25.41-11.05], respectively. Whereas CNB showed an opposite trend with LUADs showing a median of 4.26% of the genome under CNB [12.49-0.78]] and LUSCs a median of 2.96% of the genome under CNB [11.76-0.5].

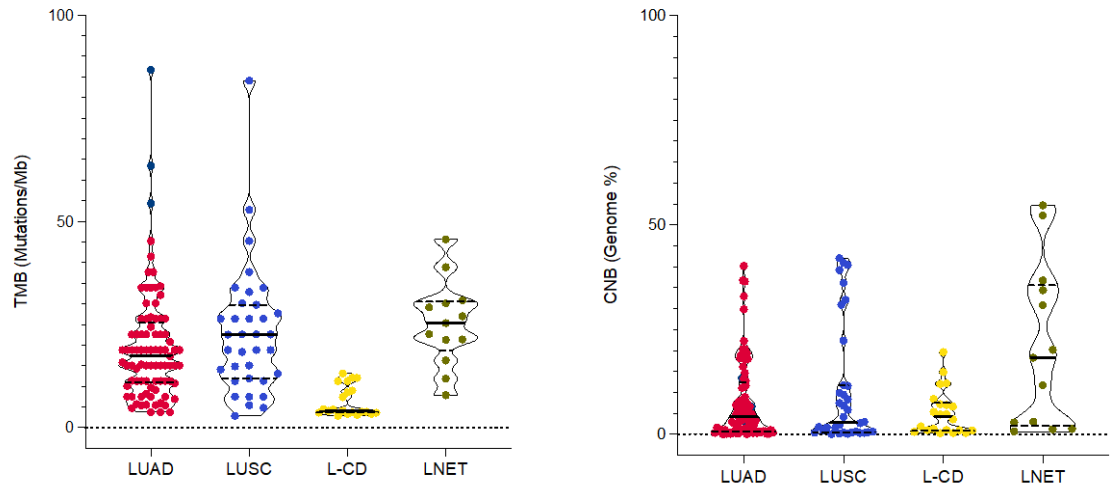


Figure 3. 12| Tumour Mutational Burden (TMB) and Copy Number Burden (CNB) detected in the four different Lung Cancer subtypes. Tumour Mutational Burden (TMB) was calculated as the total number of somatic Single Nucleotide Variants (SNVs) per Mb captured by the sequencing technology. Copy Number Burden (CNB) was calculated as percentage of genome harbouring copy number alterations. Median and quartiles are shown for each histology. Abbreviations: LUAD (Lung Adenocarcinoma); LUSC (Lung Squamous Carcinoma); LNET (Lung Neuroendocrine Tumour); L-CD (Lung Carcinoid).

3.6.2 TMB/CNB Relationship

Overall, no relationship was observed between TMB and CNB in LC samples that underwent either WES or TCS (R=0.11; *P* = 0.17). Since the number of exonic regions between the two sequencing technologies could represent an important bias, correlations between TMB and CNB were also calculated for each histology and sequencing category (Fig. 3.13). Once more no relationship was observed between TMB and CNB when examining tumour samples that were TCS (R=0.08; *P* = 0.47) or WES (R=0.17; *P* = 0.15).

Next considering the different histotypes, LUSCs showed a trend towards a positive correlation between TMB and CNB in tumour samples that underwent WES (R=0.28; *P* = 0.35) and TCS (R=0.25; *P* = 0.27), although neither of these correlations achieved statistical significance. In contrast, LUADs showed no correlation in WES tumour samples (R= -0.2; *P* = 0.34) and, to a lesser extent, TCS (R= -0.02; *P* = 0.88) again neither being statistically significant.

L-CDs and LNETs (for which only WES data was available) showed (although not statistically significant) trends towards negative and positive correlations between TMB and CNB respectively (L-CDs: R= -0.18; *P* = 0.43 and LNETs: R=0.079; *P* = 0.82).

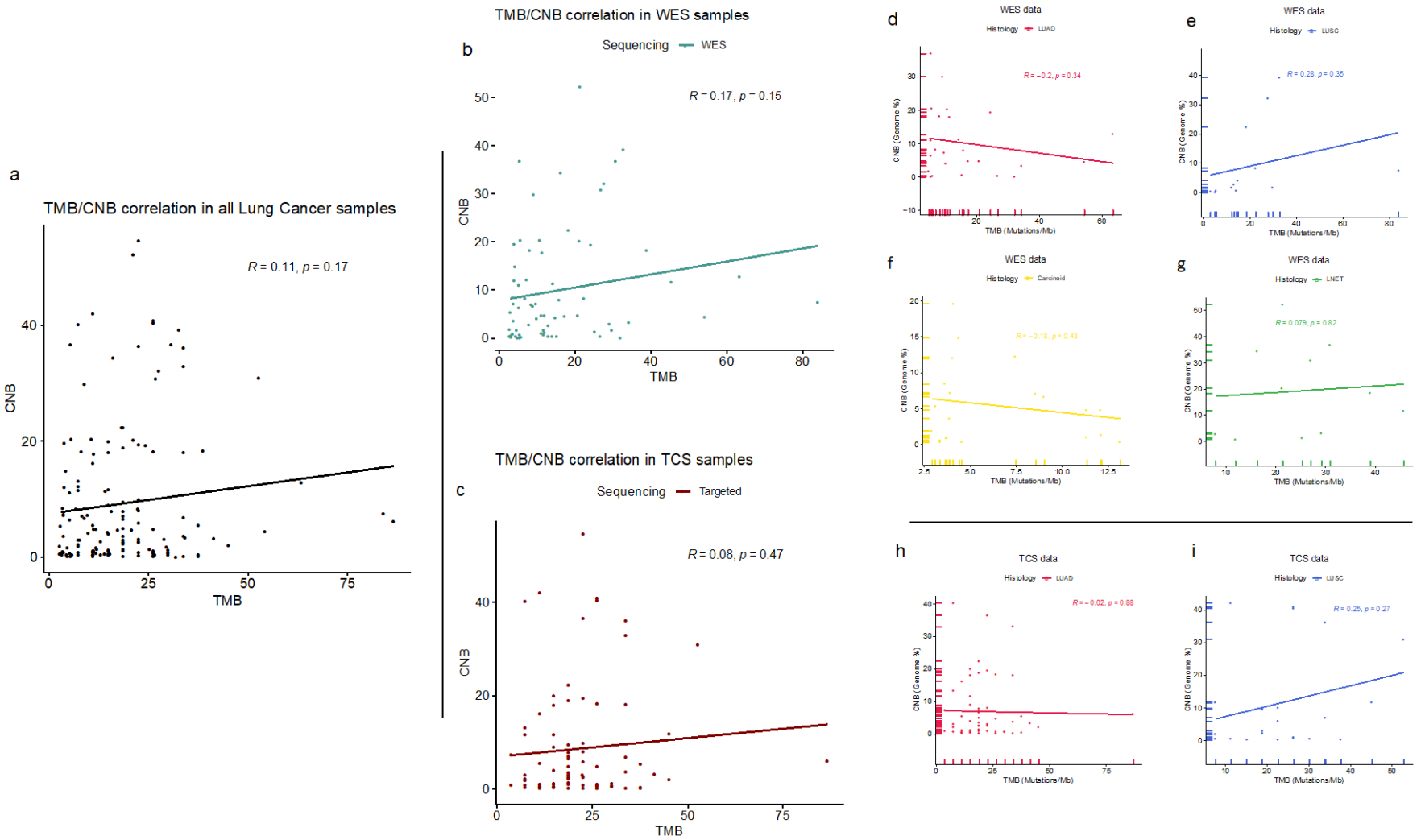


Figure 3. 13| TMB/CNB correlations of Lung Cancer tumour samples. a) All Lung Cancer samples by WES and TCS. b) Samples sequenced by WES; c) samples sequenced by TCS; d) LUAD samples sequenced by WES; e) LUSC samples sequenced by WES; f) L-CD samples sequenced by WES; g) LNET samples sequenced by WES; h) LUAD samples sequenced by TCS; i) LUSC samples sequenced by TCS. Abbreviations: TMB, Tumour Mutational Burden; CNB, Copy Number Burden; WES, Whole Exome Sequencing; TCS, Targeted Capture Sequencing; LUAD, Lung Adenocarcinoma; LUSC, Lung Squamous Carcinoma; LNET, Lung Neuroendocrine Tumour; L-CD, Lung Carcinoid.

3.6.3 Hallmark Signalling Pathways

As mentioned previously (Chapter 1, Section 1.2.2) several cancer hallmarks have been associated with genes and biological pathways. The similarities and differences in ten canonical oncogenic signalling pathways genetically altered in cancer were therefore explored for the set of 157 LC tumours that mutational and copy number data were available for (Fig. 3.14).

For the ten pathways, most LC histological subtypes showed a high number of recurrent mutations and CNAs, the exception being L-CDs which did not show alterations in four of hallmark pathways. The latter were the PI3K pathway, TGF-beta, WNT and NRF2 signalling pathways and this finding suggests there may be alternative carcinogenic mechanisms leading to tumorigenesis in this rarer cancer subtype.

Genes of the TP53 pathway and cell cycle pathways were recurrently altered across almost all LC subtypes except for L-CDs, which only appeared altered in 4.5% of tumours at the CN level (and not by somatic mutation or InDel) for the *CDKN2A* gene located on chromosome 9. The same gene was altered at high frequencies by both mutations and CNAs in LUSCs and LUADs and LNETs with frequencies of 60%, 39.3% and 15.4%, respectively. *CDKN1A* on the other hand was similarly altered in LUADs (34.6%) and LUSCs (38.5%). *ATM* was mutated only in the LUAD histotype at a frequency of 15.4%, whereas *FBXW7* appeared mutated in 23.1% of LUADs and 8.3% of LNETs.

Another gene of the cell cycle pathway, *MYC*, was recurrently altered gene in all LC subtypes with LNETs being the histological subtype with the highest number of alterations (42.9%), followed by LUSCs (42.9%), LUADs (34.8%), and L-CDs (18.2%). NSCLC tumours also harboured alterations in *MGA* and *MXI1* genes of the same pathway but at lower frequencies (15.4% and 7.7% for LUSCs and 7.7% and 3.8% in LUADs, respectively). *MYC* amplification and overexpression in lung cancer patients represents a potential opportunity for inhibiting a key player in tumour progression, since in lung tissue it specifically acts as a transcriptional coordinator for proliferation and tissue regeneration[196].

The RTK-RAS pathway was also substantially altered at the genetic level across all four LC subtypes with mutations and CNAs in *EGFR* and *FGFR1* genes, with *EGFR* being more prevalent in LUADs (36%) whilst *FGFR1* appeared highly altered in LUSCs (61.5%). The

KRAS member of the same pathway was altered in 29.2% of LUADs. Other altered genes were *ROS1*, *ERBB4* and *ALK* that, in addition to being altered in NSCLCs, were also altered in LNETs with frequencies of 15.4%, 7.7% and 7.7% respectively. LNETs also showed mutations in *MAP2K2* (8.3%).

NOTCH2 was the gene most frequently altered amongst the Notch pathway with frequencies of 46.2% in LUADs, 33.3% in LNETs and 4.5% of L-CDs. Interestingly, *NOTCH2* was not found to be mutated or amplified in LUSCs. This is consistent with recent studies showing the prognostic value of this gene in LUADs rather than in LUSCs, with the former showing more positive staining[196]. The relevance of *NOTCH2*, however, has been suggested to be masked due to its concomitant activation of *NOTCH1*[197]. Nevertheless, *NOTCH1* showed a higher number of mutations in LUSCs (11.4%) as compared to LUADs (4.5%), suggesting different carcinogenic roles between the two different NSCLC histotypes.

The WNT pathway showed alterations in five different genes of the Frizzled Class Receptor (FZDs) family, namely *FZD4*, *FZD10*, *FZD5* and *FZD9*. FZDs showed mutations in 25% of LNETs, whilst they had low mutation rates in LUADs (3.8%) and LUSCs (7.7%). Moreover, *APC*, another member of the WNT pathway, was equally mutated in LUADs and LUSCs with a frequency of 7.7% for both LC subtypes.

The NRF2 signalling pathway associated with oxidative stress responses was only altered in NSCLC, with mutations in *KEAP1* in 17.1% of LUSCs and 13.5% of LUADs. In addition, *NFE2L2* (also known as *NRF2*) was more frequently mutated in LUSCs (11.4%) than in LUADs (1.1%). KEAP1 negatively regulates NFE2L2 under unstressed conditions but when stress occurs it promotes the transcriptional activation of several genes involved in a broad of cytoprotective mechanisms. Consequently, developing drugs that promote its activation could be a potential treatment specially for chronic diseases where different aetiologies are involved. Thus, the role of NRF2 in cancer is a matter of current intense research[103].

Finally, several genes of the Hippo pathway were found mutated in LC subtypes at varying frequencies. *FAT4* was the only gene of this pathway being altered in all four histotypes: 16.7% of LNETs, 7.7% in both LUADs and LUSCs and 4.5% of L-CDs. Other tumours carried mutations in *FAT3*, with frequencies of 16.7% of LNETs, 15.4% of LUSCs and 11.5% of LUADs, with no mutations present in any L-CD tumour. *FAT1* and *FAT2* also were identified as mutated in LUSCs only each by 15.4%. Other recurrently mutated genes were

DCHS1, altered in 16.7% of LNETs, encoding a transmembrane cell adhesion molecule; and *CRB1*, altered in 15.4% of LUADs, which has recently been suggested to play an important role in tumour progression in a cancer-type specific manner[198].

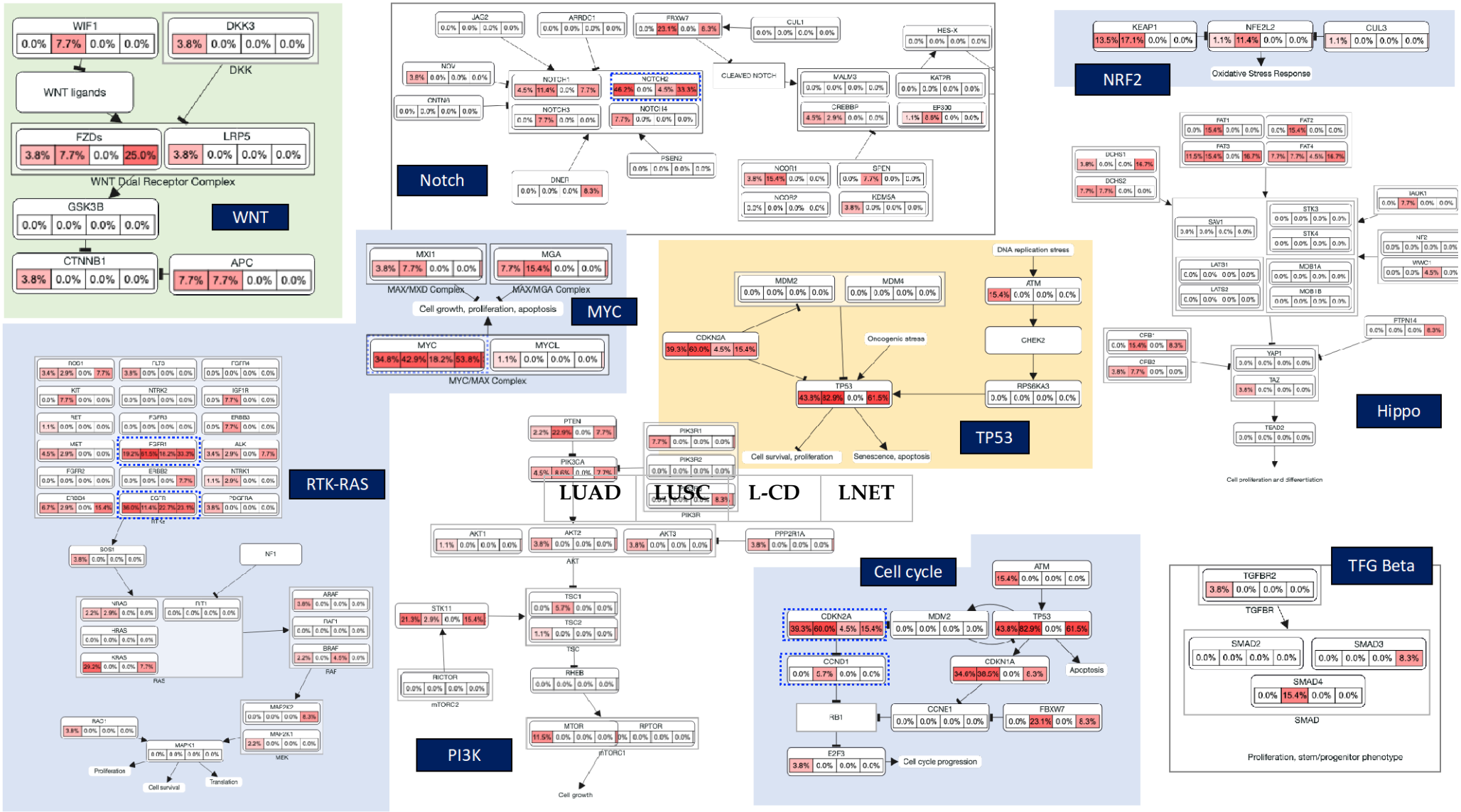


Figure 3. 14| Genes in ten hallmark cancer pathways altered in Lung Cancer histotypes. Red colour intensities indicate the average frequency of alteration by mutations, InDels and significant copy number alterations (CNAs) within the entire dataset. Genes within blue dotted boxes show genes altered by both mutations and significant CNAs. Abbreviations: LUAD, Lung Adenocarcinoma; LUSC, Lung Squamous Carcinoma; LNET, Lung Neuroendocrine Tumour; L-CD, Lung Carcinoid.

3.7 Discussion

The identification of somatic genetic alterations, including point mutations and CNAs, has greatly accelerated the understanding of cancer biology and is key for the identification of novel therapeutic targets.

In contrast to WES, TCS focused on a panel of genes known to be frequently altered in lung cancer. First, WES data was validated by using targeted sequencing on the NextSeq 550 platform in 10 NSCLC tumour and normal paired samples. This initial validation ensured that the gene panel could be successful used for the identification of somatic mutations, and that high coverage could be achieved by TCS. Moreover, the detection of two additional mutations through the gene-panel based targeted sequencing emphasised the accuracy of this study design. A possible explanation for this is the low coverage associated with WES compared to TCS approaches. Indeed, both WES and TCS showed similar frequently altered genes in NSCLCs. In addition, human lung cancers are characterized by a high number of somatic gene alterations, including mutations, copy number changes, and translocations. Consistent with this, in this study both LUAD and LUSC subtypes have shown a high tumour mutational burden, and specifically LUSCs showed more mutations per mega base than LUADs.

The WES data in this present study allowed the patterns of single base substitutions associated with particular mutational processes to be identified. Both known and *de novo* mutational signatures were identified in both the common and rare LC subtypes. In line with the different genetic alterations found in the L-CD histotype, the spectrum of mutational signatures also appeared different for this subtype with CMS 3 (signature associated with failure of DNA Double-Strand Break (DSB) repair by Homologous Recombination [HR]) being identified more frequently in these tumours. Failure of DNA double-strand break repair by homologous recombination has been found in breast, ovarian and pancreatic tumours. DNA mismatch repair deficiency generally leads to hypermutation but here a lack of recurrent mutations was found for this LC histotype. DNA damage and DNA repair jointly shape mutagenesis and alter mutation rates.

Missense mutations in the *TP53* gene were the most recurrent genetic alteration across LC subtypes and the gene appeared as one of the topmost mutated genes detected by both WES and TCS. *TP53* was consistently associated with LUSC histology in WES and TCS data,

highlighting the potential of targeted NGS for accurate detection of cancer-related genes. Different genes of the PI3K/AKT/mTOR signalling pathway were identified and significantly associated with NSCLCs histotypes with *STK11* appearing highly mutated and associated with LUAD histology, and *PTEN* for LUSC.

Additionally, the CUB and Sushi Multiple Domain 3 gene (*CSMD3)* appeared associated with LNET histology. In line with this observation, *CSMD3* has been reported as frequently mutated in lung cancers and loss in its functionality has been found to be associated with proliferation of airway epithelial cells[199]. Furthermore the gene has been found enriched in poorly differentiated neuroendocrine carcinomas compared to the more differentiated neuroendocrine tumours[200], strengthening the need for further investigation for *CSMD3* role and its therapeutic inhibition.

A key finding is the lack of recurrent mutations in L-CDs, a histotype that shows significantly longer survival time compared with patients with tumours of other histologies (*P*=0.0148; Ratio =2.457; 95% CI 1.192-5.062). Studies have historically focused on the identification of somatic mutations that are altered at high frequencies in oncogenes and TSGs, as these mutations could be potential drivers of cancer. In this present study, both mutations and CNAs were identified in the four different lung cancer subtypes (histotypes) and have enabled the identification of recurrent genes altered at the copy number level in L-CDs. Importantly, in the latter, recurrent *ROCK1* deletions and *EGFR* and *MYC* amplifications were found. Existing therapeutic agents for these three alterations have been identified in other cancers and the data from this study suggest that the agents could have potential therapeutic benefit for L-CD patients.

In this present study, different spectrums of mutational signatures between the different LC subtypes were also observed. Specifically, tumours of LUAD, LUSC and LNET histology all showed dnCMSs associated to tobacco smoking (CMS 4) and the activity of AID/APOBEC enzymes (CMS 13), causing C>T and C>A substitutions respectively. Importantly these signatures showed low weights in the L-CD histological subtype suggesting distinct mutational processes underlying the development of cancer, with potential implications for L-CD prevention and treatment. Concomitantly, the identification of CSM 3 in these rarer tumours (a signature associated with failure of DSB repair by HR) suggests an imbalance between DNA damage and repair in this LC subtype. DSBs can arise

from endogenous sources including reactive oxygen species generated during cell metabolism, collapsed replication forks and nucleases, as well as from exogenous sources like Ionizing Radiation (IR), chemical agents and Ultraviolet (UV) light that cause replication blocking lesions. These lesions are sensed by different proteins that trigger cell cycle arrest and activation of DNA repair pathways. Not all genotoxins or DNA repair deficiencies lead to unique mutational signatures[201], hence the importance of recognising the variable nature of mutagenesis. The results of this study suggest that DSB repair proteins and its regulators[202] could be used for the development of more effective chemo- and radiotherapeutic strategies for L-CD patients.

Mutational signature CMS 29 (associated with tobacco chewing) was also observed at high frequencies in nearly all LC subtypes, the exception being LUSC, suggesting different lifestyle exposures for tobacco-smoke associated damage (C>A).

Furthermore, unknown mutational signatures were detected with high weights hence *de novo* mutational signatures (dnCMSs) investigation was performed. Tumours of LUAD, LUSC and LNET histology all showed dnCMSs that were similar to tobacco smoking (CMS 4) and the activity of AID/APOBEC enzymes (CMS 13) both causing C>A and C>T substitutions, although the latter type of substitution is more typically caused by APOBEC deamination. Interestingly, LUADs and L-CD commonly showed dnCMSs associated with spontaneous deamination of 5-methylcytosine in CG motifs. The latter has been found in most cancer types and has been correlated with age[203]. This finding however is surprising in the case of L-CD, as the L-CD patients were the youngest in the present study. Nevertheless, the low TMB associated with this L-CD histotype may be an indication that non-clock-like mutational processes do not prevail in front of the natural acquisition of mutations during a lifetime. Moreover, this signature is also in line with the observed CMS 20 associated to MMR in these tumours, as different rates of mismatch repair could explain the observed mutational process because DNA replication without previous repair will convert T•G mismatches into C>T mutations. This has been suggested to be an indication of the number of cell divisions experienced serving as a "clock". This is consistent with the fact that L-CDs are well differentiated tumours and that some carcinogens contribute to cancer development by stimulating cell proliferation with a tumour promoter to stimulate proliferation of altered cells, rather than by inducing mutations[204]. Chemicals, radiation, and viral infections are

carcinogens that can be stimulated by tumour promoters, do not necessarily cause mutations, but at stable and prolonged rates can induce malignant transformation. Further research with a bigger sample size may be needed to confirm with more confidence the mutational processes underlying this L-CD histotype.

Using the SNP genotyping data in this study, ploidy levels were found to be the highest for the LNET histotype consistent with prior literatiure[205]. *SPTBN5* deletions were the most frequent CNA in LNETs and significantly enriched when compared to the other histotypes (P= 0.044). The gene encodes spectrin protein, levels of which have been shown to have cytoplasmatic positivity in carcinoids[206]. In the present study, *SPTBN5* deletions were detected in only 4.55% of L-CDs indicating a potential marker for the differential diagnosis of L-CDs.

BCHE amplifications (Fisher's exact test: P= 2.029x10^{-10}) and *RAD50* deletions (Fisher's exact test: P = 0.0025) were significantly enriched in LUSCs (Section 3.5.6) in line with previous findings[207]. *EGFR* amplifications were highly frequent in LUADs although this enrichment was not statistically significant relative to other histotypes (Fisher's exact test: P = 0.091). *BCHE* amplifications have already been frequently observed in lung squamous carcinoma[207], as well as in leukaemia[208] and ovarian carcinomas[209].

With the data generated in this chapter, genetic alterations in ten hallmark signalling pathways associated with genes that control critical biological processes were explored for the different LC subtypes by integrating mutational and CN data thereby allowing interpretation of the identified alterations. As for many other cancer types, pathways regulating cell cycle progression, survival, proliferation and apoptosis were found to be recurrently altered across LC subtypes especially in LUADs, LUSCs and LNETs, whereas L-CDs showed a lack of mutations in these important pathways. This is in line with the observed overall longer survival in L-CD patients as compared to the other histotypes (P = 0.0148). Furthermore, no genetic alteration was observed for the L-CD histotype in the PI3K, TFG-beta, WNT and NRF2 signalling pathways. This suggests the existence of alternative carcinogenic mechanisms leading to tumorigenesis in this rarer cancer subtype. In contrast, *MYC* was recurrently altered gene in all LC subtypes and thus represents a potential therapeutic opportunity for inhibiting a key player in both NSCLC and neuroendocrine lung tumours. Thus, further research may be beneficial for treat LC patients with Myc overexpression.

Consistent with previous research (Chapter 1, Section 1.2.2), all the hallmark oncogenic signalling pathways were found in the present study to be highly altered in LUAD, LUSC and LNET tumours but interestingly not in L-CDs. Here the integration of both mutations and larger amplifications and deletions have allowed increased accuracy in identifying different mechanisms leading to tumorigenesis and the key genes affecting these molecular processes in common and rarer LC subtypes. These findings provide novel insights into the genetic alterations and the cellular pathways commonly and differentially affected between LC histotypes. This potentially opens up new avenues for both biomarker selection and treatment.

Chapter 4: Epigenomic Landscape of Lung Tumours

4.1 Introduction: The DNA Methylomes of Cancer

Few clinical phenomena can be explained by a single, fully penetrant genetic lesion alone[210]. In 1942 Waddington [211] first proposed that acquired phenotypic changes could become heritable under certain conditions. He observed that fly wing phenotypes induced by a heat-shock treatment during development were passed through generations in the absence of any further treatment. This observation led him to coin the concept of epigenesis.

Nowadays, we also know that every cell type in unicellular and multicellular organisms relies on different gene expression programmes. This differential activation and repression of specific genes is interposed between the genotype and the environment, the latter exerting an influence through epigenetic regulation[107]. This fine control is facilitated by the organization of the genetic material[212], which is malleable and depends on intrinsic and external factors that define cell differentiation, development, physiology and ultimately phenotype and behaviour[213]. To add another layer of complexity, every cell type has its own epigenome. Thus it can be said that humans as multicellular organisms display a single, unique genome but many epigenomes[110].

The accumulation of somatic mutations[214–216] and the progressive alteration of the epigenetic landscape and nuclear organisation[217,218] are at the basis of health and disease. Epigenetic changes can cause activation or silencing of TSGs, potentially leading to cancer. Understanding the epigenetic changes that occur in cancer offers a new layer of information that must be integrated with genetic alterations to provide a full picture of the molecular alterations that drive the cancer phenotype.

One of the most studied epigenetic alterations is DNA methylation[219]. In vertebrates, DNA methylation mainly occurs at cytosines in a CpG dinucleotide context. With the advent of NGS technologies, genome-wide DNA methylation maps have been obtained at a high resolution, allowing comprehensive accurate and quantitative estimates of DNA methylation levels across the genome to be obtained[220].

Lung Cancer (LC) is a set of heterogeneous diseases, where both genetic and epigenetic alterations have been implicated in their development and progression[119,221]. Nevertheless, a genome-wide comparison of the DNA methylation changes that typify Non-Small Cell Lung

Cancers (NSCLC) and Lung Carcinoids (L-CDs) is missing and might further elucidate the cancer biology behind LCs, ultimately improving stratification and therapy for these patients.

4.2 Research Samples Demographics

Out of the 159 LC samples that were DNA sequenced (Table 3.1), a subset of 79 LC tumours underwent Whole-Genome Bisulfite Sequencing (WGBS) to generate a snapshot of the DNA methylation profiles in tumour and matched unaffected tissue (Chapter 2, Section 2.6.2). Specifically, all L-CD pairs (Chapter 3, Section 3.1), 23 NSCLCs pairs, and a NSCLC and LNET unpaired samples were bisulfite sequenced.

Specifically, WGBS was performed for tumour samples from 40 patients (Chapter 2, Section 2.6.2). Out of these patients, 18 had Lung Adenocarcinoma (LUAD), 15 patients had Lung Carcinoid (LC), 6 patients had Lung Squamous (LUSC) and 1 patient had combined Small Cell and Large Cell Carcinoma (LNET). A set of 17 male NSCLCs who had showed a loss of Y chromosome expression[1] will be discussed in detail later in Chapter 5.

The analysed samples had a tumour content varying between 25 and 95%. Patients' age ranged between 28 and 82 years (mean 65.68 years-old; standard deviation (SD) 12.14), with L-CD patients being the youngest (mean 60.73 years-old; SD 16.74). Whilst L-CD patients were significantly younger than NSCLC ($P = 0.0437$), this difference appeared to be driven by a minority of patients (Fig. 4.1).

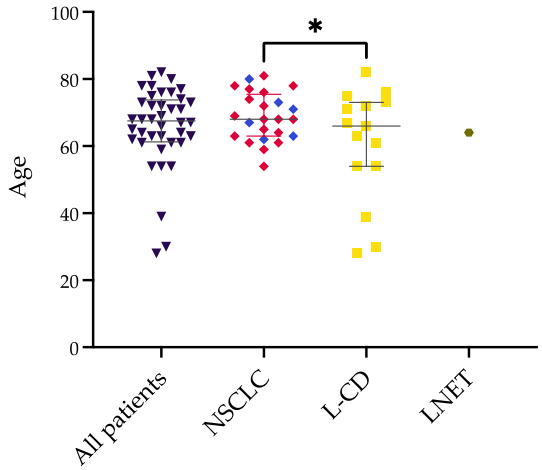


Figure 4. 1| Age of Lung Cancer (LC) patients whose samples underwent Whole Genome Bisulfite Sequencing (WGBS). Graph shows age (in years) distributions for all patients (n=40). Median and interquartile range (IQR) of age are shown for each LC histotype. Abbreviations: NSCLC, Non-Small Cell Lung Cancer (includes Lung Squamous Carcinoma [LUSC] and Lung Adenocarcinoma [LUAD]); L-CD, Lung Carcinoids; LNET, Lung Neuroendocrine Tumours.

Gender proportion was also significantly different (Fisher's exact test, $P = 4.774 \times 10^{-5}$) between NSCLC and L-CD patients. Almost three quarters (73%) of L-CD patients were female, compared with only 8.3% of NSCLC patients. Those patients with LUSCs and NETs were all male.

4.3 Assessment of WGBS Sequencing Depth and Coverage

A key consideration in genomic analyses are sequencing depth and coverage (Fig. 4.2). Assuming that reads are randomly distributed across the genome, sequencing depth is the average number of times that a nucleotide is covered by a high-quality aligned read from a sequencing experiment. Sequencing breadth or coverage denotes the percentage of target bases that are sequenced at a given depth.

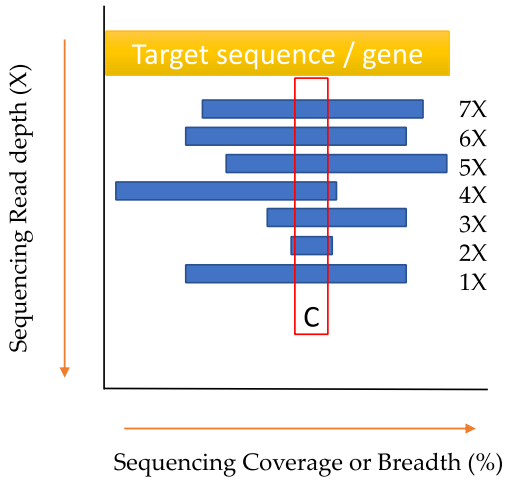


Figure 4. 2| Sequencing read coverage or breadth and sequencing read depth. Sequencing breadth or coverage (on the *X* axis) denotes the percentage of target bases (%) that have been sequenced at a given depth; while sequencing read depth (on the *Y* axis) is defined as the average number of times that a nucleotide is covered by a high-quality aligned read (X coverage) from a sequencing experiment.

To exclude sequencing performance as an explanation for between-sample methylation differences genome-wide coverage differences amongst the full 79 samples were looked for. This served as a QC before calling Differentially Methylated Regions (DMRs). For this purpose, mosdepth (Chapter 2, Section 2.6.6) allowed read sequencing depth from WGBS BAM files to be assessed. Specifically, the proportion of total bases that were covered for at

least a given coverage value (cumulative distribution) was obtained for each sample and genomic coordinate.

As evident in Figure 4.3, tumour and normal matched samples showed similar coverage, of ~30X on average for at least 40% of the genome. Thus, it could be concluded that sequencing performance was good and most importantly not different between tumour and matched normal lung tissue samples. Under this assumption, differences in DNA methylation could then be investigated confidently without any findings being attributable to technical confounders.

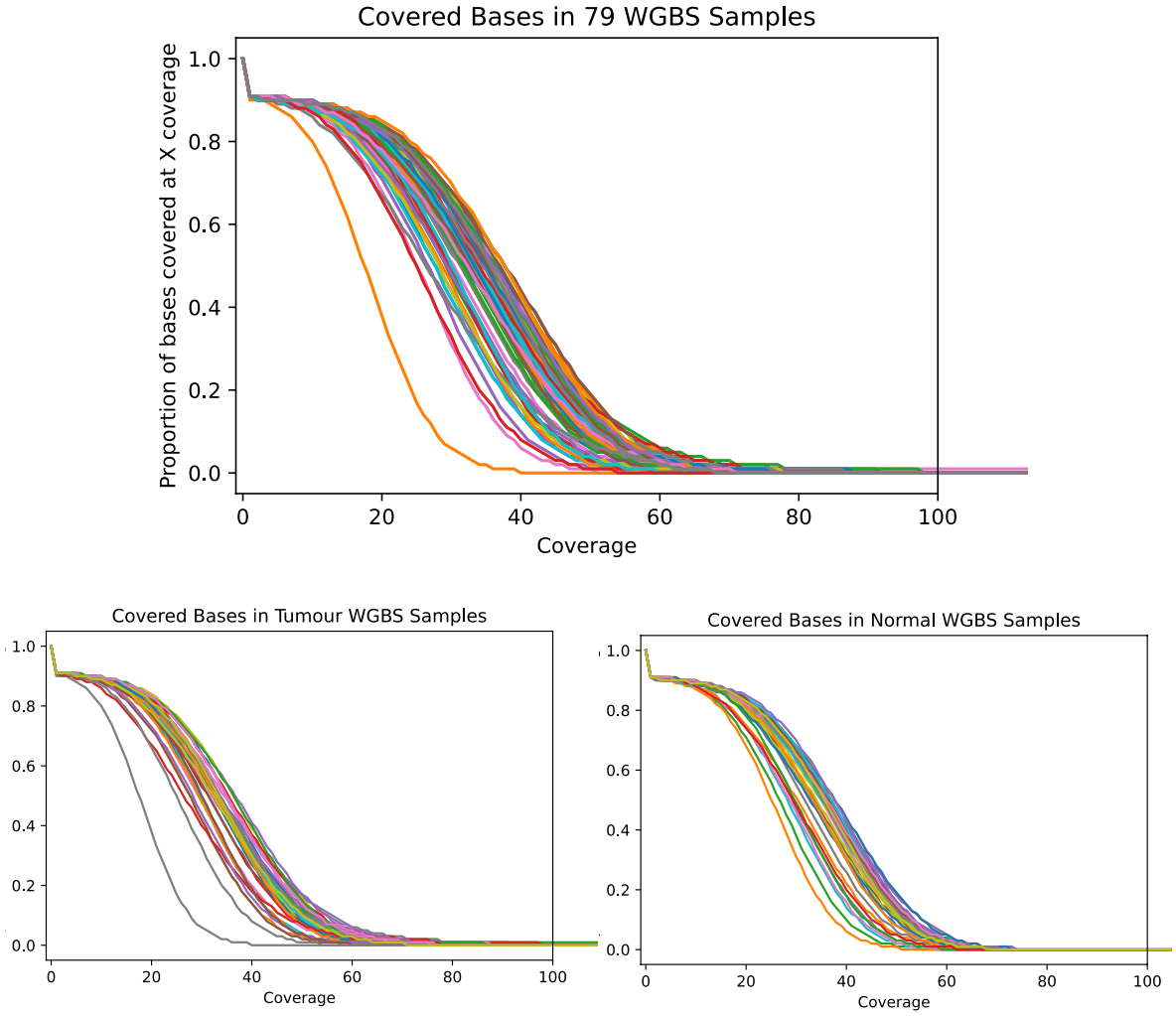


Figure 4. 3| Cumulative distribution of covered bases of whole genome bisulfite sequenced (WGBS) samples. a) All 79 WGBS samples; b) Tumour samples; c) Normal samples. Tumour and normal matched samples show similar coverage of >27X on average for most of the genome.

4.4 Pre-processing of WGBS Data

A median per-base coverage of ~27X was obtained for all 79 WGBS samples after normalization and filtering by read coverage (>10X and <99.9th percentile). The next step involved merging of CpG DNA methylation data for all the samples to enable different comparative analyses to be performed.

4.5 Graphical Representations of High-dimensional Methylation Data

For the purpose of exploratory analysis, and to determine groups for differential methylation comparisons, a Principal Component Analysis (PCA) analysis was performed based on per base CpG DNA methylation. As can be seen in Figure 4.4, PCA of the WGBS data distinguished the different lung cancer histological subtypes as well as tumour from healthy tissue. The normal tissues had similar DNA methylation patterns irrespective of the cancer patient subtype.

Notably the combined small cell and large cell neuroendocrine sample (LNET) clearly separated from the L-CDs and showed more similarity to NSCLCs based on their DNA methylation profiles.

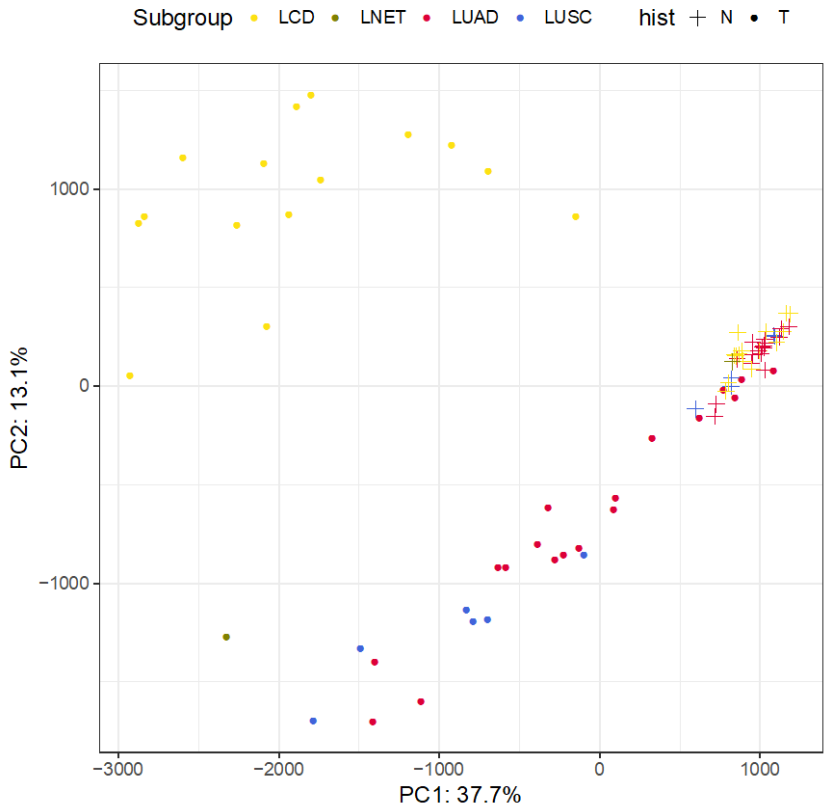


Figure 4. 4|Principal Component Analysis of whole genome CpG DNA methylation data differentiated L-CD and NSCLC tumours and tumour samples from healthy matched tissue. L-CD tumours are shown in yellow; NSCLC tumours are shown in red for LUAD and blue for LUSC; and LNET is shown in khaki. Tumour or normal matched tissue are shown as dots and crosses, respectively. Abbreviations: L-CD, Lung Carcinoids; NSCLC, Non-Small Cell Lung Cancer (includes Lung Squamous Carcinoma [LUSC] and Lung Adenocarcinoma [LUAD]); LNET, Lung Neuroendocrine Tumours.

4.6 Genomic Binning of DNA Methylation Data by Annotations and Chromatin States

PCA was able to simplify the complexity of the WGBS data, while retaining trends and patterns, and alone distinguished the different LC histological subtypes as well as tumour from healthy tissue. The human genome, however, is extensive and complex. The next aim therefore was to investigate which genomic regions and chromatin states carried most of the DNA methylation variance.

To achieve this the genome was binned into functional classes and the first two principal components (PC1 and PC2) examined to infer the regions in which most of the between-group variance was concentrated.

Specifically, the PCs obtained by analysing four different subsets of samples were compared: 1) all 79 samples with tumour and normal WGBS data from NSCLC, L-CD and 1 LNET; 2) a subset containing all tumours except the LNET, 3) a subset containing NSCLC tumour and normal data, and 4) a subset containing L-CD tumours and normal WGBS data. The results of the analyses are summarised in Table 4.1.

In the first subset, most of the variance was explained by the epigenetic mark H3K36me3 (27.63%), repeat elements (26.58%) and H3K9me3 (26.54%). In the second subset containing both NSCLC and L-CD tumours, H3K36me3 similarly accounted for most of the variance (34.06%), followed by H3K9me3 (30.3%). Between NSCLC tumours and healthy tissue, enhancers (22.82%) and H3K27me3 (22.72%) accumulated most of the variance, whereas H3K36me3 (39.71%) and H3K9me3 (38.75%) concentrated most of the variance in L-CDs.

These data suggest that epigenetic marks account for most of the variance in DNA methylation and indicate that distinct DNA methylation landscapes exist both between NSCLCs and L-CDs, and between tumours and their normal matched tissues. The high percentage of variance at PC1 and PC2 for L-CDs across the different genomic categories were also suggestive of a greater level of dysregulation at the epigenetic level in L-CD tumours as compared to NSCLCs. This observation was reinforced by the high percentage of variance accounted for by the two first PCs when comparing NSCLC against L-CD tumours.

	All 79 samples		NSCLC and L-CD *(Tumours)*		NSCLC *(Tumour and Normal)*		L-CD *(Tumour and Normal)*	
	PC1	*PC2*	*PC1*	*PC2*	*PC1*	*PC2*	*PC1*	*PC2*
Promoter	19.9	6.34	23.5	9.76	11.94	8.432	29.68	10.77
Introns	24.97	4.57	30.1	6.08	12.03	6.19	36.57	8.25
Exons	23.52	4.27	27.75	6.24	11.95	5.601	32.66	7.862
Repeats	26.58	3.11	30.25	5.68	13.66	6.691	37.88	8.381
Enhancers	24.37	10.89	28	13.16	22.82	10.9	34.58	16.71
CGI	19.46	7.56	23.66	9.83	14.15	8.408	27.95	11.68
LncRNA	24.46	6.34	26.73	8.99	14.95	9.299	36.16	11.26
CTCF	23.78	5.22	28.06	7.47	12.11	7.35	34.04	8.723
H3K4me3	13.12	11.47	17.62	15.61	17.03	6.79	21.52	18.7
H3K27ac	18.96	8.92	23.39	11.98	10.8	8.58	27.89	13.98
H3K27me3	14.59	13.15	20.29	15.08	22.72	6.63	21.17	20.63
H3K36me3	27.63	4.06	34.06	5.27	13.18	6.135	39.71	8.14
H3K4me1	18.79	9.44	22.98	13.01	13.52	8.455	27.91	15.2
H3K9me3	26.54	7.73	30.3	10.33	17.58	9.971	38.75	11.41

Table 4. 1| Percentage of variance explained by the two first Principal Components (PC1 and PC2) based on CpG DNA methylation data at different genomic regions in the four subsets of WGBS data. Abbreviations: PC, Principal Component; L-CD, Lung Carcinoids; NSCLC, Non-Small Cell Lung Cancer (includes Lung Squamous Carcinoma [LUSC] and Lung Adenocarcinoma [LUAD]); LNET, Lung Neuroendocrine Tumours; CGI, CpG Island; LncRNA, Long non-coding RNA; CTCF, CCCTC-binding factor; H3, histone H3; K, lysine residue; me, methylation at the indicated lysine residue.

4.7 Analysis of Differentially Methylated Regions

4.7.1 Descriptive Statistics of Samples

Median CpG DNA methylation percentage was calculated for each tumour and unaffected tissue from LC patients from different histological subtypes. For this analysis, the focus was on the NSCLCs and L-CDs histological types (n=40), with the aim of identifying differences in CpG DNA methylation between these tumours and their normal counterparts, as well as inter-tumour differences for NSCLC and L-CD tumours. LNET data, due to their limited availability, are displayed in Figure 4.5 for context only. To limit the confounding influence of sex, differential methylation in the autosomes only was examined.

Autosomal DNA methylation levels were significantly reduced in tumour samples from L-CD ($P = 4.136 \times 10^{-5}$) and NSCLC ($P = 1.694 \times 10^{-8}$) histologies as compared with paired normal lung tissues (Fig. 4.5). Global DNA methylation loss is a common feature of cancer and ageing that has been generally observed in cancers with genomic instability and in cancers with rather stable genomes, such as Chronic Lymphocytic Leukaemia (CLL) or renal clear cell carcinoma (KIRC). In this present study, a relative reduction in DNA methylation levels genome-wide in NSCLC tumours with high TMB and CNB, and in L-CDs rather harbouring only a few alterations at the genetic level was detected.

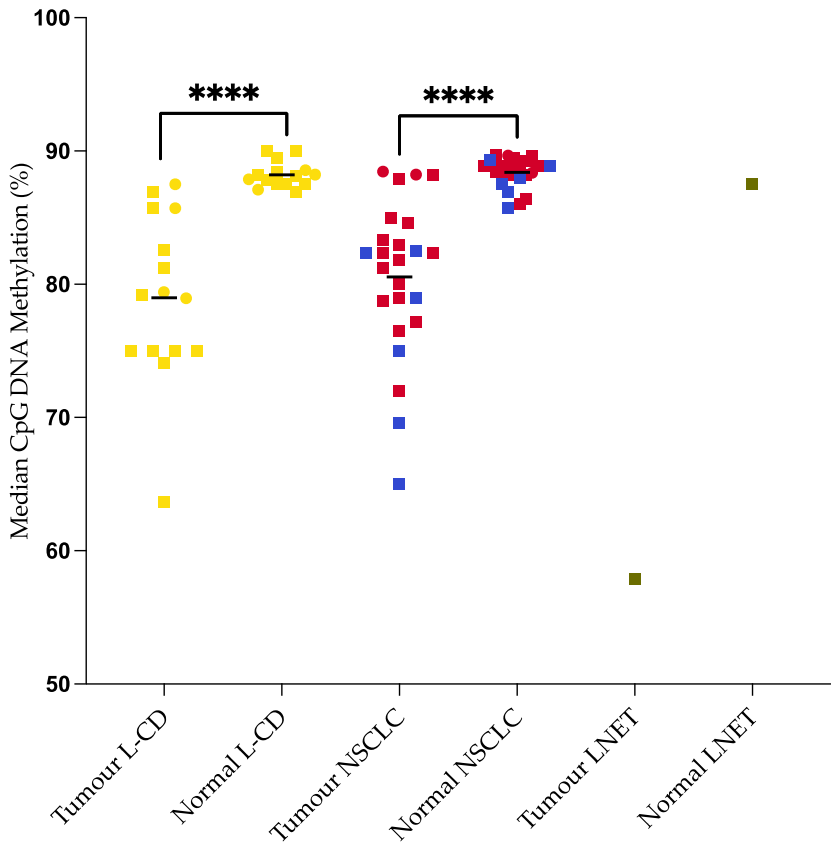


Figure 4. 5| Median CpG DNA methylation per sample. The figure shows median DNA CpG methylation percentage per sample in L-CD samples in yellow, NSCLCs (in red for LUADs and blue for LUSCs), and LNETs in khaki. Males and females appear as squares and circles, respectively. Median CpG DNA methylation is significantly lower in tumour L-CDs ($P = 4.136x10^{-5}$) and NSCLCs ($P = 1.694x10^{-8}$) as compared to normal tissue. L-CD, Lung Carcinoids; NSCLC, Non-Small Cell Lung Cancer (includes Lung Squamous Carcinoma [LUSC] and Lung Adenocarcinoma [LUAD]); LNET, Lung Neuroendocrine Tumours.

There was a number of trends observed between DNA methylation and TMB, CNB and age, although these were not significant in all combined or for the individual histotypes (Fig. 4.6). For example, median DNA methylation levels showed a negative trend with TMB (R=-0.066; $P = 0.76$), CNB (R= -0.53; $P = 0.0091$) and age (R =-0.13; $P = 0.55$) for NSCLCs, whilst L-CDs showed a slightly positive trend for CNB (R = 0.18; $P = 0.53$) and age (R = 0.081; $P = 0.77$). TMB also showed a negative trend (R = -0.28; $P = 0.34$) with DNA methylation in L-CDs. This data suggested the heterogenous genetic and epigenetic landscapes amongst LC subtypes.

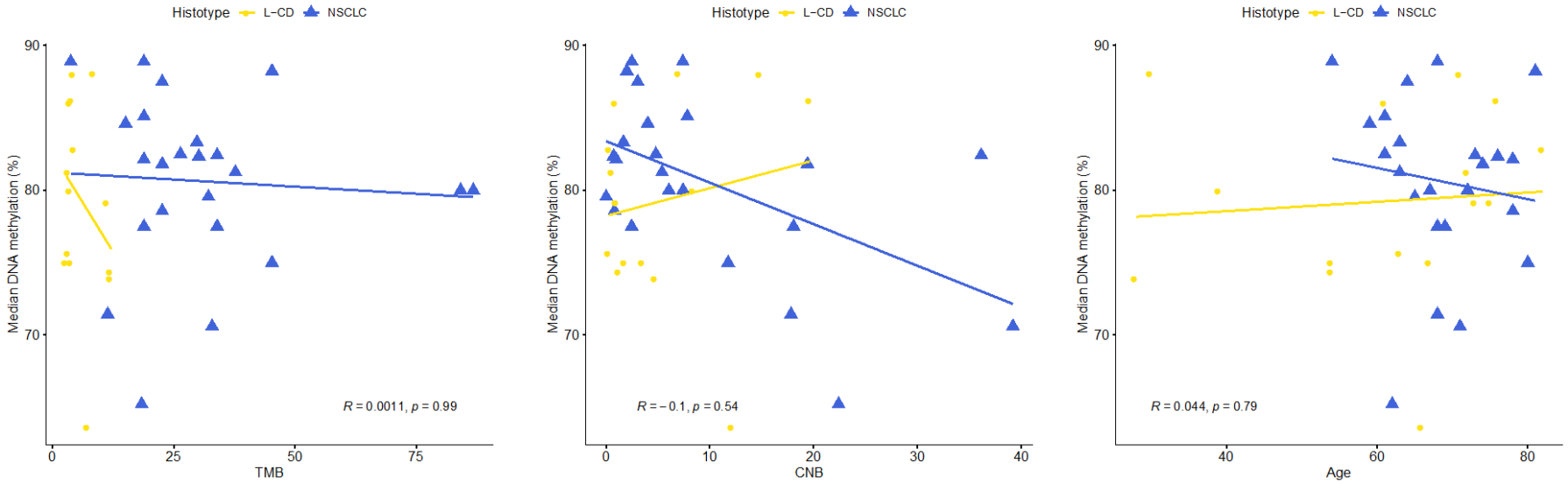


Figure 4. 6| Scatter plots of the correlation between CNB, TMB and age with median genome wide CpG DNA methylation per sample. Spearman correlation coefficients and *P*-values are shown for the combined associations in both NSCLCs and L-CDs datasets; and age is shown in years. Abbreviations: CNB, Copy Number Burden; TMB, Tumour Mutation Burden; L-CD, Lung Carcinoids; NSCLC, Non-Small Cell Lung Cancer (includes Lung Squamous Carcinoma [LUSC] and Lung Adenocarcinoma [LUAD]); LNET, Lung Neuroendocrine Tumours.

Next the number of DMRs per chromosome for each of the three comparisons (T vs N within each histotype, and T vs T between histotypes) was determined. As can be seen in Figure 4.7, L-CDs harboured a particularly high number of DMRs in chromosomes 4 and 8 when compared to normal lung tissue. Interestingly, NSCLCs exhibited roughly half the number of DMRs per chromosome detected in L-CDs, with chromosome 7 harbouring the greatest number of differentially methylated regions. Nevertheless, an independence test showed that the differences were statistically not significant (two-sided Fisher's hybrid test: $P = 1$) When comparing between tumours, the most DMRs were detected in chromosomes 1 and 4.

Furthermore, the amount of DMRs were more stable for the T vs T comparison, with a coefficient of variation of 42.94%. Interestingly, NSCLC tumours showed a very high coefficient of variation as compared to L-CDs, with coefficients of variation of 77.98% and 46.31%, respectively. This suggests more stable but substantial genome-wide epigenetic deregulation in L-CDs, whilst a more intermittent/discontinuous variation of DNA methylation levels in NSCLCs.

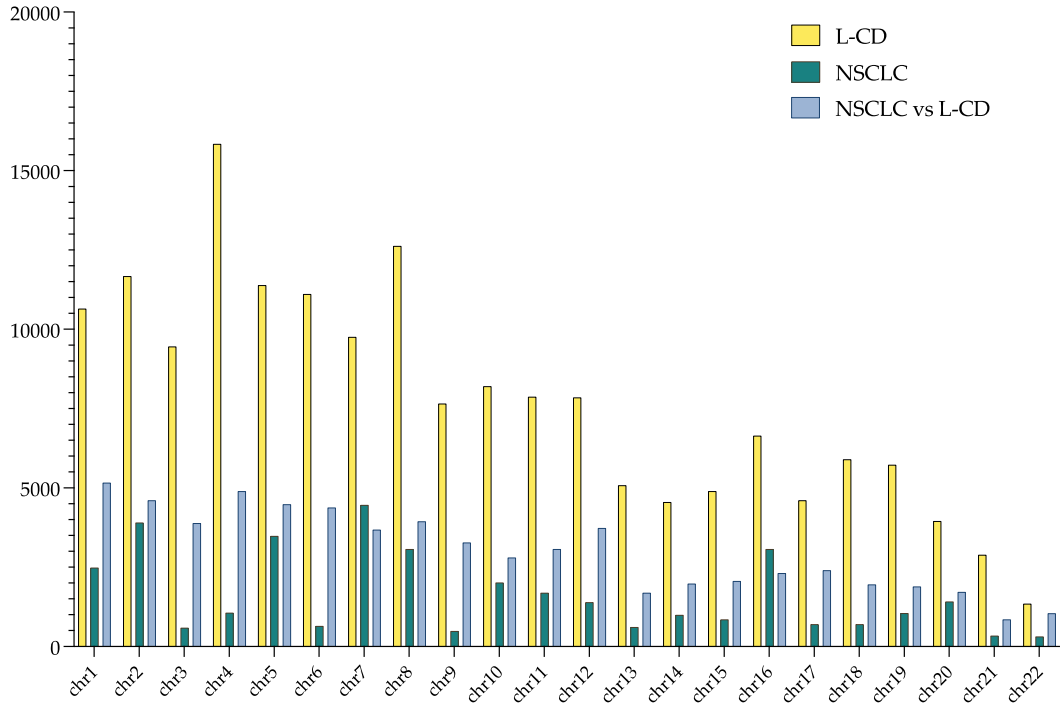


Figure 4. 7| Number of DMRs per chromosome in three different contrasts. Colours indicate different tumours types/comparisons as indicated by the key. X axis shows chromosomes and Y axis shows the number of DMRs. Abbreviations: DMR, Differentially Methylated Region; L-CD, Lung Carcinoids; NSCLC, Non-Small Cell Lung Cancer (includes Lung Squamous Carcinoma [LUSC] and Lung Adenocarcinoma [LUAD]); LNET, Lung Neuroendocrine Tumours.

4.7.2 DMR Annotation

To put these differences in context, integration of the regions obtained through differential methylation analysis with genome annotation datasets available with the Annotatr R package was conducted. The annotations datasets included both CpG annotations, such as CpG islands (CGIs), shores and shelves, and intergenic regions, as well as common genic annotations such as promoter, exonic, intronic and untranslated regions (UTRs), between 1 and up to 5 Kb upstream of the transcription start site.

The total number of DMRs was substantially higher in L-CDs when comparing against the normal tissue, with number of DMRs equalling 246,621 followed by 98,121 DMRs for the inter-tumour NSCLC versus L-CD comparison and finally 89,689 DMRs when comparing NSCLC tumour against normal tissue (Supplementary Table 4.1).

Interestingly, the pattern of regions accumulating DMRs was the same in all three comparisons as can be seen in Figure 4.8. Despite not being included in the below figure, the number of intergenic regions with DMRs was huge for L-CDs (192,996) followed by 58,566 DMRs for NSCLC vs L-CD and 74,342 DMRs for NSCLC.

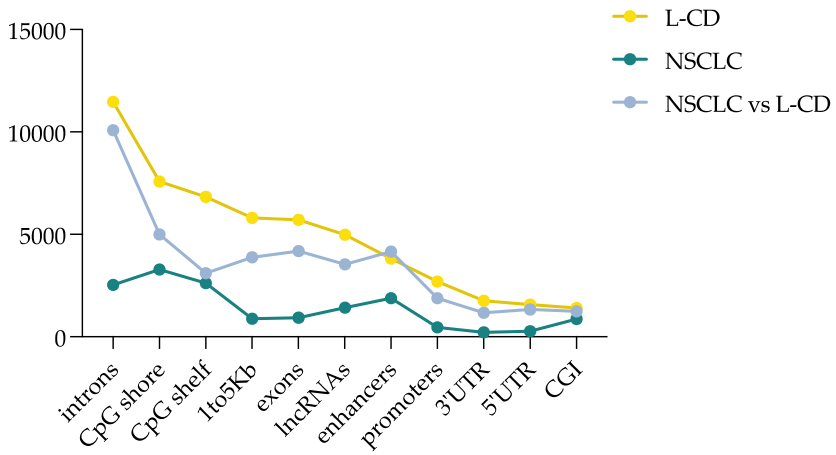


Figure 4. 8| Number of DMRs at different genic and CpG annotation classes for the three different comparative analyses. Abbreviations: DMR, Differentially Methylated Region; L-CD, Lung Carcinoids; NSCLC, Non-Small Cell Lung Cancer (includes Lung Squamous Carcinoma [LUSC] and Lung Adenocarcinoma [LUAD]); LNET, Lung Neuroendocrine Tumours; CGI, CpG Island; LncRNA, Long non-coding RNA; UTR, Untranslated region; Kb, Kilo base.

After intergenic regions, introns, CpG shores and shelves were the annotations concentrating the highest numbers of DMRs (Figs. 4.9 - 4.10). Interestingly, intronic regions were the most different between L-CD tumour and normal (*n* DMRs = 11,473) and between L-CD and NSCLC tumours (*n* DMRs = 10,078). This suggests that these regions may be of particular relevance for differentiating between these tumours. This observation mirrors the results obtained from the PCA (Section 4.5 above and Table 4.1), in which introns and the epigenetic mark related to alternative splicing, H3K36me3, were both found to explain a high amount of variance - 36.57% and 39.71%, respectively.

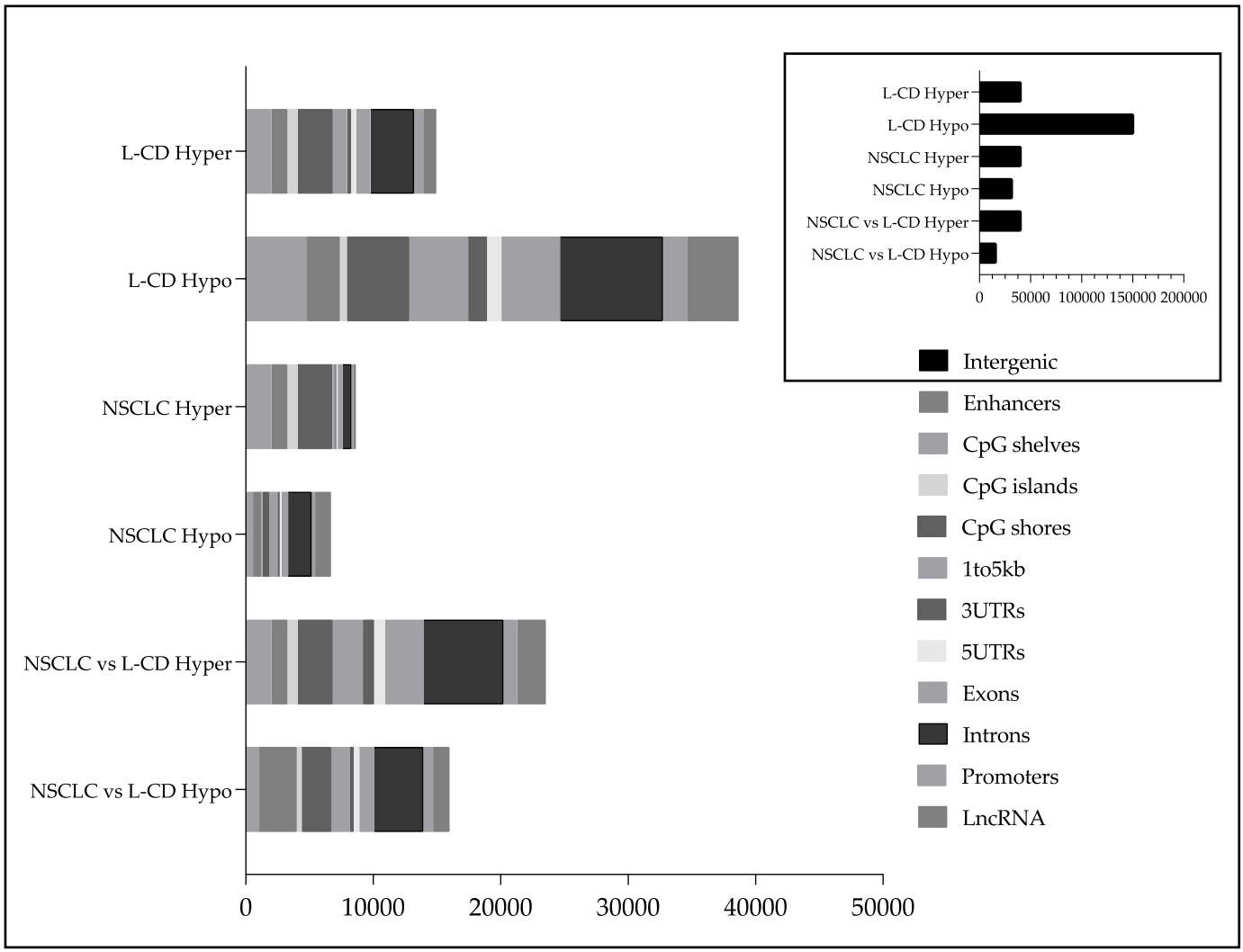


Figure 4. 9| Number of hypomethylated (Hypo) and hypermethylated (Hyper) Differentially Methylated Regions at each genomic annotation category in three different comparative analyses. From top to bottom, hypermethylated and hypomethylated DMRs detected when comparing L-CD tumour versus normal and NSCLC tumour versus normal; and finally comparing NSCLC tumours versus L-CD tumours. Number of intergenic DMRs are shown on the top right corner in black for the same comparative analyses.

Furthermore, the number of DMRs in L-CDs was much higher for hypomethylation than for hypermethylation events in L-CDs as compared to normal tissue. In NSCLCs, hypomethylation and hypermethylation events were more even, although leaning towards a higher number of hypermethylated regions or DMRs.

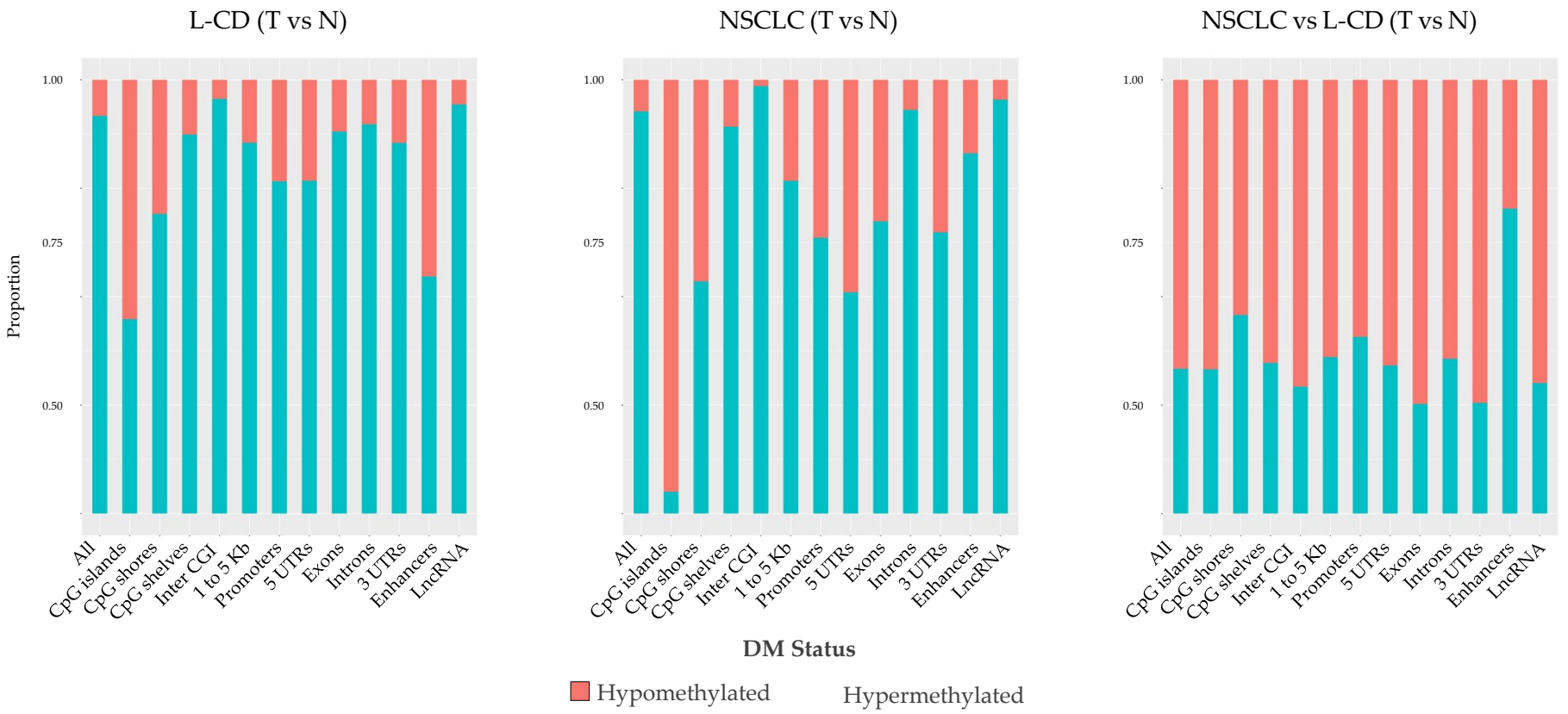


Figure 4. 10| Proportion of hypomethylated (Hypo) and hypermethylated (Hyper) Differentially Methylated Regions at each genomic annotation category in the three different comparative analyses. Figures show the same data as in the previous Figure 4.9 separate for each comparison to aid interpretation.

4.8 Pathway Analysis of Promoter Genes with DMRs and Comparative Inference

Next, to get mechanistic insights from the gene lists generated from the DMR, the Reactome and g:profiler databases were used to scan for gene promoters and other subsets of genes of interest that will be detailed for each particular contrast next in this chapter. A detailed summary of the pathways enriched in each comparison is provided in Supplementary Tables 4.2 a-c.

4.8.1 L-CD: Tumour vs Normal

A total of 767 promoters were found to be hypermethylated in L-CD as compared with normal tissue, with 1,925 hypomethylated. Similarly, regions 1 to 5 Kb upstream of 1,157 and 4,650 genes were found hyper- and hypomethylated respectively.

Pathways enriched in hypomethylated promoter genes and upstream regions were associated with olfactory signalling and sensory perception pathways, whereas hypermethylated regions were enriched for pathways related to apoptosis, the attenuation of the heat shock transcriptional response and the activation of HSF1 by stress, pathways related to neuronal development (myelinization, axon repulsion and semaphoring interactions) and gene expression during endocrine differentiation in the developing pancreas.

4.8.2 NSCLC: Tumour versus Normal

A total of 277 promoters were found to be hypomethylated in NSCLC as compared with normal tissue, and 177 hypermethylated. Similarly, regions 1 to 5 Kb upstream of 202 and 684 genes were found to be hyper- and hypomethylated respectively.

Pathways enriched in hypomethylated promoter genes and upstream regions included the olfactory signalling pathway, sensory perception, RUNX mediated transcription, activation of matrix metalloproteinases and downregulation of ERBB2 signalling, regulation of insulin secretion and formation of the cornified envelope during terminal differentiation of keratinocytes and *TP53* regulates transcription of death receptors, among others.

Pathways enriched in hypermethylated promoters and upstream regions were related to activation of *HOX* genes during differentiation, hyaluronan biosynthesis and export, and repression of genes related to differentiation and developmental biology.

4.8.3 NSCLC versus L-CDs: Inter-tumour Comparison

A total of 1,901 promoters were hypermethylated in NSCLC as compared with L-CDs, and 790 were hypomethylated. Similarly, regions 1 to 5 Kb upstream of 2,409 and 1,470 genes were found to be hypermethylated and hypomethylated respectively.

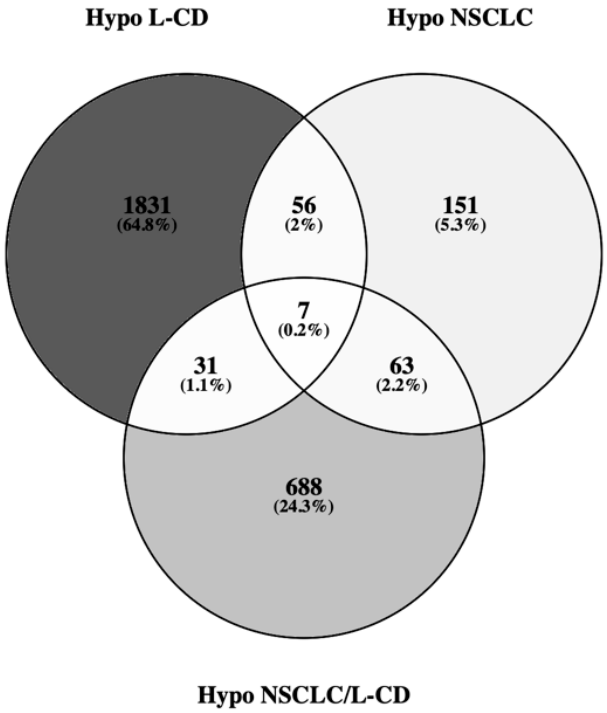
Pathways enriched in hypomethylated promoter genes and upstream regions related to themes of inflammation, necrosis and transcriptional regulation, the activation of HSF1 by stress and myelinization process. Conversely, hypermethylated regions showed processes related to transcriptional activity of *MECP2* and *HOX* genes, immune response (antigen presentation and interferon signalling), Notch1 signalling and hyaluronan biosynthesis.

4.9 Comparison of DMRs in Promoter Regions across Comparison Group

Next an analysis was performed to determine which genes were commonly differentially methylated at the promoter level across histologies or exclusively differentially methylated in a single histology only. Only a very small number of promoter genes were commonly hypomethylated (*n* 56, 2%) or hypermethylated (*n* 22, 1.2%) in both L-CDs and NSCLCs lung cancer histological subtypes (Fig. 4.11) out of the total genes hypo- and hypermethylated in each histology (L-CD n_{hypo} =1,925 and n_{hyper} = 766; NSCLC n_{hypo}=277 and nhyper=177).

In addition, a few genes were commonly hypomethylated in NSCLC tumours as compared to their normal matched tissue and in NSCLC tumours relative to L-CD tumours (2.2%). These observations suggest that distinct epigenetic programmes are dysregulated in each histological LC type as compared to the normal lung tissue. Similarly, only 1.2% of the genes hypermethylated at the promoter level were commonly detected in both LC tumour types with 3.7% commonly hypermethylated in NSCLC tumour respective to the normal lung tissue and in NSCLC tumours as compared to L-CD tumours.

a



b

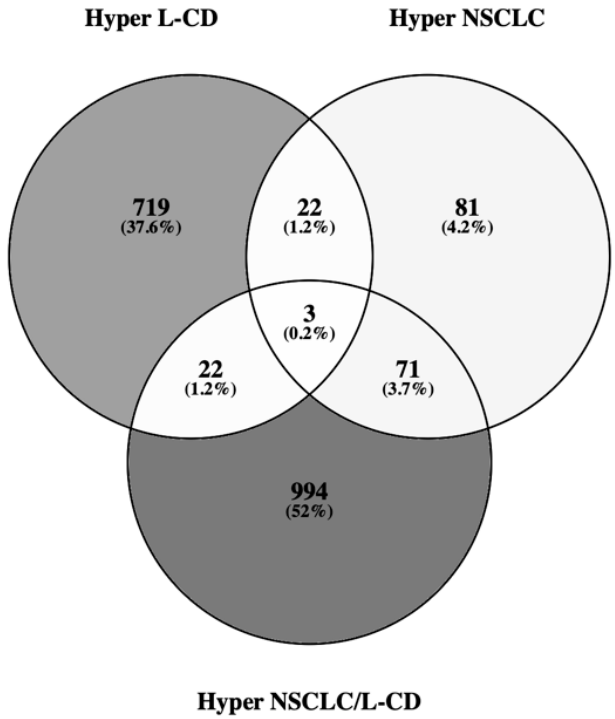


Figure 4. 11| Venn diagram of the hypomethylated (a) and hypermethylated (b) DMRs in gene promoters across the three different comparison groups.

Among common hypomethylated promoters between L-CDs and NSCLC were thirteen olfactory receptor (OR) genes: *OR10T2*, *OR14A16*, *OR1C1*, *OR1S1*, *OR2A5*, *OR2M3*, *OR2T6*, *OR5AS1*, *OR6F1*, *OR6Q1*, *OR8B2*, *OR8B4*, *OR9I1* (Appendix, Supplementary Data 4.3). ORs are predominantly expressed in the olfactory sensory neurons but they have also been observed in all non-olfactory human tissues including the lung[222,223]. In the gastrointestinal system, they can sense external stimulants from the environment as well as odorous chemicals internally[224]. In the pancreas ORs regulate insulin secretion[225], among other functions[222], when serotonin is linked to intracellular GTPases. Additionally, ORs have been suggested as markers of several cancers such as prostate cancers[226] and somatostatin receptor SSTR-negative lung carcinoid tumours[227]. Some of their most prominent functions are presented in Figure 4.10.

Olfactory pathway gene methylation marks have been associated with BMI / obesity / diet[228] all of which in turn relate to LC risk[229]. Nevertheless, in this present study DNA methylation levels at promoters or exonic regions were found not to be associated with BMI of LC patients (Appendix, Supplementary Figure 4.2).

Whether the proteins encoded by genes whose promoters were commonly hypomethylated could form a Protein-Protein Interaction (PPI) network was investigated next using the STRING database[230]. Indeed, the PPI network (Fig. 4.12) identified contained significantly more interactions than expected by chance (PPI enrichment $P = 0.0106$).

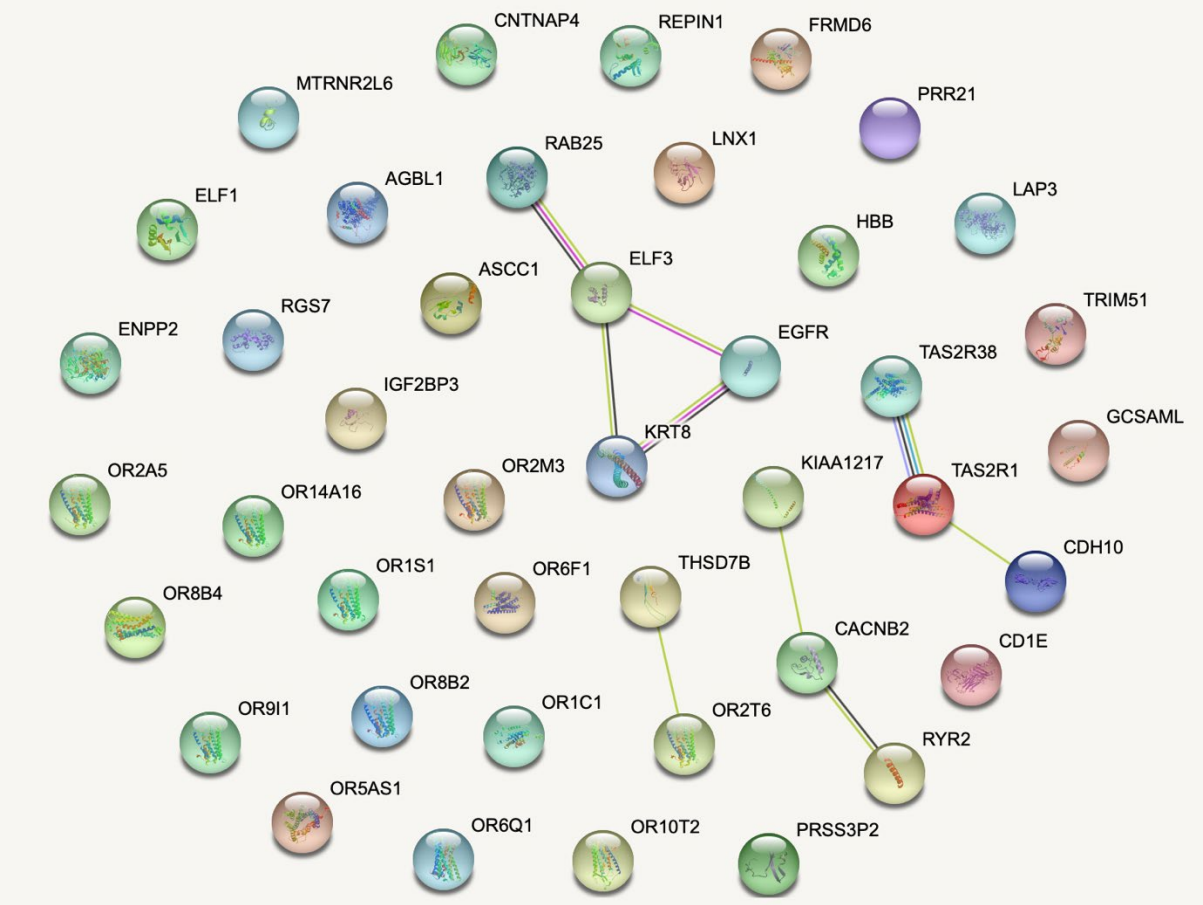


Figure 4. 12| STRING protein-protein interaction network of proteins encoded by genes whose promoters were commonly hypomethylated in L-CDs and NSCLCs. Network nodes represent proteins. Coloured nodes indicate first shell of interactions and white nodes second shell of interactions. Node content is empty for proteins of unknown structure or filled for proteins whose 3D structure is known or predicted. Edges represent protein-protein associations and are coloured to indicate different levels of support for the displayed interaction. Light blue edge for known interactions from curated databases; pink edges for experimentally determined interaction; green for predicted gene neighbourhoods; red for predicted gene fusion; blue navy for predicted gene co-occurrence; lime green for text mining; black for co-expressed proteins and purple for protein homology.

Next cBioPortal was examined to see which other commonly hypomethylated genes, from the TCGA data, were relevant in LC cohorts. Interestingly *RYR2*, *EGFR*, *CDH10*, *TAS2R1*, *THSD7B*, *LINC01020*, *RGS7*, *CD1E*, *CNTNAP4* appeared as frequently altered by somatic mutation or copy number change with frequencies ranging between 8-35% in NSCLC and LNETs respectively. In this study they were consistently identified as commonly significantly hypomethylated, suggesting that the expression of these genes may be frequently disrupted by both genetic and/or epigenetic events. Moreover, several long

intergenic non-coding RNAs (lincRNAs) and micro-RNAs (miRs) were also detected (Appendix, Supplementary Data 4.3) hypomethylated in the two LC types.

The *RAB25* gene, encoding for the Ras-related protein Rab-25 which is involved in the regulation of cell survival and associated with lung cancer invasiveness[231] and tumour acquired radio resistance in NSCLC[232,233], was also commonly found hypomethylated.

Finally, this present study also found non-coding RNAs to also be commonly hypomethylated and some have already been found to be dysregulated in other cancer types or associated with measures of lung function. For instance, variants in LINC00917 associate with Forced Expired Volume in 1 second (FEV_1)[234] and LINC01020 has been found to be differentially expressed in Kidney Renal Clear Cell Carcinoma (KIRC) where it also related to overall survival[235]. Interestingly both were identified as part of a six-lncRNA signature, high expression of which was associated with shorter survival in breast cancer[235]. Also commonly hypomethylated were several micro RNAs (miRNAs) including miR-518b, whose overexpression serves as a biomarker in NSCLC[236], and miR-520e which is upregulated in metastatic NSCLC tumour tissues as compared with non-metastatic ones[237].

Common hypermethylated promoters were found for *APAF1*, *CA3*, *CLDN11*, *COLEC11*, *CPEB1*, *CUL1*, *DRD4*, *LIMD2*, *LIMS1*, *MAPK8IP2*, *MEIS1*, *MIR30B*, *MIR3150B*, *NKX2-8*, *PDZD2*, *RRN3P2*, *SERPING1*, *TAC1*, *TBX15*, *TBX4*, *TERT* and *ZNF106*. Among these, *TERT*, *PDZD2* and *CLDN11* amplifications have been found as the most frequently genetically altered genes based upon lung cancer data generated by the TCGA Research Network: https://www.cancer.gov/tcga (accessed through cBioPortal)[238,239].

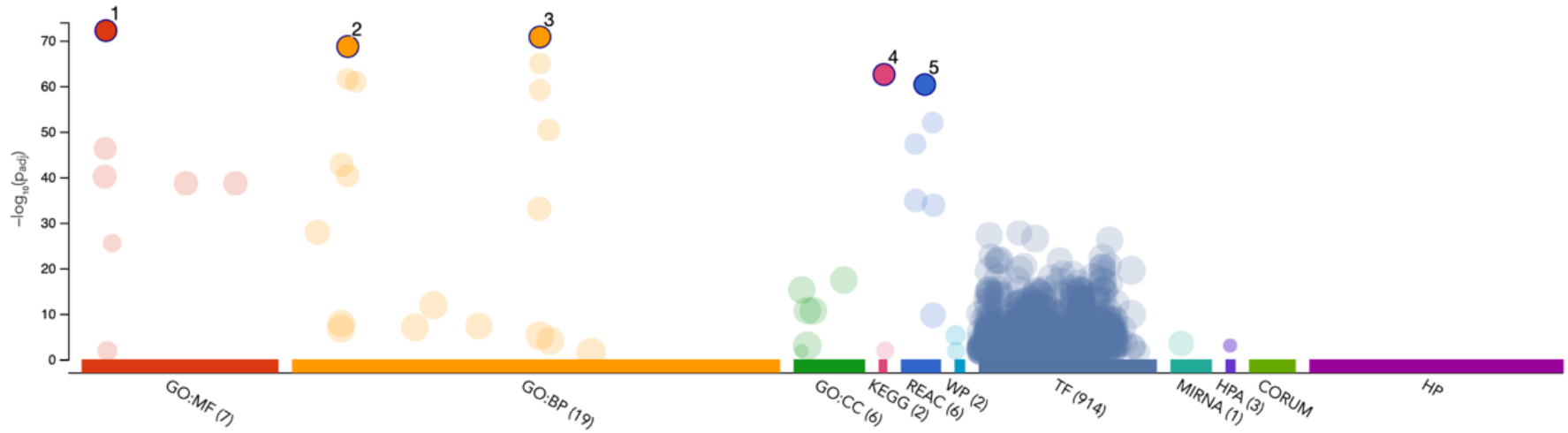
The *TERT* gene encodes for a ribonucleoprotein polymerase that acts as a cancer-protection mechanism when repressed, since telomere erosion to a critical size and dysfunction triggers the activation of the DNA damage response (DDR) pathway and subsequent replicative senescence. Thus, telomerase activation/induction confers unlimited proliferation potential by maintaining telomere ends. The most frequent mechanism for TERT expression is via promoter mutation and focal amplification[240,241]. Nevertheless, epigenetic alterations have been recently discovered to be key players of *TERT* transcriptional regulation. Importantly DNA hypermethylation within *TERT*'s promoter has been revealed to be required for *TERT* expression and telomerase activation in cancer cells[242]. PDZD2 (PDZ domain-containing 2) is a protein, with multi-PDZ domains, which is expressed in several

tissues including the lung and has been suggested as a TSG due to its anti-tumorigenic and anti-proliferative effects in LUAD[243]. CLDN11 is a tight junction associated protein and its silencing through promoter hypermethylation has been found to be associated with nasopharyngeal carcinoma progression[244]. The transcriptional silencing of *CLDN11* through hypermethylation promotes migration by derepressing tubulin polymerization.

Histology-specific DM Gene Promoters

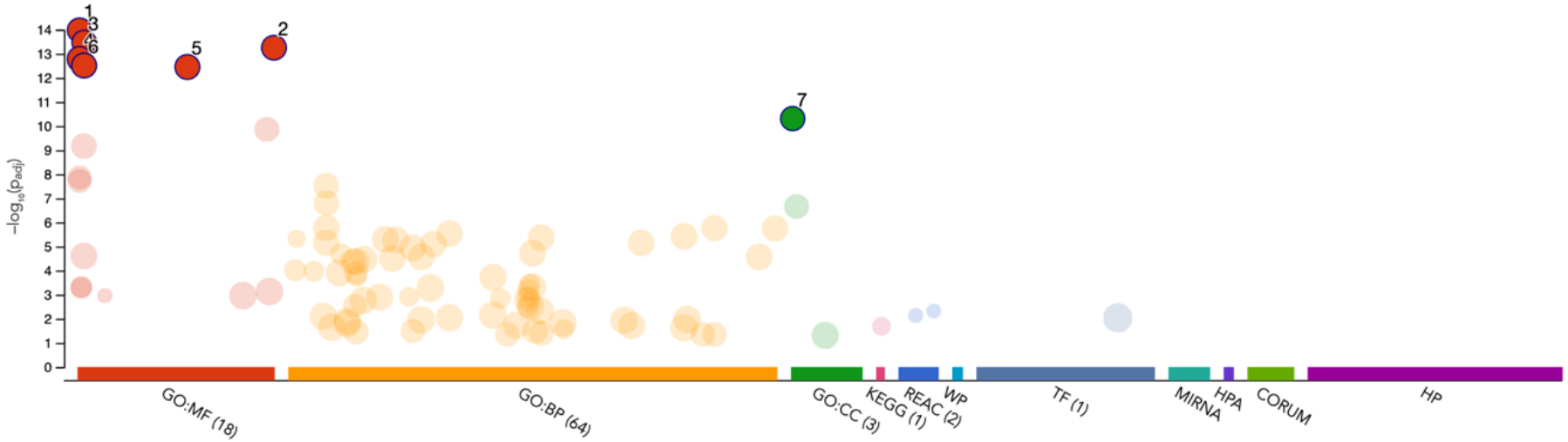
Promoters of 1,831 genes were uniquely detected as being hypomethylated in the L-CDs. Enrichment analysis with g:profiler showed again olfactory signalling related pathways as the most enriched pathways amongst DM promoters (adj.$P \leq 44.62 \times 10^{-61}$) (Fig. 4.13). Other signalling pathways significantly overrepresented were the relaxin receptors (adj.$P = 2.58 \times 10^{-2}$); RUNX3 which regulates immune response and cell migration (adj.$P = 5.11 \times 10^{-2}$); MECP2 which regulates TFs (adj.$P = 5.11 \times 10^{-2}$), loss of function of FBXW7 in cancer and NOTCH1 signalling (adj.$P = 5.13 \times 10^{-2}$) and beta defensins (n=10 genes, adj.$P = 5.4 \times 10^{-2}$). Other abundant families of genes exclusively found to be hypomethylated in L-CDs included genes encoding for Zinc-Finger proteins (ZNFs), Solute-Carrier (SLC) superfamily proteins, small nucleolar RNAs (SNORs), interferon (IFN) family members, genes from the human leukocyte antigen (HLA) complex, and proteins of the High-Mobility Group (HMG) family (Appendix, Supplementary Data 4.3).

Regarding promoters exclusively hypomethylated in NSCLCs as compared to the paired normal tissue, promoters of 151 genes were identified. Only one, the olfactory receptor activity pathway, was found significantly enriched (adj.$P = 5.195 \times 10^{-3}$). On the other hand, 81 genes exclusively hypermethylated at the promoter level in NSCLCs tumours were enriched in pathways associated to transcriptional activity (Fig. 4.14).



ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:MF	GO:0004984	olfactory receptor activity	6.681×10^{-73}
2	GO:BP	GO:0007608	sensory perception of smell	2.084×10^{-69}
3	GO:BP	GO:0050911	detection of chemical stimulus involved in sensory perception of smell	1.646×10^{-71}
4	KEGG	KEGG:04740	Olfactory transduction	2.848×10^{-63}
5	REAC	REAC:R-HSA-3...	Olfactory Signaling Pathway	4.621×10^{-61}

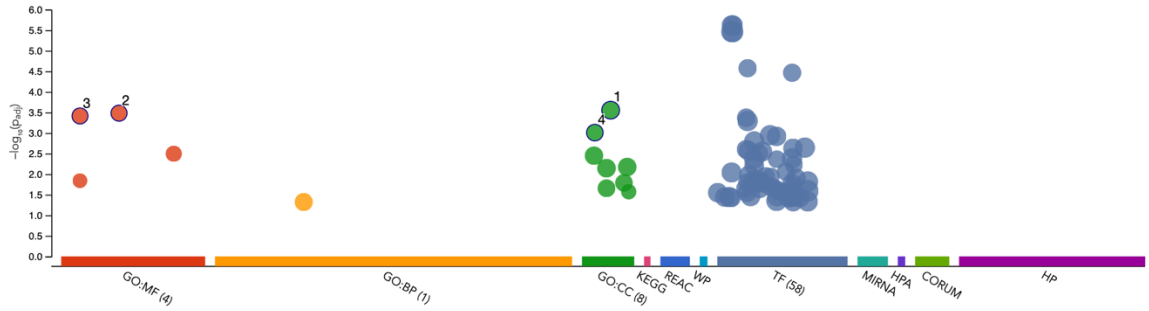
Figure 4. 13| | g:GOSt Manhattan plot of the significantly enriched pathways in hypomethylated promoters in L-CDs as compared to their normal matched tissue obtained with g:Profiler web server (https://biit.cs.ut.ee/gprofiler). The functional terms from Gene Ontology (GO) subcategories are grouped and colour-coded by data source are shown on the *X* axis, and significance of the enrichment is shown on the *Y* axis in negative log10 scale. Abbreviations: GO:MF, Gene Ontology: Molecular Function; GO:BP, Gene Ontology: Biological Process; GO:CC, Gene ontology: Cellular Component; KEGG, Kyoto Encyclopedia of Genes and Genomes; Reac, Reactome; WP, Wiki pathway; TF, Transcription Factor; HPA, Human Protein Atlas.; HP, Human Phenotype Ontology.



ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:MF	GO:0000981	DNA-binding transcription factor activity, RNA polymerase II-specific	1.049×10^{-14}
2	GO:MF	GO:1990837	sequence-specific double-stranded DNA binding	5.590×10^{-14}
3	GO:MF	GO:0003700	DNA-binding transcription factor activity	3.418×10^{-14}
4	GO:MF	GO:0000976	transcription cis-regulatory region binding	1.630×10^{-13}
5	GO:MF	GO:0043565	sequence-specific DNA binding	3.523×10^{-13}
6	GO:MF	GO:0003690	double-stranded DNA binding	3.068×10^{-13}
7	GO:CC	GO:0000785	chromatin	4.899×10^{-11}

Figure 4. 14| g:GOSt Manhattan plot of the significantly enriched pathways in hypermethylated promoters in NSCLCs as compared to their normal matched tissue obtained with g:Profiler web server (https://biit.cs.ut.ee/gprofiler). The functional terms from Gene Ontology (GO) subcategories are grouped and colour-coded by data source are shown on the *X* axis, and significance of the enrichment is shown on the *Y* axis in negative log10 scale. Abbreviations: GO:MF, Gene Ontology; Molecular Function; GO:BP, Gene Ontology: Biological Process; GO:CC, Gene ontology: Cellular Component; KEGG, Kyoto Encyclopedia of Genes and Genomes; Reac, Reactome; WP, Wiki pathway; TF, Transcription Factor; HPA, Human Protein Atlas.; HP, Human Phenotype Ontology.

Finally, genes found uniquely hypermethylated in NSCLC tumours as compared to L-CD tumours were associated to GTPase activity and synapse and axonal cellular components (Fig. 4.15). Although interesting, because these pathways have previously been found to be dysregulated in cancer[245], it should be noted that the pathways failed to achieve significance at the FDR. Nevertheless, REACTOME pathways detected as enriched were associated to MECP2 regulates TFs (P = 7.81x10^{-3}), loss of function of FBXW7 in cancer and NOTCH1 signalling (P = 1.15x10^{-2}), ERBB2 regulation of cell motility (P = 1.65 x10^{-2}) and the circadian clock (P = 1.76 x10^{-2}).



ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:CC	GO:0045202	synapse	2.786×10^{-4}
2	GO:MF	GO:0030695	GTPase regulator activity	3.290×10^{-4}
3	GO:MF	GO:0005096	GTPase activator activity	3.849×10^{-4}
4	GO:CC	GO:0030424	axon	9.788×10^{-4}

Figure 4. 15| g:GOSt Manhattan plot of the significantly enriched pathways in hypermethylated promoters in NSCLs tumours as compared to L-CD tumours obtained with g:Profiler web server (https://biit.cs.ut.ee/gprofiler). The functional terms from Gene Ontology (GO) subcategories are grouped and colour-coded by data source are shown on the *X* axis, and significance of the enrichment is shown on the *Y* axis in negative log10 scale. Abbreviations: GO:MF, Gene Ontology; Molecular Function; GO:BP, Gene Ontology: Biological Process; GO:CC, Gene ontology: Cellular Component; KEGG, Kyoto Encyclopedia of Genes and Genomes; Reac, Reactome; WP, Wiki pathway; TF, Transcription Factor; HPA, Human Protein Atlas.; HP, Human Phenotype Ontology.

4.10 Integration of Mutational, Copy Number and DNA Methylation Data

For this integrative analysis, the subset of samples that underwent WGBS data analysis (for which SNP and DNA sequencing data was also available) was used.

4.10.1 L-CDs

First looking at L-CDs (*n*=15), no overlap was found between genes that carried CNA and mutations (Fig. 4.16).

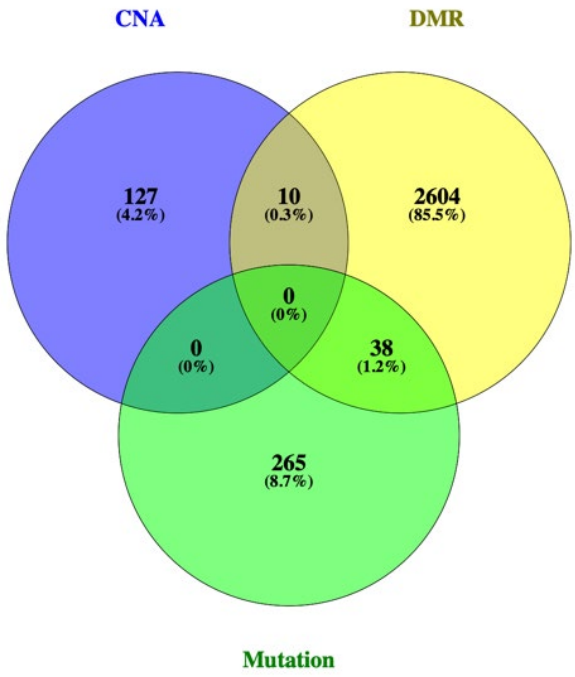


Figure 4. 16| Venn diagram for the genes altered by somatic mutation, copy number alteration and/or DNA methylation changes at the promoter level in 15 L-CD patients. Abbreviations: CNA, Copy Number Alterations; DMR, Differentially Methylated Regions.

Ten genes, however, were affected at the copy number and DNA methylation level, namely *INSL6*, *EGFR*, *TERT*, *BRD9*, *PDLIM2*, *TGIF1*, *DLGAP1*, *ANK1*, *NKX2-8* and *NMU*. A description of the genes and their alterations is given in Table 4.2.

DM & CN status	Genes	Description
Hypo & Amp	*ANK1*	Epigenetic disruption in lung and pancreatic cancers
Hypo & Amp	*BRD9*	Role in chromatin remodelling and regulation of transcription
Hypo & Amp	*DLGAP1*	Oncogenic role in multiple cancer types, by regulating tumour cell growth
Hypo & Amp	*EGFR*	Regulates epithelial tissue development and homeostasis; lung and breast cancer driver
Hypo & Amp	*NMU*	Increased expression of this gene was observed in renal, pancreatic and lung cancers
Hyper & Del	*PDLIM2*	Tumour suppressor particularly important for lung cancer therapeutic responses
Hypo & Del	*INSL6*	Mainly expressed in testis, with lower levels of expression detectable in a variety of other tissues including intestine, thymus, kidney, uterus, ovary, spleen, breast, lung, and liver
Hyper & Amp	*NKX2-8*	Overexpressed in some lung cancers and is linked to poor patient survival, possibly due to its resistance to cisplatin. Aberrantly methylated in pancreatic cancer, deleted in squamous cell lung carcinomas, and acts as a tumour suppressor in oesophageal cancer
Hyper & Amp	*TERT*	Roles in ageing and considered antiapoptotic. Active in progenitor and cancer cells
Hyper & Amp	*TGIF1*	Inhibits 9-cis-retinoic acid-dependent RXR alpha transcription activation; active transcriptional co-repressor of SMAD2 and may participate in the transmission of nuclear signals

Table 4. 2| Genes altered at Copy Number and DNA methylation level. Differential Methylation (DM) and Copy Number (CN) status are shown. Genes showing a trend towards an increase of expression by amplification (CN level) and hypomethylation (DM level) are coloured in red, whereas genes that show a trend towards a decreased expression level by deletion and hypermethylation are coloured in blue. Uncoloured cells indicate disagreement between CN and DM status.

In addition, thirty-eight genes harboured DMRs were found to be mutated (Supplementary Table 4.4). Genes included *TJP1* (tight junction protein), *HOXB3* (differentiation), *TRIB1* (cell cycle and immune regulatory functions) and *CYFIP2* (cell survival).

4.10.2 NSCLCs

Interestingly, overall a higher number of genes (Fig. 4.17) were found altered at the three different genomic levels in NSCLC tumours (*n*=25) when compared to the normal matched tissue.

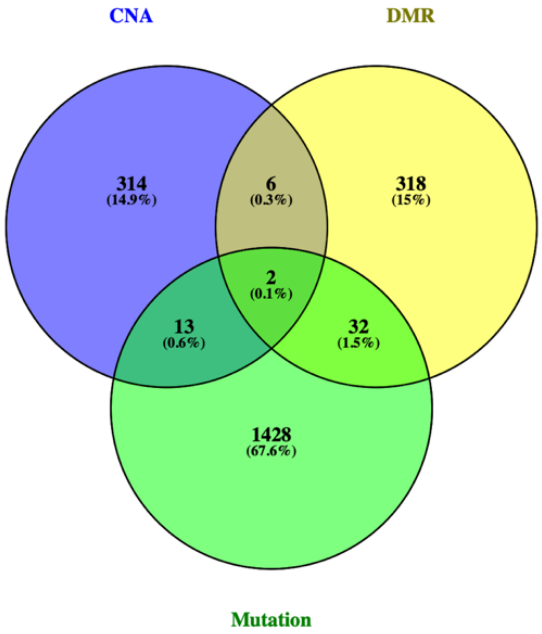


Figure 4. 17| Venn diagram for the genes altered by somatic mutation, copy number alteration or DNA methylation changes at the promoter level in 25 NSCLCs (*n*=18 LUADs and *n*=6 LUSCs). Abbreviations: CNA, Copy Number Alterations; DMR, Differentially Methylated Regions.

- 6 genes harboured both CNAs and promoter DMRs: *MIR1204*, *IRX4*, *S100A11*, *PIP5K1A*, *TXNIP* and *NKX2-8*.
- 32 genes harboured DMRs and somatic mutations (Supplementary Table 4.5), including *MET*, *PAX3*, *HOXD3* and *CDKN2A*.
- 13 genes harboured CNAs and somatic mutations in NSCLC tumours, *MKRN3*, *MYC*, *BRD9*, *CLPTM1L*, *ARNT*, *ITGA10*, *PIAS3*, *CELF3*, *ZNF229*, *ZNF467*, *ANK1*, *KAT6A* and *KDR*.
- Importantly, 2 genes harboured alterations via mutation, copy number change and differential methylation at the promoter level: *EGFR* and *SLC12A7*.

The *EGFR* gene encodes the epidermal growth factor receptor (EGFR) tyrosine kinase which is the most well-established driver mutation in NSCLC. This driver is observed in 10-

15% of Caucasian patients and up to 50% of East-Asian patients with NSCLC, with a higher incidence in females and those who have never smoked or light smokers[246,247]. Somatic mutations in this gene lead to constitutive ligand-independent receptor activation and subsequent downstream signalling promoting cell survival and proliferation.

Looking at the whole NSCLC dataset, 11 LUAD patients (12.4%) and 1 LUSC patient (2.8%) harboured *EGFR* mutations.

On the other hand, *SLC12A7* gene encodes for a transmembrane protein that acts as a solute carrier to regulate cell volume. *SLC12A7* (also known as *KCC4*) has been found overexpressed in several cancer types and associated with tumour cell growth and invasion[248–250]. Exploration of the 4,767 Lung Cancer samples from the TCGA data portal similarly showed *SLC12A7* mutations and copy number gains exclusively for LUAD and LUSC tumours (Supplementary Figure 4.6).

4.10.3 Analysis of Cis-regulatory Regions

After introns and intergenic regions, enhancer regions were the third genomic region most hypomethylated between the two tumour subtypes (see Fig. 4.18).

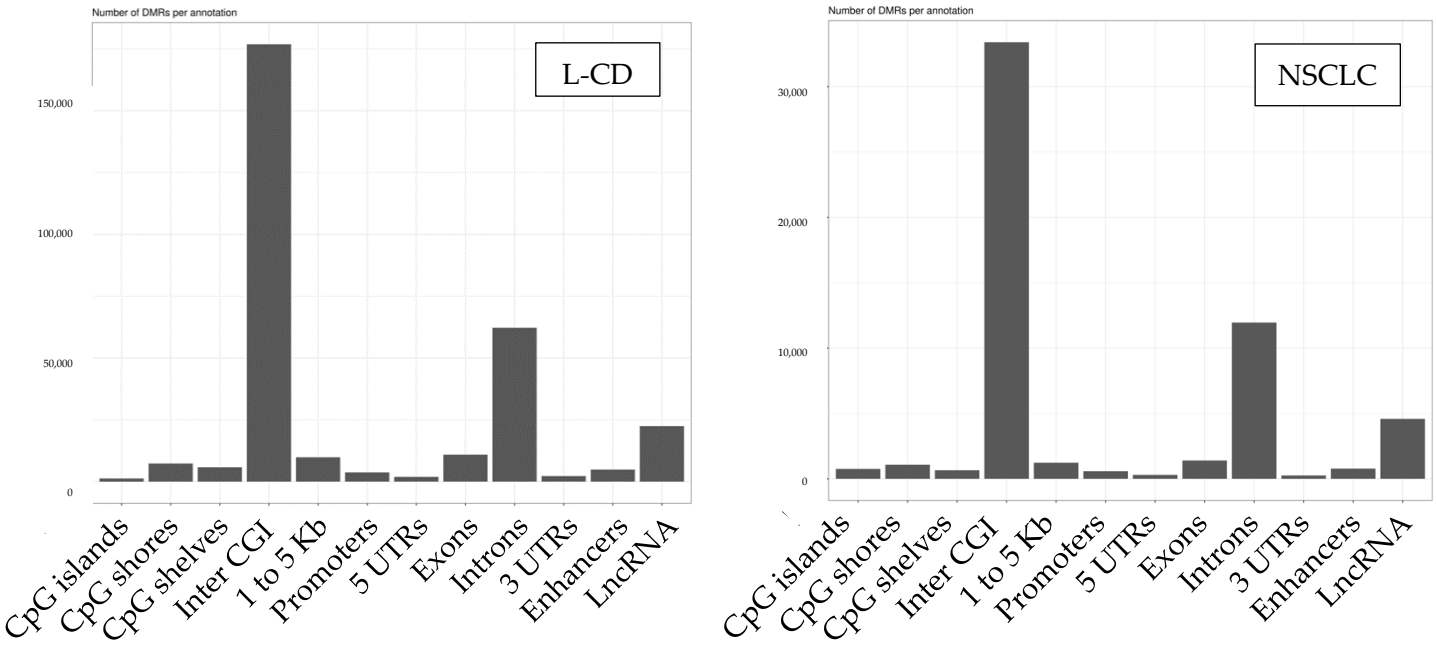


Figure 4. 18| Number of DMRs for each type of genomic annotation obtained with Annotatr. Note the difference in the *Y* axis scale between the two graphs.

Enhancer regions have been found to have a stronger correlation with gene expression than promoters[124]. Consequently, the functional significance of DM enhancers (eDMRs) by using the Genomic Regions for Enrichment (GREAT) tool for each comparison group was investigated next.

To get the most meaningful results of the pathways associated to eDMRs, the GREAT Gene Ontology (GO) results were filtered based on a normalised enrichment scores of >2 and multiple hypothesis testing corrected *P* values of <0.01, for both the binomial and the hypergeometric distribution-based tests (Appendix Supplementary Table 4.2).

L-CD Tumour versus Normal

In L-CDs a total of 2,580 and 1,235 enhancers were hypo- and hypermethylated respectively as compared to their paired normal tissues. Pathways associated to hypermethylated eDMRs were related to regulation of hemopoiesis, response to Transforming Growth Factor Beta (TGF-B), myeloid and leukocyte differentiation, angiogenesis, negative regulation of collagen biosynthesis and metabolic process, amongst others. Pathways associated to hypomethylated eDMRs included regulation of lymphocyte and T cell differentiation, regulation of dopaminergic neuron differentiation, mesoderm morphogenesis and development. The GO results are given in detail in Supplementary Table 4.2.a) in the Appendix.

NSCLC versus L-CD

In NSCLC a total of 2,925 enhancers were hypomethylated in NSCLC as compared to L-CDs, whereas the number of hypermethylated enhancers was 1,235. Hypomethylated regions in enhancers (eDMRs) were enriched in biological processes related to myeloid and leukocyte differentiation, negative regulation of ERK1 and ERK2 cascades, and peptidyl-tyrosine phosphorylation. Other significant pathways included regulation of ROS, pathways related to cell adhesion and Fc receptor mediated stimulatory signalling leading to activation of immune responses (Appendix, Supplementary Table 4.2.b). In contrast, pathways enriched in hypermethylated eDMRs included regulation of lymphocyte and T cell differentiation, regulation of circadian rhythm, angiogenesis and regulation of dopaminergic neuron differentiation, amongst others.

NSCLC Tumour versus Normal

A total of 647 and 1,235 enhancers were hypo- and hypermethylated respectively as compared to paired normal tissue. Pathways associated to hypermethylated eDMRs were related to development and tissue morphogenesis, whereas pathways associated to hypomethylated eDMRs included regulation of muscle adaptation, endoderm development and regulation of leukocyte cell-cell adhesion. The GO results are given in detail in Supplementary Table 4.2.c) in the Appendix.

4.10.4 Transposable Element (TE) Content in DMRs

Genomic sequencing has revealed that the genomes of prokaryotes and eukaryotes contain a variety of TEs. This is as a result of insertional events that occurred during evolution[251] (See Chapter 1, Section 1.3.3). In humans, these elements make up almost half of the nuclear DNA. The integration of these sequences into new sites create target site duplications and double-strand breaks. This leads to the activation of DNA repair mechanisms of the host cells to enable repair and gaps to be filled.

Since TE activity is well known to be under epigenetic control, the extent of DMRs that were enriched for TEs was investigated. Using the genomic coordinates of the four main classes of TEs (SINEs, LINEs, LTRs and DNA transposons), DMRs were overlapped and classified based on their TE content into TE-rich and TE-free DMRs (Fig. 4.18).

L-CD tumours carried the highest proportion of TE-free DMRs (48%), whereas hypomethylated DMRs overlapped more with TEs in NSCLCs (50.75%). Hypermethylated DMRs similarly had a low overlap with TEs in both LC types, with 0.45% and 2.04% in NSCLC and L-CDs respectively.

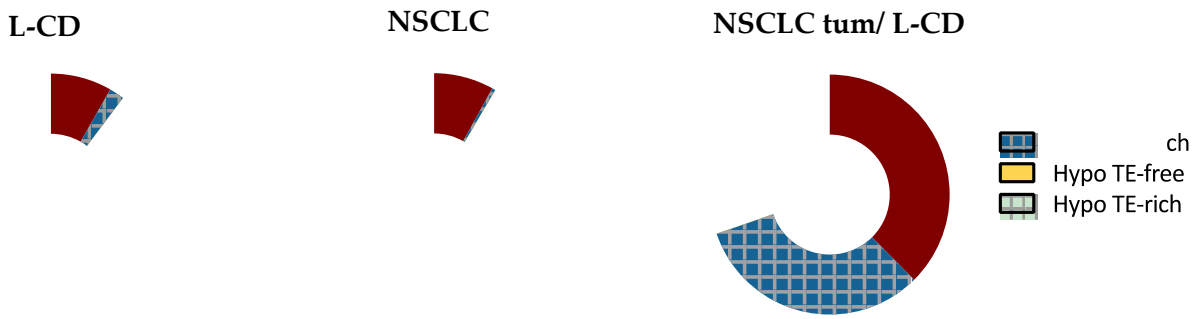


Figure 4. 19| Proportion of repeat-free and repeat-rich DMRs. L-CD and NSCLC pie charts show tumour-normal comparisons while NSCLC tum/L-CD tum show tumour- tumour comparison. In yellow: hypomethylated and in red: hypermethylated DMRs with low TE content; and in light green squared pattern: hypomethylated and in blue squared pattern: hypermethylated DMRs with high TE content. Abbreviations: TE-free, Transposable Element-free; TE-rich Transposable Element -rich; Hypo, hypomethylated DMR, Hyper, hypermethylated DMR; L-CD, Lung Carcinoids; NSCLC, Non-Small Cell Lung Cancer (includes Lung Squamous Carcinoma [LUSC] and Lung Adenocarcinoma [LUAD]).

Looking deeper into which categories showed a stronger overlap with DMRs (Table 4.3) intergenic regions showed a substantial enrichment for DMRs, as expected since most of the genome is non-coding and where the vast majority of TEs are found. Intronic DMRs were the next genomic category showing a substantial proportion of TEs. Whilst present in NSCLC, these were a prominent feature in L-CDs with a sharp distinction between the two tumour classes.

Interestingly hypermethylated DMRs in both LC tumours were found concentrated at TE-free regions. This suggests regulatory functions not related to the epigenetic silencing of TEs in cancer. Similarly, hypomethylated DMRs were more evenly distributed across overlapping and non-overlapping TE regions in both cancers, indicative that TE expression may be dysregulated through hypomethylation at TE regions in both tumour types but to a higher extent in L-CDs.

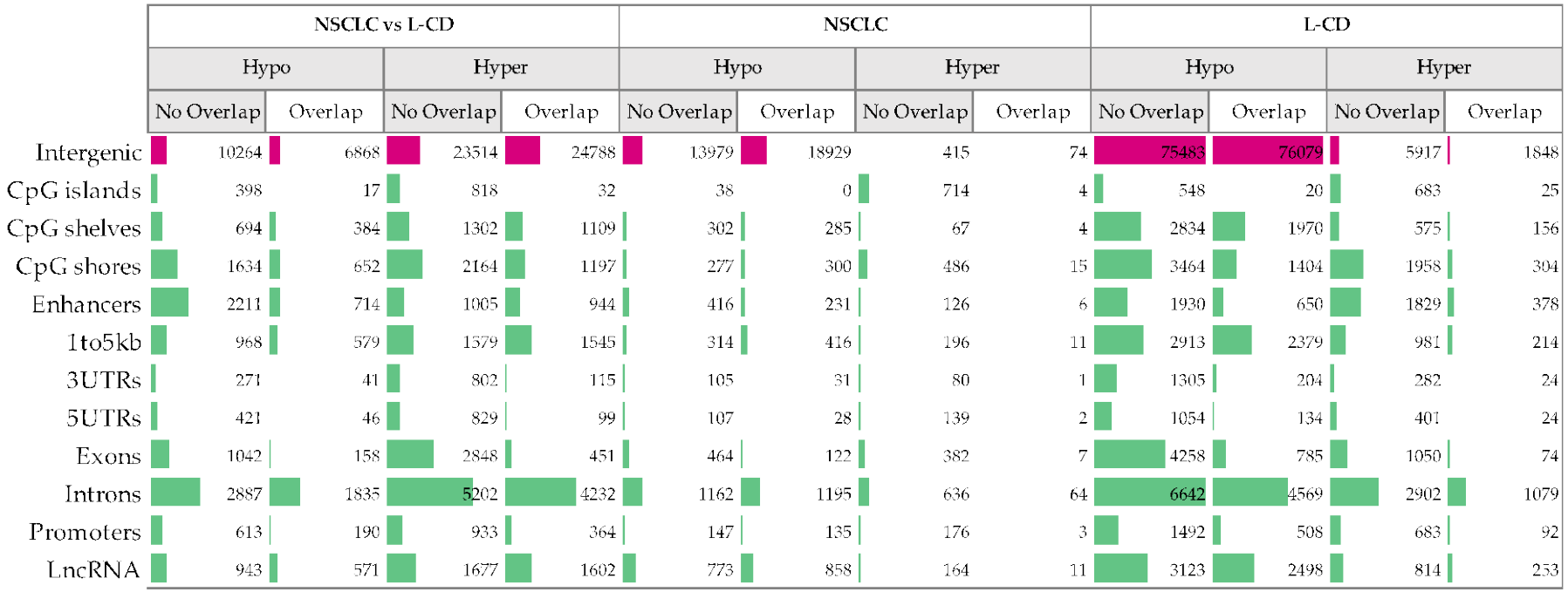


Table 4. 3| Number of Transposable Elements overlapping DMRs for each genomic category. Colour bars represent the proportion of DMRs in each category relative to the highest number. Intergenic regions are coloured independently for better proportion visualization of the rest of DMR categories.

4.11 Discussion

The work conducted in this chapter has revealed that DNA methylation at the whole-genome level distinguished NSCLC from L-CD tumours and tumours from healthy lung. Furthermore, PCA of DNA methylome data indicated that these data could *alone* be used to stratify patients and to distinguish cancer from healthy tissue. Importantly, targeted DNA methylation experiments for exonic regions could be used for cancer classification purposes, reducing the costs and time required for generating WGBS data.

Global DNA hypomethylation was detected in both L-CD and NSCLC. Importantly this confirms that tumours with both stable and non-stable genomes harbour global hypomethylation. Considering the low mutational and copy number burden in L-CDs as compared to NSCLCs (discussed in Chapter 3) one can conclude that DNA methylation loss is a characteristic of lung cancer regardless of the level of genetic instability.

Not only promoter but also DNA methylation changes at intergenic regions and gene bodies were found in this present study to be common features of both L-CD and NSCLC tumours. Other genome-wide DNA methylation studies have also provided data supporting this observation and consequently, the potential role of these alterations is starting to be explored. DNA methylation is generally suggested to be in part associated with gene expression by regulating processes related to transcription. For instance, DNA methylation plays a role in the regulation of alternative promoter usage, Alternative Splicing (AS) and polyadenylation and the expression of non-coding RNAs. In fact, the work conducted in this chapter found that lncRNAs formed the third most abundant category of aberrant methylation. In addition, many promoters of lncRNAs and miRs were detected as being differentially methylated.

Furthermore, the fact that intronic regions and H3K36me3 marks accumulated high percentages of variance could suggest that alternative splicing (AS) may be a feature of L-CDs. The promoters of *MEF2C* and *RBFOX1* were consistently hypomethylated in L-CD as compared to NSCLC tumours (Supplementary Data 4.3). These genes are involved in the regulation of AS by MeCP2. Knockdown of *MeCP2* or treatment that reduces DNA methylation (lowering MeCP2 binding to the DNA) results in the aberrant exclusion of alternative exons[252].

Previous studies have shown that methylation levels at enhancer regions correlates better with gene expression than that in promoter regions[124]. The present study has identified and associated differentially methylated enhancer regions (eDMRs) with *cis* regulatory functions by performing Gene Ontology analysis with genes physically located 50 to 500 Kb upstream or downstream of eDMRs. In both L-CD and NSCLC tumours, eDMRs were associated with developmental and differentiation programmes. This is consistent with the study by Ziller *et al.*[125], where low DNA methylation levels were commonly found enriched for cell-type-specific TF binding sites across different developmental stages and in cancer cells. Altogether, this suggests that DNA methylation patterns are associated with the expression of cell type specific TFs in both LC types and that deregulation of developmental regulatory programmes are a common feature in both. Further investigations are therefore warranted to confirm if these alterations lead to gene expression changes and such investigations will be addressed in the subsequent chapters of this thesis.

EGFR appears as a common differentially methylated promoter in both LC histotypes, together with mutations and gene amplifications (Chapter 3, Fig. 3.14), suggesting that its expression and/or function can be altered at different levels, confirming its importance in LC carcinogenesis. Notably *EGFR* promoter hypomethylation has also been reported for gastric, colorectal, breast, head and neck squamous cell carcinoma and lung tumours[253]. The fact that different regulatory mechanisms are altered for this gene strengthens its potential as a therapeutic target for tyrosine kinase inhibitors (TKIs).

Interestingly, in this present study, circadian genes were found to be hypermethylated including *PER1*. The latter has already been reported previouslyto be hypermethylated in NSCLC[254]. Moreover *PER2*, *CRY2* and *RORA* are known circadian genes altered in cancer[255] via promoter hypermethylation and have been detected also in this study for NSCLC. Circadian clocks can be disrupted by genetic, environmental and internal factors, all of which can in turn also disrupt cellular processes related to tumorigenesis (for example metabolic reprogramming, redox imbalance, chronic inflammation, etc.). In addition, many circadian clock proteins physically interact with proteins that participate in pathways relevant to cancer. In this way modulation of circadian clock function or expression of clock proteins may protect or promote cancer, opening a new line of study and therapeutic options.

Several DNA methylation changes occur across a cell's lifespan. Embryonic stem cells show the highest levels of DNA methylation followed by primary cells and finally aged and cancerous cells have the lowest levels of DNA methylation. In this study, DMRs overlapping with TEs were mainly found at hypomethylated regions suggesting TE reactivation in both NSCLC and L-CD tumours. In contrast, hypermethylated DMRs were not colocalised with TEs. This is consistent with patterns observed in age-related diseases and cancer, where a decline in repeat element DNA methylation has been observed and CpG islands may become methylated[256].

Limitations of the present study include the heterogeneity in cellular composition that can directly cofound tumour-specific DNA methylation levels. It has become clear that cell type is a potential cofounder in DNA methylation studies, particularly in complex mixtures such as blood or tumour samples. Normal lung tissue is composed of many different cell types including cells of the epithelium, interstitial connective tissue, blood vessels, immune cells etc, and together with recent studies looking at immune cell composition in NSCLC[257], there is a level of sample heterogeneity at the cellular level. Thus, further DNA methylation investigations although costly should be performed at the single cell level to confirm observations to date for whole tissue samples as well as identify further DNA methylation changes in cancerous cells.

In addition, the repetitive nature of TEs remains a challenge for their accurate detection from high-throughput sequencing data. TEs account for more than half of the genome and can be even higher in abundance due to the insertion of active classes across the genome due to several biological processes. Consequently alignment and accurate detection can be confounded with mapping artefacts and requires high sequencing read depth.

There is accumulating evidence on the effect of biological sex in in DNA methylation patterns during aging and disease[258–261]. In the setting of exploratory PCA analysis (Section 4.5), only sex-mediated differences on the sex-chromosomes could be excluded as LC histotypes were clearly clustering in different groups when accounting only for DNA methylation data on autosomes (data not shown). This however cannot remediate the cofounding factor of sex-mediated differences as a whole in the different histotypes. Being aware of this limitation and to control for this factor as it may mask DNA methylation

signatures, DNA methylation data was adjusted for biological sex for calling DMRs in this present study.

Nonetheless larger datasets, with higher representation of both genders in each LC histological group, would be required to determine if sex-driven differences are not explaining the dissimilarity seen in the methylomes.

Chapter 5: Marked Loss of Y Expression in NSCLC

This chapter describes and extends the research paper "Y disruption, autosomal hypomethylation and poor male lung cancer survival", published in Nature Scientific Reports in 2021[1] (full details of which are given on page 6).

5.1 Introduction

Through network analysis, Willis-Owen *et al.*[1] identified weakening of a male-specific gene expression in 27.7% of male NSCLC tumours, accompanied by poor survival. Specifically, a Weighted Gene Co-expression Network Analysis (WGCNA)[262] was used to summarise gene expression separately in tumour and matched normal tissues allowing comparison at the system level. WGCNA specifies modules or networks of co-varying transcripts. These transcripts can vary in the same or opposing directions. In this way, these networks contain genes that are related at the functional level, for example forming part of the same signalling pathway or regulated by similar TFs. Additionally, prior evidence indicates that genes that are highly interconnected or central to a gene expression network typically have the highest impact when disrupted. As a result, the application of WGCNA can identify common and divergent gene co-expression networks between tissues and specify key genes within these.

For this study, WGCNA analysis (carried out by Dr. S.A.G. Willis-Owen) was used to identify gene co-expression networks in expression data generated with the Affymetrix HuGene 1.1 ST microarray. Research samples consisted of tumour and normal lung tissue from 126 NSCLC patients with Stage IA – IV lung cancer. DNA and RNA extractions and expression data had been generated by prior members of the Asmarley Centre for Genomic Medicine (NHLI, Imperial College London) and their collaborators[2].

To study common and unique networks between tumour and normal NSCLC samples, networks were first constructed for tumour and normal samples together (consensus networks). Then, networks were constructed in tumour and normal samples separately to allow to specification of tissue-specific and common networks.

One network specific to the normal lung tissue showed a significant relationship with biological sex (P = 3.72x10^{-28}) and lacked assignment to a consensus network, indicating a different co-expression pattern between tumour and normal lung tissue samples in NSCLCs.

Furthermore, hierarchical clustering of samples and transcripts of this network distinguished an individual branch featuring a tumour-specific disturbance related to a reduction of male-specific gene expression (Fig. 5.1a). This network included over a quarter of male NSCLC tumours (including both LUAD and LUSC tumour histologies). This observation was also reproduced in a replication dataset comprising microarray expression data of 123 samples from 69 patients with either LUAD or LUSC histology (Fig. 5.1b).

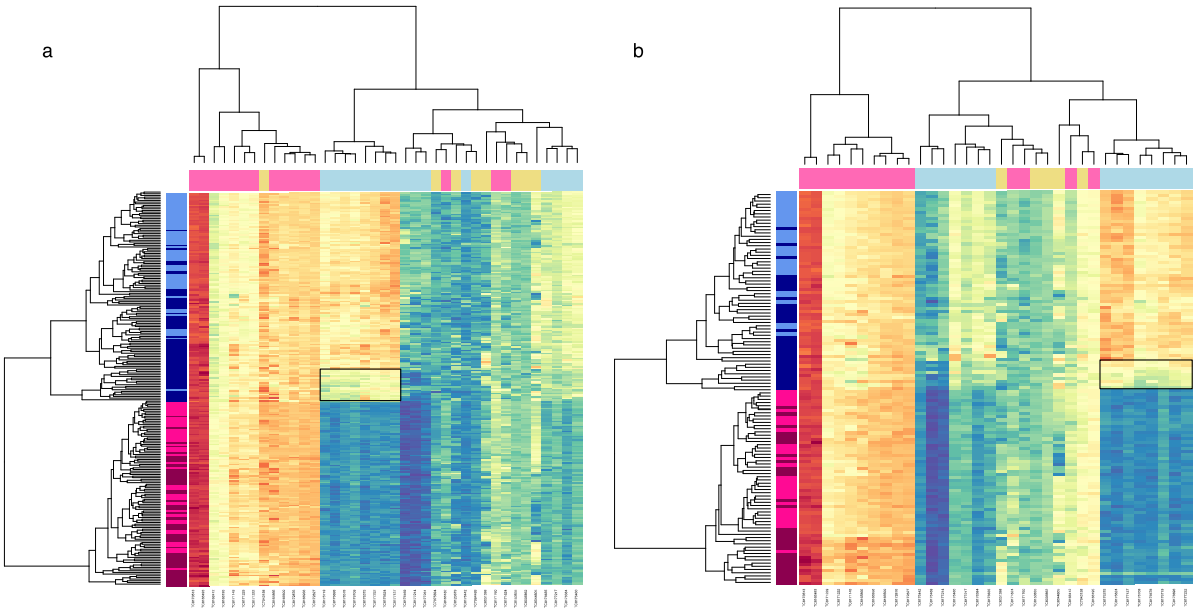


Figure 5. 1| Hierarchical clustering of transcripts assigned to a normal-specific sex associated co-expression network in a) discovery and b) replication datasets. Samples are shown on the *y* axis with transcripts on the *x* axis. Expression is shown on a continuous colour scale from blue (low) to red (high). Sample colour (*y* axis) reflects tissue type (light – histologically normal, dark - tumour) and sex (blue - male, pink - female). Transcript colour (*x* axis) reflects chromosome class (yellow - autosomal, pink – X, blue – Y). Low Y sample / Transcript Clusters (TCs) are highlighted by a solid black box. The figure is reproduced with permission from Willis-Owen *et al.*[1].

Most of the transcripts forming this network mapped to the sex chromosomes. Specifically, 15 transcripts mapped to chromosome X and 16 transcripts mapped to chromosome Y. The remainder mapped to autosomes, with many previously found to show sex-biased gene expression (e.g. *DDX43*, *NOX5*, *NLRP2*)[263,264]. Intriguingly, most of the transcripts mapping to the Y chromosome lacked assignment to a tumour network (*n*=12, 75%), indicating a specific loss rather than restructuring of the gene networks in tumour tissue.

This tumour-specific disturbance in sex-related gene co-expression was characterised by a substantial reduction of Y-chromosome transcripts encoded by the genes *DDX3Y*, *EIF1AY*, *KDM5D*, *RPS4Y1*, *TXLNGY*, *USP9Y* and *UTY*.

Following these observations, mechanistic insights for the low expression of Y chromosome (LYE) transcripts in these NSCLC tumours were sought through the analysis of DNA sequencing and WGBS sequencing data.

5.2 Material and Methods

The DNA sequencing and WGBS sequencing data was generated as detailed in Chapter 2.

5.3 Results

5.3.1 LYE Tumours Exhibit Low Read Depth Consistent with Somatic Loss of Y

To try and gain mechanistic insights for the deficiency of Y chromosome gene expression that had been observed, first somatic loss of Y was queried through read depth analysis of WES and WGBS data.

A subset of male tumour samples exhibiting low Y chromosome expression (*n* WES = 6, WGBS = 17) were compared with matched unaffected tissue from the same patients and with a subset of male tumour samples lacking this feature (*n* WES = 9, WGBS = 5; see Supplementary Tables 5.1 and 5.2) including all such samples for whom sufficient template was available. Consistent with tumour-specific LYE, normalised read depth was significantly lower in LYE tumours as compared with unaffected samples from the same patients (P_{WES} = 0.0108; $P_{WGBS} = 2.01 \times 10^{-20}$). In male tumours lacking the low Y gene expression signature, this was not seen (P_{WES} = 0.99; P_{WGBS} = 0.97).

Correspondingly the percentage loss was significantly greater in males with low Y-expressing tumours than in males lacking this feature (P_{WES} = 0.027; P_{WGBS} = 0.046) (Fig. 5.2).

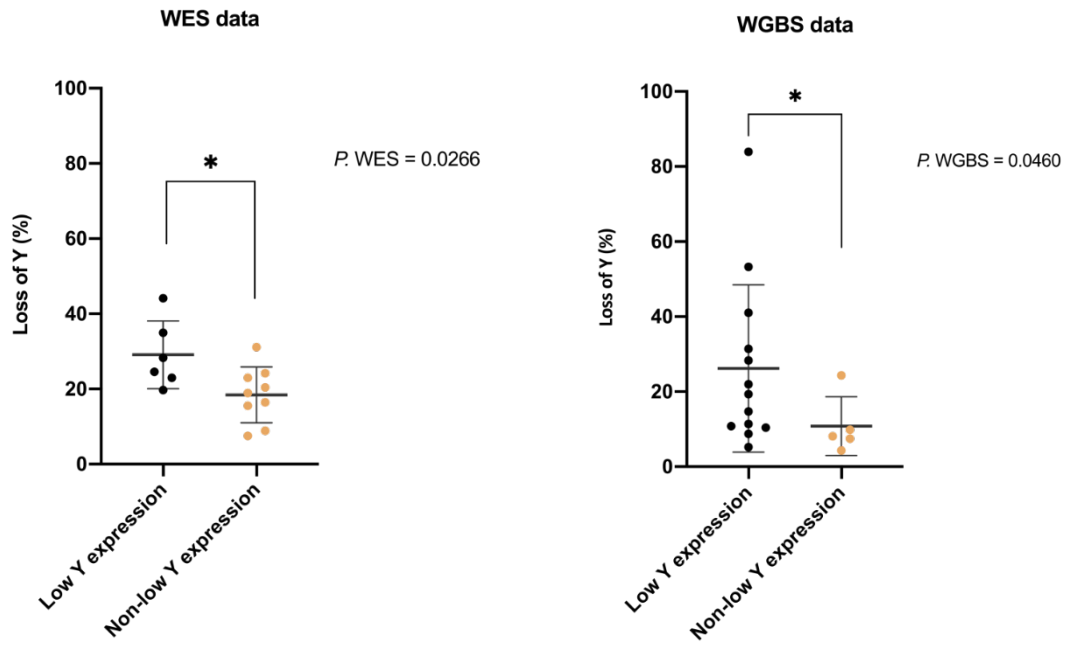


Figure 5. 2| Validation of somatic LYE through sequence read depth analysis from WES (a) and WGBS (b) data. Box plots of the percentage of Y loss in male tumours with low Y-expression and in males lacking this feature. Error bars represent standard deviations from the mean. Magnitude of significance is denoted with asterisks with *P* values as shown (*).

5.3.2 LYE Tumours Show a DNA Hypomethylation Signature

Amongst the transcripts within the sex-associated co-expression network, *KDM5D* showed the highest Module Membership (MM), a concept closely relative to intra-network connectivity as highlighted in the introduction of this chapter (Section 5.1).

The lysine Demethylase 5D *(KDM5D)* encodes a male-specific demethylase targeting trimethylated H3K4 (H3K4me3). This chromatin landmark is generally detected near the start site of transcriptionally active genes and can exhibit pronounced sex bias that can be translated to differences in gene expression between male and females[265]. Whilst histone and DNA methylation pathways involve distinct enzymes and chemical reactions, these pathways are interconnected, with complex dependency relationships[266]. Amongst histone methylation marks, H3K4me3 specifically is anti-correlated with DNA methylation[267] and mutations in the X-linked *KDM5D* homolog (*KDM5C*) have been linked with multi-locus DNA methylation loss[268] providing evidence of functional inter-dependency.

In the present study, a pronounced DNA methylation loss signature was observed in male tumours with the low Y gene expression phenotype (Fig. 5.3). Relative to paired unaffected tissues, median autosomal DNA methylation levels were significantly reduced (P = 3.12 x10^{-6}). This relative reduction was not reproduced in male tumours lacking the low Y gene expression feature (P = 0.0625) indicating that extensive hypomethylation is a characteristic of the low Y pulmonary tumour state and potentially therefore also a latent factor contributing to lung cancer-related methylation changes previously reported elsewhere[269].

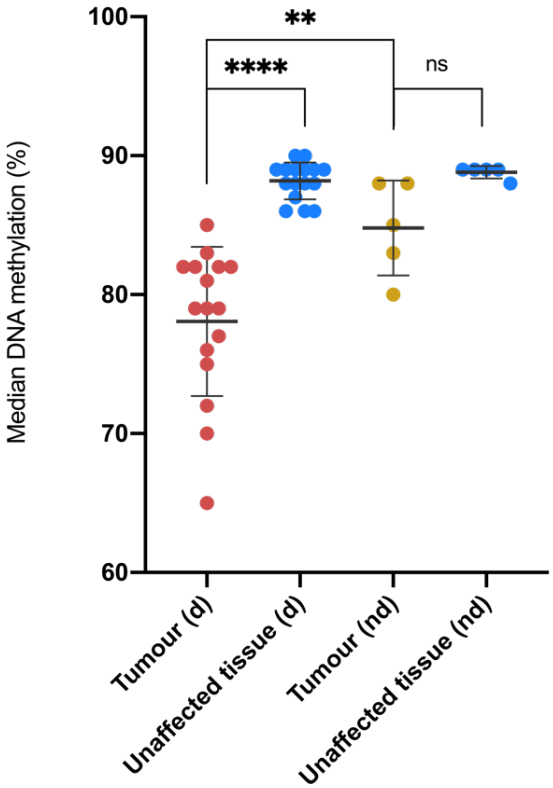


Figure 5. 3| Median CpG DNA methylation percentage per sample. The figure shows median DNA methylation percentage per sample in males with deficient Y chromosome gene expression (d) and males lacking this feature (nd) for tumours with normal paired data available (n=21). Data is shown for both tumour and histologically normal tissue. Normality was assessed with a Shapiro Wilk test. Differences in DNA methylation between paired tumour and histologically normal tissues were assessed using a two-tailed paired t-test (low Y group), and a Wilcoxon test (non-low Y group). A two-tailed unpaired Mann-Whitney test was used to assess differences in DNA methylation between the two tumours groups. Error bars represent standard deviation from the mean. Magnitude of significance is denoted with asterisks (*) ** P = 0.0082, **** P = 3.12 x10-6. Abbreviations: d (deficient chromosome Y gene expression), nd (non-deficient chromosome Y gene expression), ns (non-significant). The figure is reproduced from Willis-Owen *et al.*[1].

Autosomal DNA methylation levels were also significantly lower in LYE male tumours as compared with non-LYE tumours (P = 0.0082). These results demonstrated coincidence between reduced Y chromosome gene expression and widespread autosomal DNA hypomethylation in the same patients. The results also suggest reduced KDM5D activity as a potential mechanism leading to the widespread DNA methylation loss observed here.

Through the examination of individual regions, showing significant differential methylation between LYE tumours and unaffected paired tissues, the cancer-associated changes in DNA methylation were found to be strongly biased in favour of hypomethylation. For instance, promoter regions 1Kb upstream of 1,728 genes were found to be hypomethylated in LYE tumours with methylation differences exceeding 20%. As shown in Table 5.1 these regions showed significant enrichment for multiple motifs relating to the dimeric Activating Protein 1 (AP-1) transcription factor complex. The latter has established roles in malignant transformation and invasion[270].

Hypomethylated Genomic Regions

Motif Name	Consensus	*P*-value	Target Sequences with Motif (%)
AP-1(bZIP)/ThioMac-PU.1-ChIP-Seq(GSE21512)/Homer	VTGACTCATC	1.83E-06	23.37%
Atf3(bZIP)/GBM-ATF3-ChIP-Seq(GSE33912)/Homer	DATGASTCATHN	1.74E-06	21.38%
BATF(bZIP)/Th17-BATF-ChIP-Seq(GSE39756)/Homer	DATGASTCAT	2.95E-05	20.47%
Fos(bZIP)/TSC-Fos-ChIP-Seq(GSE110950)/Homer	NDATGASTCAYN	1.65E-07	20.23%
Fra1(bZIP)/BT549-Fra1-ChIP-Seq(GSE46166)/Homer	NNATGASTCATH	3.37E-06	18.72%
JunB(bZIP)/DendriticCells-Junb-ChIP-Seq(GSE36099)/Homer	RATGASTCAT	9.19E-07	18.30%
Fra2(bZIP)/Striatum-Fra2-ChIP-Seq(GSE43429)/Homer	GGATGACTCATC	2.95E-08	16.73%
Oct6(POU,Homeobox)/NPC-Pou3f1-ChIP-Seq(GSE35496)/Homer	WATGCAAATGAG	5.84E-05	16.18%
Fosl2(bZIP)/3T3L1-Fosl2-ChIP-Seq(GSE56872)/Homer	NATGASTCABNN	3.31E-07	11.41%
Six1(Homeobox)/Myoblast-Six1-ChIP-Chip(GSE20150)/Homer	GKVTCADRTTWC	5.24E-04	8.82%
Jun-AP1(bZIP)/K562-cJun-ChIP-Seq(GSE31477)/Homer	GATGASTCATCN	7.62E-08	8.51%
Bach2(bZIP)/OCILy7-Bach2-ChIP-Seq(GSE44420)/Homer	TGCTGAGTCA	8.21E-06	7.07%
Six4(Homeobox)/MCF7-SIX4-ChIP-Seq(Encode)/Homer	TGWAAYCTGABACCB	6.76E-04	2.23%

Hypermethylated Genomic Regions

Motif Name	Consensus	*P*-value	Target Sequences with Motif (%)
X-box(HTH)/NPC-H3K4me1-ChIP-Seq(GSE16256)/Homer	GGTTGCCATGGCAA	0.009168	12.00%

Table 5. 1| HOMER known motifs enriched in hypomethylated (a) and hypermethylated (b) promoter regions. Motif names are designated by the transcription factor name and its DNA binding domain, followed by the GEO Accession number for the public Chromatin Immunoprecipitation Sequencing (ChIP-Seq) experiment from the Genomic Spatial Event (GSE) database. Abbreviations: basic Leucine Zipper Domain (bZIP); POU, Homeobox (POU-domain homeobox transcription factor). This table is reproduced from Willis-Owen, S. A. G. *et al*[1].

In addition to this hypomethylation signature, a total of 473 promoter regions were significantly hypermethylated. These sites showed significant enrichment for an X-box motif (Table 5.1) that is recognised by RFX transcription factors and has functions in cellular specialization and terminal differentiation with particular relevance to ciliogenesis[271].

Increase of methylation at CpG islands has been previously associated with aging and cancer whilst decreased methylation has been observed in CpG oceans or intergenic regions[272], and a similar trend was observed in LYE tumours as compared to their normal matched tissue (Supplementary Fig. 5.1).

The findings presented next for this study are not (unlike the above) included in the publication[1] but show further research that support the conclusions reached by Willis-Owen et al. through the integration of data from the Targeted Capture Sequencing (TCS)(see Chapter 3).

5.3.3 Molecular Characterization of LYE Tumours

Somatic mutation variant calling as previously described (Chapter 2, Section 2.4.2) enabled comparison of somatic alterations present in low Y-expressing tumours (*n*=16) and tumours lacking this feature (*n*=102).

The average number of variants per sample appeared much higher in the LYE group, with an average of 102.94 mutations per sample compared to a mean of 40.49 mutations per sample in tumours lacking the feature. In addition, the topmost frequently mutated genes amongst those scanned through the targeted capture gene panel (*n*=52 genes) were different between LYE samples as compared to non-LYE samples (Fig. 5.4). *TP53* was still commonly the most recurrently mutated gene in both groups, followed by *NF1* (25%), *RBM10* (19%) and *KRAS* (19%) in LYE tumours; and *KRAS* (25%), *STK11* (17%) and *KEAP1* (16%) in non-LYE tumours.

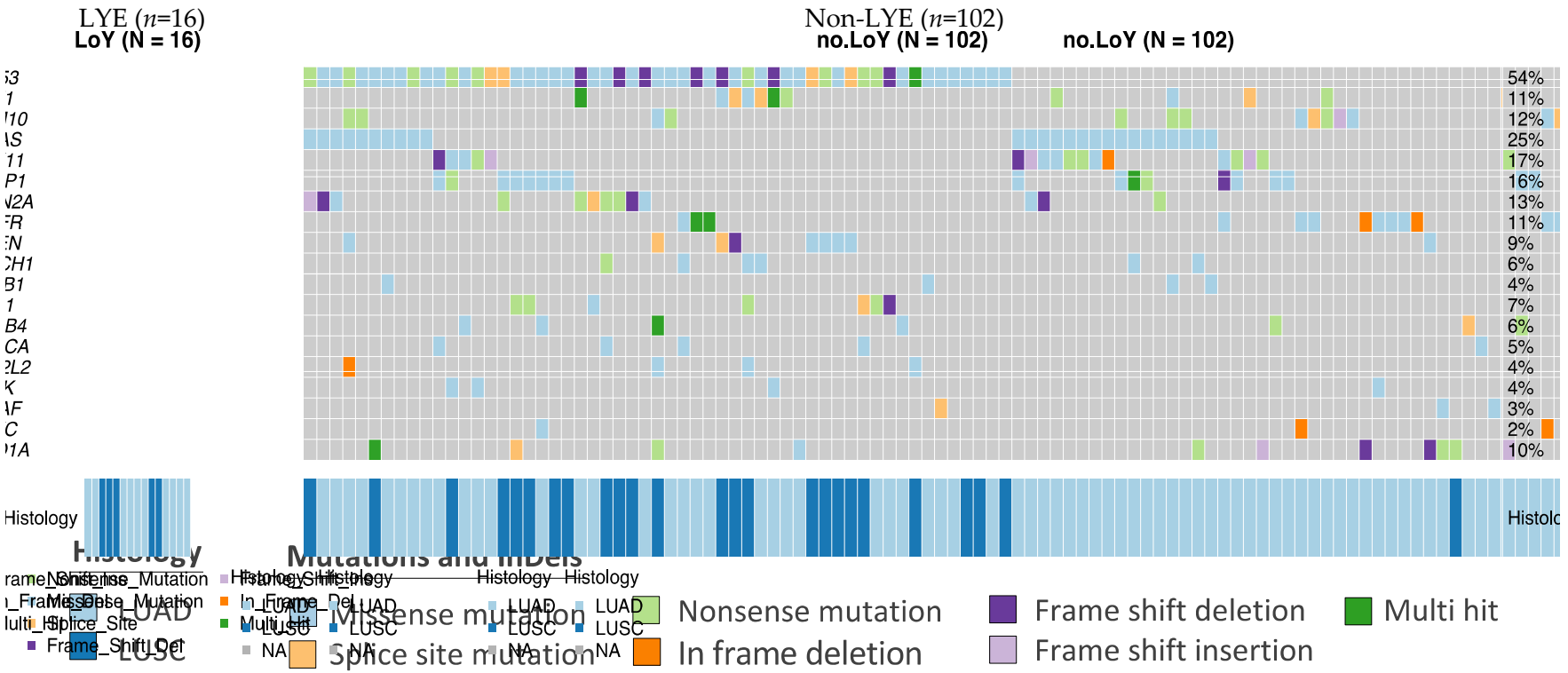


Figure 5. 4 | Oncoplots of the top mutated genes detected by TCS and WES in LYE (left) and non-LYE (right) NSCLC tumours. Each column represents a different patient's tumour and in rows are listed the 19 genes altered out of the total 52 genes of the panel. Genomic alterations are coloured by type of mutation and InDel (as per colour key); and percentages of patients with each altered gene are shown for the two groups on the right hand-side of the oncoplots. Bottom bar indicates LUAD and LUSC histotype as per the colour key shown. Abbreviations: LUAD, Lung Adenocarcinoma; LUSC, Lung Squamous Cell Carcinoma.

In addition, when including WES data (Chapter 2, Section 2.5.1) three other genes appeared frequently mutated in the LYE group. Specifically, *TTN* (31%), *CSMD3* (25%) and *USH2A* (19%). This finding is consistent with what has been observed and reported by The Cancer Genome Atlas[273].

The human gene *TTN* encodes for the TITIN protein. It is known as a major mutation gene (with the second highest rate after *TP53* genes) in many types of tumours including NSCLC[273]. Similarly the *CSMD3* gene, encoding for the CUB And Sushi Multiple Domains 3 involved in dendrite development, has been reported as the second most frequently mutated gene in an alternative NSCLC dataset by Liu *et al.*[274]. *USH2A* provides instructions for the usherin protein, that serves as a structural component of basement membranes and has been associated with increased risk of breast cancer. Nevertheless, *USH2A* is no longer regarded a driver gene given its association with Usher syndrome and a high discovery rate in mutation calling[275].

Next the two groups of tumours (LYE and non-LYE) were compared to detect significantly differentially mutated genes. This analysis identified *TP53* and *TTN* to be significantly differentially mutated between LYE and non-LYE tumours (P_{TP53} = 0.013; OR= 5.91 [95% CI= 1.26-56.16]), (P_{TTN} = 0.023; OR= 4.61 [95% CI= 1.03-18.99]) as shown in Figure 5.5.

These observations are in line with previous data (by Thomson *et al.*) exploring genetic predisposition to mosaic loss of chromosome Y (LOY) in blood, where genetic variants in genes involved in cell cycle and mitotic processes (including *TP53*) were associated with LOY by using GWAS. This association suggests that clonal expansion of LOY cells requires a permissive environment (or 'soil') in which proteins that are involved in sensing and activating cell death signalling cascades are dysregulated[276]. The Thomson *et al.* study[276] adds to prior studies[277,278] and provides evidence that LOY can be a biomarker of cell-cycle efficiency and the DDR in leukocytes. As a result, the molecular alterations uncovered in this present study could be part of an underlying mechanism leading to genomic instability and cancer across this cell type.

Following these premises, next (see Section below) the altered signalling pathways in LYE and non-LYE tumours were examined to investigate whether genes in pathways related to the cell cycle, programmed cell death and DDR were altered at the genetic and copy number (CN) level in tumours with deficient Y chromosome gene expression.

5.3.4 Oncogenic Signalling Pathways Enriched in LYE and Non-LYE Tumours

TP53 and RTK-RAS pathways appeared as the signalling pathways most altered by mutations and InDels in both groups (LYE *n*=16 and non-LYE *n*=102, Fig. 5.5). The TP53 pathway, however, was more altered in the LYE group (87.5%) than in the non-LYE (81.3%), whereas RTK-RAS was more altered in the non-LYE group (70.59%) as compared to the LYE group (57.84%). Furthermore, the PI3K pathway was similarly frequently altered in both groups of NSCLC male tumours with 37.5% and 38.2% in the LYE and non-LYE group respectively.

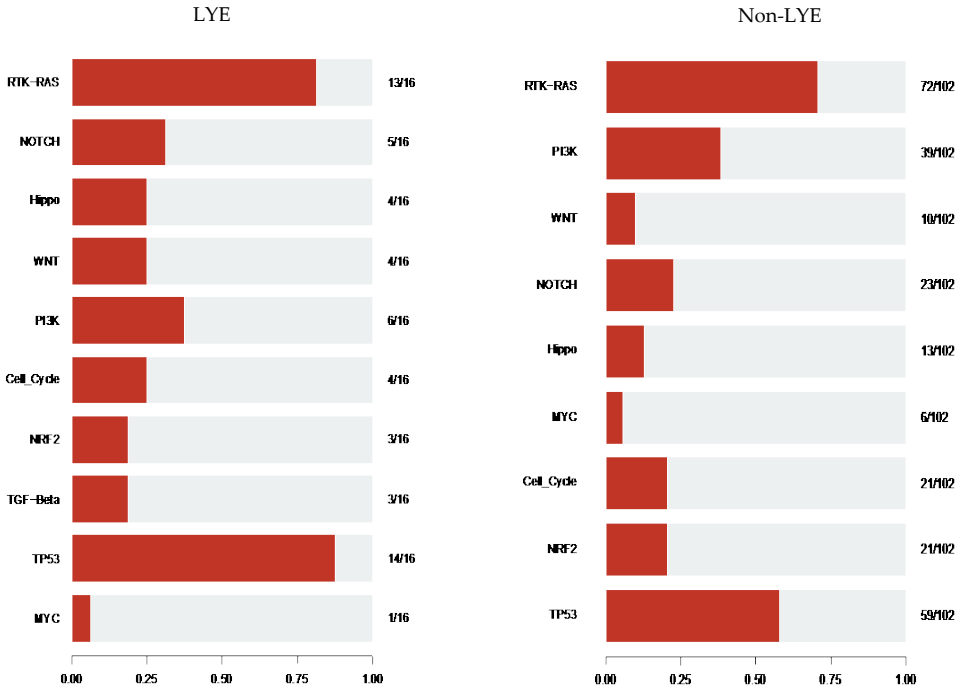


Figure 5. 5| Enrichment of known Oncogenic Signalling pathways in LYE (*n*=16) and non-LYE NSCLC tumours (*n*=102). Fraction of samples with mutations and/or InDels in genes of each signalling pathway are shown.

Looking at the genes altered within the RTK-RAS pathway, the LYE group showed more mutations in *NF1* (25%) than in *KRAS* (18.7%), followed by *EGFR* (12.5%). Conversely, non-LYE samples mostly showed mutations in *KRAS* (25.5%), *EGFR* and *NF1* (10.8% each).

5.3.5 Genetic Alterations in DNA Damage Repair Genes

KDM5D is associated with augmented cell cycling and accumulation of stalled replication forks, culminating in DNA-replication stress and activation of the DDR pathway[279,280]. Next mutations in DDR genes were sought for by scanning the sequencing data generated for the DDR genes included in the updated table of Human DNA repair genes cited in Wood *et al.* [281].

p53 is a key player in the DDR pathway acting through promotion of cell cycle arrest and thereby allowing for DNA repair, senescence or apoptosis. As mentioned above (Section 5.3.3) *TP53* appeared significantly differentially mutated between males with low Y-expressing tumours and tumours lacking this feature (88% in LYE v/s 54% in non-LYE; P = 0.013). Additionally, *PTEN* was also detected mutated at a slightly higher frequency in the LYE group of male tumours with a frequency of 12% compared to 9% in the non-LYE group of males (Fig 5.5). *PTEN* helps in DSB repair and nucleotide excision repair[282] as well as interacting with p53[283] .

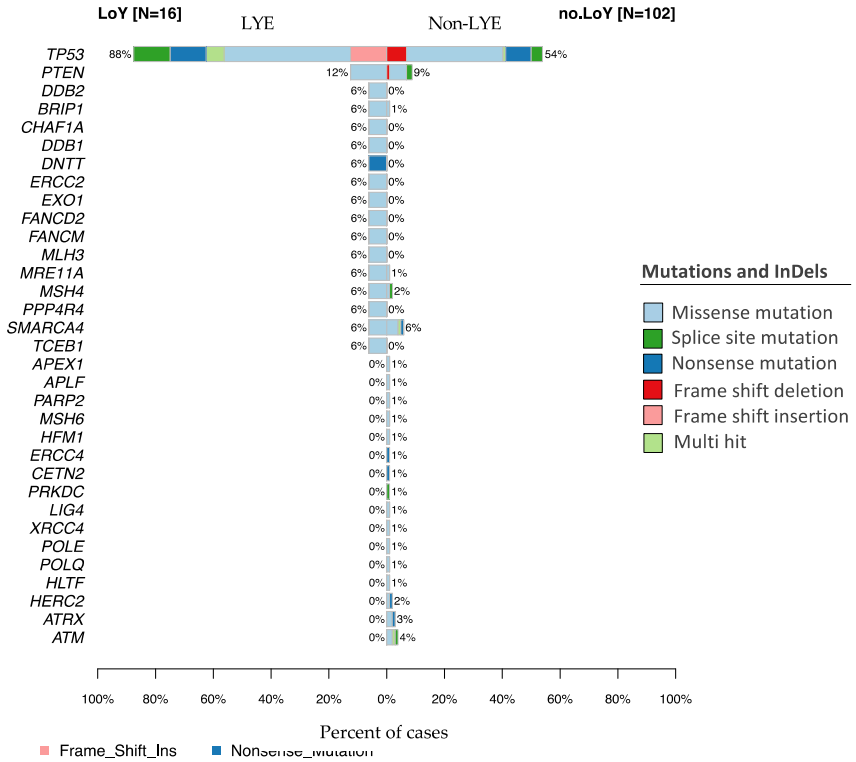


Figure 5. 6| Frequency of mutations and InDels in DNA damage repair genes in LYE (*n*=16) and non-LYE group of male tumours (*n*=102). Types of mutation and insertions and deletions coloured as per colour code key.

As shown in Figure 5.6, several other DDR genes were found to be mutated in LYE tumours with frequencies of 6% that were notably not mutated in the non-LYE group.

Furthermore, Copy Number Alterations (CNA) in DDR genes were identified through the analysis of SNP genotyping data (see Chapter 3). The number of homozygous and heterozygous calls were queried for each tumour sample. CNAs were detected in DDR genes at high frequencies. This is consistent with accumulated evidence from various tumour classes that the redistribution, or perturbation, of DNA methylation mainly occurs upon copy number alteration[272] and that such redistribution can be induced by oxidative damage[284].

The average CN status of DDR genes was calculated based on the heterozygous/homozygous deletions and amplifications that ranged from -2 for heterozygous deletions to +2 for homozygous amplifications. The CN status of DDR genes was significantly different (P = 0.0021) between the two groups of male tumours (Fig. 5.7; Supplementary Fig. 5.7).

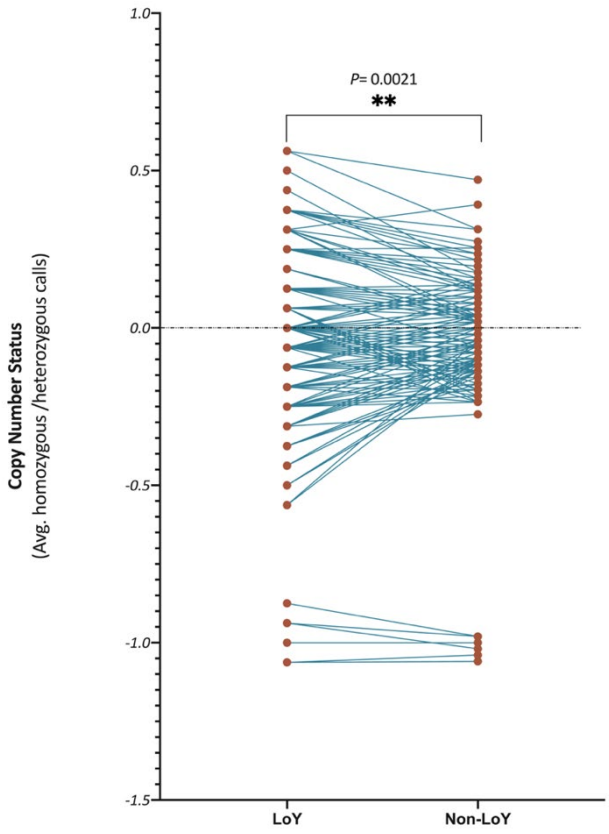


Figure 5. 7|Copy Number status of DDR genes in LYE and non-LYE NSCLC tumours. Dots represent the average CN status at each gene based on the number of heterozygous/homozygous calls.

5.3.6 APOBEC Enrichment

APOBEC enzymes are a well-known source of mutations and DNA damage. *TP53* mutation status and CNAs have been associated with *APOBEC* mutagenesis, intratumor heterogeneity[285] and metastatis[286] in NSCLCs. In line with these findings, in this present study enrichment of *APOBEC* related mutations in 50% of LYE tumours in contrast with 6.67% in tumours lacking this feature was observed. Known and *de novo* COSMIC mutational signatures (CMS) were then identified for the subset of samples that underwent WES (Chapter 3, Sections 3.3.1-3.3.2) allowing more accurate inference of mutational profiles through whole-exome sequencing data (*n* LYE=6; *n* non-LYE=15).

Half of the LYE tumours carried signature CMS 13 (identified with deconstuctSigs), a signature attributed to activity of AID/APOBEC family of cytidine deaminases (Supplementary Figures 5.2). Furthermore, *de novo* mutational signatures were found to be similar to CMS 4 (cosine similarity of 0.932) and CMS 2 (cosine similarity of 0.867), signatures that are related to exposure to tobacco mutagens and APOBEC cytidine deaminases respectively (Supplementary Fig. 5.4).

In contrast only 26% of tumours showing normal Y expression harboured CMS 13, and none exhibited the *de novo* signatures associated with APOBEC activity but instead were related to signature CMS 1 of unknown aetiology (cosine similarity of 0.787) and CMS 4 (cosine similarity of 0915) associated with smoking (Supplementary Figs. 5.3 and 5.4).

5.3.7 TMB/CNB and LYE Relationship

In light of the different mutational signatures and spectrum of genomic alterations observed between LYE and non-LYE NSCLC tumours, overall Tumour Mutational Burden (TMB) and copy number burden (CNB) were investigated (Fig. 5.8).

TMB was nominally raised in LYE tumours (median TMB = 30.97) relative to non-LYE tumours (median TMB = 14.12), although this difference did not quite reach significance (P = 0.0548). CNB was also modestly higher in LYE tumours (median CNB = 12.63) as compared with tumours lacking this feature (median CNB = 4.69) but again this was not statistically significant (P = 0.5693).

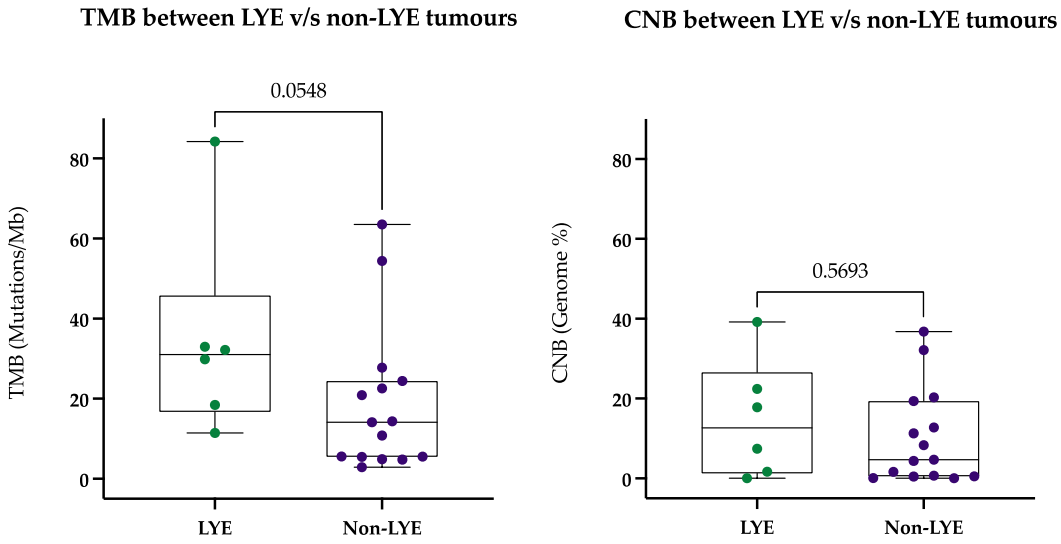


Figure 5. 8| TMB and CNB in male NSCLC tumours. The figure shows TMB and CNB per sample in males with deficient Y chromosome gene expression (LYE) and males lacking this feature (non-LYE). Normality of the data was examined through Shapiro-Wilk normality tests. Two-tailed unpaired Mann-Whitney tests were performed to assess differences in TMB and CNB between the two tumour groups. The samples shown correspond to samples that underwent WES for more accuracy (*n* LYE=6; *n* non-LYE=15).

TMB and CNB were, however, both negatively associated with *KDM5D* expression as shown in Figure 5.9, and this association attained significance for the TMB-*KDM5D* expression relationship (P = 0.027). These data further demonstrated that a decrease in *KDM5D* activity was associated with an increased genomic instability in LYE tumours.

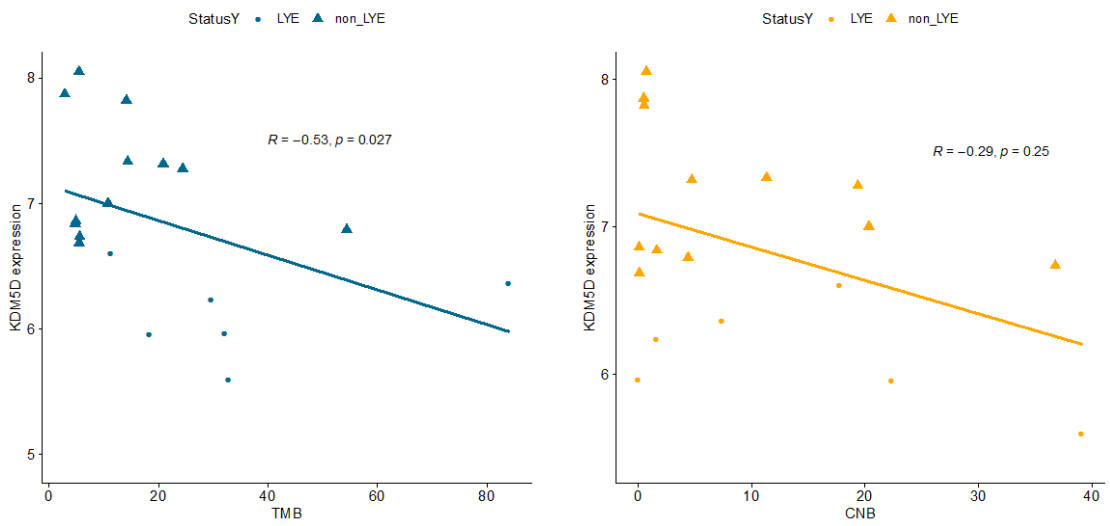


Figure 5. 9| KDM5D expression correlations with TMB and CNB in NSCLC male tumour samples. Spearman correlation coefficients and two-tailed *P*-values are shown for each association. Tumours with low Y expression are shown with circles and tumours lacking this feature are shown in triangles.

5.4 Discussion

The paper by Willis-Owen *et al.*[1] (that included some of the data detailed in this present chapter) identified a gene co-expression network that was associated with biological sex in a group of NSCLC tumours. The male specific H3K4 demethylase, *KDM5D*, was found to be central to this network with significant prognostic value. This was validated in a further 1,100 male lung cancer samples. These observations are consistent with increased mortality and higher risk of cancer found in men with loss of Y chromosome in peripheral blood[287,288].

Integration of DNA sequencing and WGBS sequencing, as described in this chapter, reveals a link between low Y chromosome expression, DNA methylation loss and genomic DDR and APOBEC signatures. Analysis of read depth from WES and WGBS experiments showed that the reduced expression in Y-chromosome transcripts co-occurs with, and therefore likely results from, a somatic loss of Y chromosome. A polymerase chain reaction (PCR) -based chromosome deletion detection assay further corroborated partial somatic deletion of the Y chromosome in these tumours.

Expanding upon these observations, the genes *TP53* and *TTN* are significantly differentially mutated in tumours with deficient Y gene expression. This observation together

with DDR and APOBEC mutational signatures may create a permissive environment for cells to proliferate, thus strengthening the link between LYE and increased mortality in male NSCLC patients. For instance, *TTN*/*TP53* co-mutation has previously been suggested as an effective predictor for OS and chemotherapy response in lung cancer[289], as well as a potential predictive marker of immunotherapy for patients with LUAD[290]. *TP53* is also a well-known DDR gene with critical roles in cell-cycle and survival, and in addition LOY has been previously associated with DNA-replication stress and activation of the DDR[291,292].

Nevertheless, APOBEC-related mutations were found to be enriched in LYE tumours. Cigarette smoke contains established pulmonary free radicals, carcinogens, mutagens and tumour promoters, that induce a variety of oxidative damage, inflammation, DNA adducts and single- and double-strand DNA breaks (SSBs, DSBs)[293,294]. SSBs and DSBs are substrates for APOBEC enzymes which are known to be highly mutagenic[295,296]. Exposure to tobacco within a more permissive environment could potentially facilitate genomic instability. Hence, the *TP53* mutations and CNAs in DDR genes together with an APOBEC mutational signature may partially explain deficient chromosome Y expression and the higher TMB detected in LYE tumours as compared with non-LYE tumours.

LYE tumours also feature a DNA methylation loss signature. Amongst histone methylation marks, H3K4me3 specifically is anti-correlated with DNA methylation[267] and mutations in the X-linked *KDM5D* homolog (*KDM5C*) have been linked with multi-locus DNA methylation loss[268] providing a plausible explanation for the DNA methylation loss detected in LYE tumours in this present study. The functional role of DNA methylation, for example whether DNA methylation is a passive mark of transcription activity or an active regulator that modifies gene expression, is however still under debate as discussed previously in Chapter 4. Evidence accumulated from various tumour classes also points to a redistribution or perturbation of DNA methylation upon CNAs[272], and oxidative damage has also been shown to be at the basis of widespread chromosomal loss of DNA methylation thereby contributing towards an unstable genome[284].

The spectrum of mutations, CNAs and mutational signatures identified in this present study suggests the potential role of *KDM5D* status to predict response to chemotherapeutic agents targeting DNA damage and repair pathways. For instance, low expression of *KDM5D*

is associated with a reduced sensitivity to cisplatin and increased sensitivity to inhibitors of the ataxia telangiectasia and Rad3-related protein (ATR) in PC cell lines[279]. Additionally inactivation of p53 has been found to circumvent APOBEC3B-induced cell cycle arrest and maintain a kataegic mutational signature and DDR biomarkers, and sensitize cells to a platinum salt, cisplatin used in combination with ATR inhibitors (ATRi) as well as PARP inhibitors used in combination with ATRi[297–299].

In addition, in this present study low tumour *KDM5D* expression was associated with an increased relative hazard of death as compared with males with normal *KDM5D* expression (HR 4.92 [95% CI 1.46,16.55], $P = 0.01$), and when compared with both males and females (HR 3.80 [95% CI 1.40 - 10.3], $P = 0.009$). These results were also validated in 1,100 male tumours from 11 independent LC mRNA gene expression datasets[1].

Finally, *TTN*/*TP53* co-mutation was suggested as an effective predictor for OS and chemotherapy response in lung cancer, as well as a potential predictive marker of immunotherapy for patients with LUAD[290].

Several important limitations may impact the results described here. These include small sample size for the identification of DNA methylation changes and a lack of information describing cell type composition and purity of tumour samples. Studying different types of -omic data to infer both tumour purity and cell type composition has been shown to be a promising approach and may avoid its cofounding effect in future studies. This study has also analysed two histological subtypes of NSCLC which, as outlined in Chapter 3, show a different repertoire of mutations, InDels and CNAs. Nevertheless, the reduced Y chromosome gene expression feature reported here occurs in both LUAD and LUSC histologies independent of stage and histology. Further research should therefore investigate this phenomenon in larger sample sizes, accounting for tumour purity, cell composition, stage and histological subtype. Finally, the SNP genotyping arrays used for this study may not fully represent the copy number burden genome-wide because the majority of probes on the arrays are designed for the purposes of SNP detection. Recently developed methods now enable the detection of the same set of SNPs included in this analysis plus an additional set of probes targeting non-polymorphic positions. This may improve accuracy of CNA detection and show whether LYE tumours also lean towards genomic instability following CN changes.

Chapter 6: Lung Carcinoid Molecular Subtypes from Transcriptomic Data

This chapter describes and extends the preprint "Distinct pancreatic and neuronal Lung Carcinoid molecular subtypes revealed by integrative omic analysis"[300] currently under peer-review process (full details of which are given on page 6).

6.1 Introduction

The World Health Organization (WHO) of Thoracic Tumours[4] currently classifies L-CDs under the umbrella of Neuroendocrine Neoplasms (NENs) together with Small Cell Lung Cancer (SCLC) and Large Cell Neuroendocrine Carcinoma (LCNEC). L-CDs occur frequently in never-smokers and are subclassified as Typical Carcinoids (TC) and Atypical Carcinoids (ACs). ACs grow a little faster and are more likely to metastasize to other organs. TCs grow slowly and rarely spread beyond the lungs. Thus TCs and ACs are considered low grade and intermediate grade respectively and classified under the term NETs. LCNECs and SCLC classify as high-grade neuroendocrine carcinomas under the term Neuroendocrine Carcinomas (NECs).

These groupings, based on histopathology, have been supported by recent genomic and mutational studies showing that similar genes are altered, although at different prevalence between NENs with L-CDs consistently showing the lowest number of molecular alterations[49,63,144,301,302]. As a result there is controversy around the possibility that high-grade neuroendocrine tumours could arise from pre-existing L-CDs[49,303]. Recent studies, however, have argued that L-CDs may not be early progenitor lesions of other neuroendocrine tumours since some altered genes and pathways do not overlap between these classifications[304,305]. Furthermore, differentiating TCs from ACs is challenging and involves a board of experts due to the disagreement for the optimal diagnosis[306] and interobserver variability on mitotic count and degree of necrosis[307]. This lack of consensus results in unspecific treatment and lack of disease specific support. Consequently many patients with rare NETs experience a "no clear pathway" of care in their cancer journey[308].

Despite the central role of epigenetic alterations in L-CDs, there are no studies that have investigated DNA methylation alterations genome-wide and integrated the data with genetic, transcriptomic and clinical data in a comprehensive manner. Some studies have integrated genetic (WGS or WES or TCS) with transcriptomics data[305] and a recent study integrated methylation array (850K) data with transcriptomic data comparing TCs, ACs and LCNECs using a Machine Learning (ML) approach[62]. While most LCNECs clustered in a separate subtype, TCs and ACs were indistinctively allocated to the same groups. This highlights the need for further research to refine the current histopathological classification.

L-CDs remain relatively understudied despite their increasing incidence[309–315]. An integrated and comprehensive characterization at the molecular level of these malignancies has the potential to provide novel deeper insights on the distinct mechanisms of dysregulation as well as opening up new avenues for biomarker selection and treatment opportunities for L-CD patients.

The identification of CNAs and DNA methylation changes in L-CDs pinpoints alternative non-mutation processes leading to carcinogenesis. Thus, the aims of this chapter were to take the data generated previously (Chapters 3 and 4) and integrate it with expression and clinical features.

6.2 Molecular Classification from Transcriptomic Data

Unsupervised clustering and Principal Components Analysis (PCA) analysis of RNA sequencing data (Chapter 2, Section 2.7) from 15 L-CD tumours (comprising 3 ACs and 12 TCs) was conducted and clearly differentiated the 15 tumours into two approximately equally sized groups (n=8 and 7 respectively). These two groups were reproducibly identified using data from the top 500 most variable genes or all sequenced genes (n 25,764) (Fig. 6.1a-b).

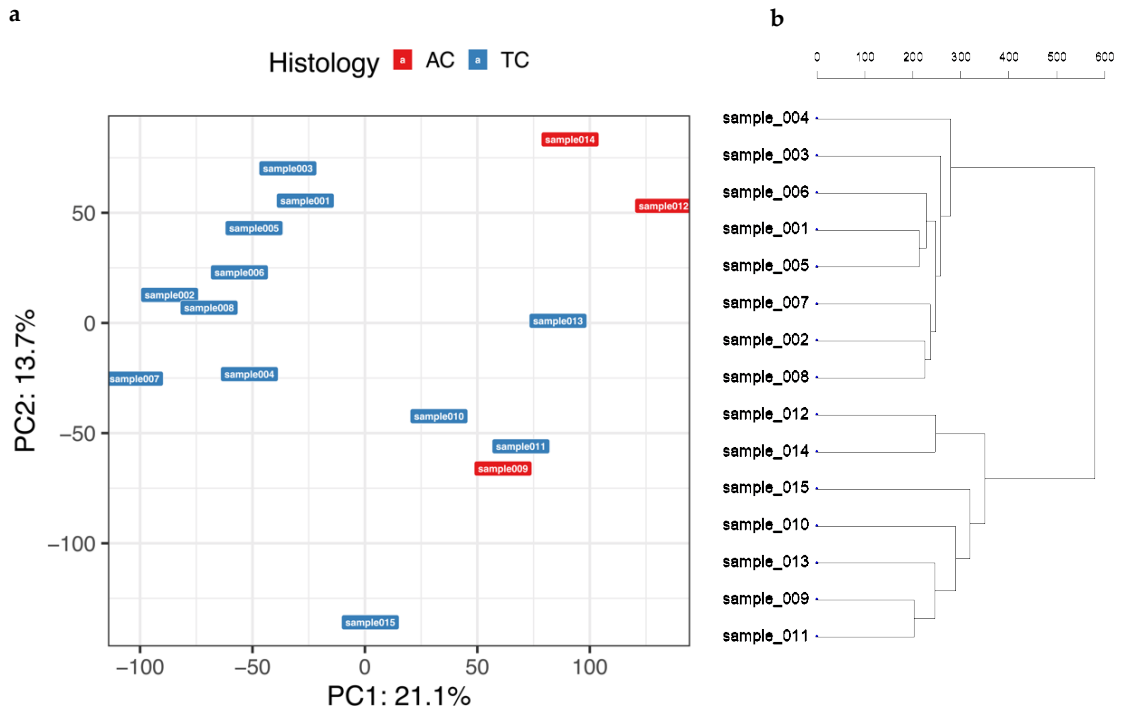


Figure 6. 1| Clustering of L-CD tumour RNA-sequencing data. a) Principal Component Analysis and b) Dendrogram of sample similarity based on all sequenced genes (*n* 25,764) obtained using the War.D2 algorithm[316] with *hclust*. Two groups of L-CD tumours clustered into two separate groups based on their expression profiles. Abbreviations: TC (Typical Carcinoid; AC (Atypical Carcinoid); PC (Principal Component).

The clustering results were further evaluated and validated with the clValid R package (Chapter 2, Section 2.7.8). Specifically, internal and stability measures were obtained by using the top 500 most variable genes, together with the optimal number of clusters (Table 6.1).

	Measure	Score	Method	Optimal Number of Clusters
Top 500 most variable genes	Connectivity	6.0833	hierarchical	2
	Dunn	0.9540	hierarchical	2
	Silhouette	0.3639	hierarchical	2

Table 6. 1| Hierarchical clustering validation measures and optimal number of clusters identified using top 500 most variable genes using the clValid R Package. The optimal number of clusters using the hierarchical clustering algorithm was 2, thus validating the results obtained with *hclust* and PCA methods.

Looking next at the two groups (Group 1: L-CD-PanC and Group 2: L-CD-NeU – assigned labels explained below) and gene expression levels (Chapter 2, Section 2.7.6) there was substantial DE between the two groups. There were 1,924 DE transcripts that achieved significance at a 1% FDR threshold (Fig. 6.2). The top 20 differentially expressed transcripts are listed in Table 6.2.

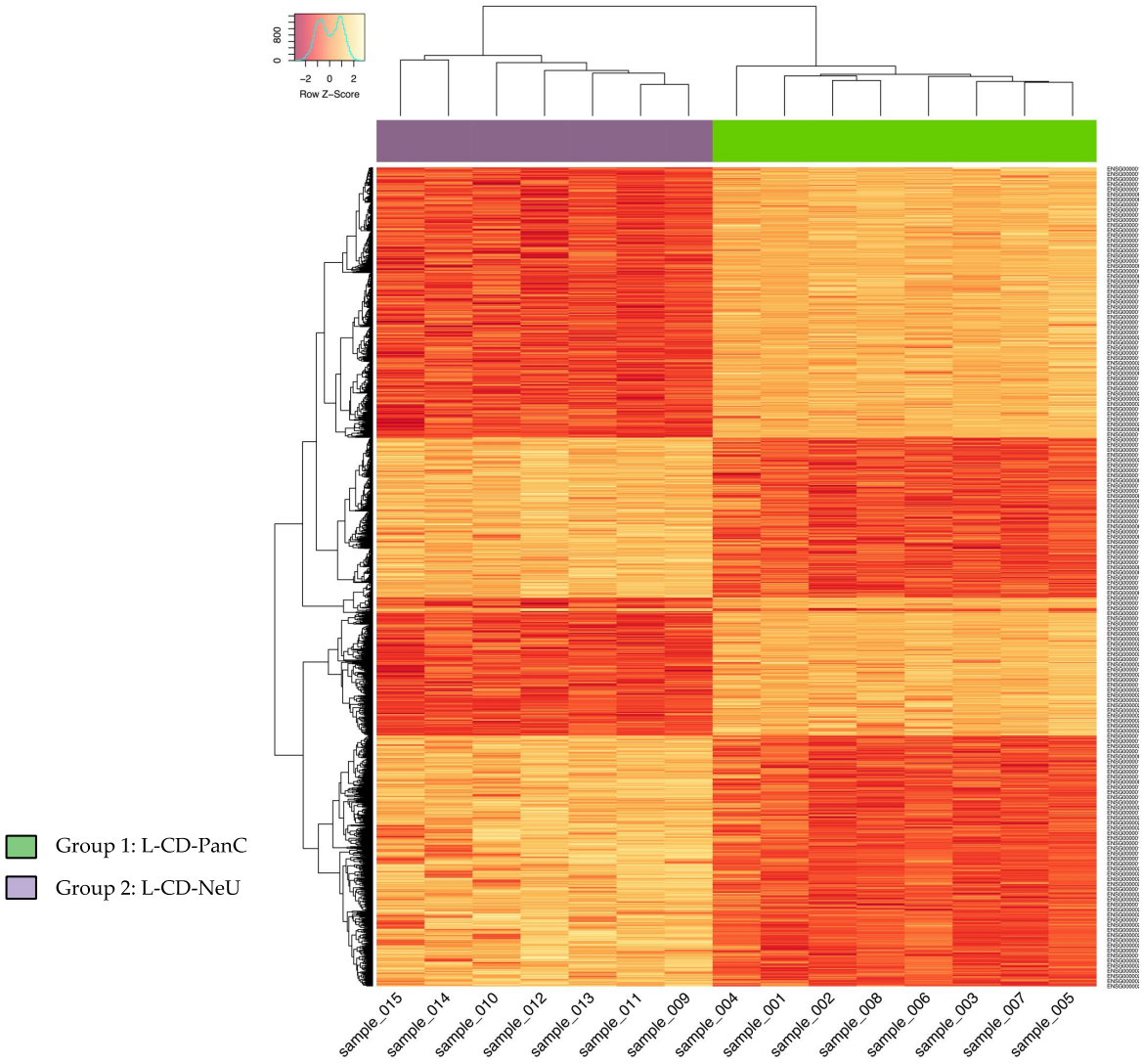


Figure 6. 2| Heatmap of the significantly differentially expressed genes (*P* <0.01) between L-CD subtypes. Figure displays heat map and dendrograms with hierarchical clustering of 15 tumour lung carcinoids (on the *X* axis) and genes (on the *Y* axis) that were significantly differentially expressed. Top bar indicates lung carcinoid group membership. Abbreviations: L-CD-PanC, Lung-Carcinoids with pancreatic expression profiles; L-CD-NeU, Lung-Carcinoids with neuronal expression profiles.

Gene	*Log₂FC (CI)*	*AveExpr*	*P.Value*	*adj.P*	*PanC*	*NeU*
A1CF	10.63(6.53,14.73)	3.24	2.94E-13	1.89E-09		
TM4SF5	10.50(10.25,10.73)	1.46	2.02E-14	3.39E-10		
SERPINA10	9.77(8.37,11.17)	1.12	4.39E-12	1.03E-08		
RDH12	9.65(8.00,11.30)	4.20	8.05E-13	3.69E-09		
RFX6	9.42(8.65,10.20)	1.44	2.12E-13	1.82E-09		
HNF1A	8.33(8.04,8.63)	1.16	2.63E-14	3.39E-10		
C2orf72	7.30(6.79,7.81)	3.41	2.02E-12	5.79E-09		
FOXA3	7.24(6.91,7.58)	0.60	8.18E-12	1.76E-08		
ARHGEF10L	3.02(2.54,3.51)	6.78	1.00E-12	3.69E-09		
GLYCTK	2.91(2.21,3.62)	4.82	1.18E-12	3.79E-09		
RFC3	-2.44(-4.40, -0.47)	4.29	1.09E-08	3.65E-06		
AUTS2	-2.51(-2.91, -2.11)	4.99	1.51E-08	4.62E-06		
ATP8A1	-2.89(-3.67, -2.11)	5.83	3.62E-09	1.56E-06		
DPYSL3	-5.08(-6.09, -4.06)	7.31	1.10E-07	2.08E-05		
DOK7	-5.53(-5.99, -5.07)	-1.07	1.30E-08	4.23E-06		
PRUNE2	-5.58(-6.22, -4.94)	4.63	4.71E-08	1.18E-05		
FAM3B	-6.21(-7.12, -5.29)	0.98	8.76E-08	1.76E-05		
HS3ST6	-6.27(-6.71, -5.82)	-0.39	2.82E-08	7.66E-06		
KCNK10	-9.52(-9.86, -9.18)	0.84	3.03E-09	1.40E-06		
RALYL	-9.70(-10.31, -9.08)	0.21	1.73E-10	1.78E-07		

Table 6. 2| The top 20 transcripts differentially expressed between PanC and NeU L-CDs (adj.*P* <0.01). Abbreviations: log₂FC (log₂ fold change); CI (confidence interval); AveExpr (Average expression across all samples in log2-counts per million); adj.*P* (Benjamini-Hochberg false discovery rate adjusted *P*-value). Coloured cells indicate relatively higher (red) or lower (blue) gene expression between subtypes.

Gene Set Enrichment Analysis (GSEA – Chapter 2, Section 2.7.7) revealed several metabolic pathways and hallmarks of pancreatic beta cells significantly enriched in Group 1 (Fig. 6.3), peaking at regulation of beta cell development (Normalised Enrichment Score [NES] 2.19, *P*< 0.01). In addition, Wnt signalling pathway genes, including *DKK4*, *KREMEN2* and *RNF43*; and genes related with aflatoxin activation and detoxification such as *GGT5*, *CYP34A*, *CYP3A5* and *DPEP1*, were also found upregulated in this group. Consistent with these themes, the genes showing the strongest evidence of differential expression, both with relatively raised expression in this group, were *TM4SF5* (log₂ fold change [log₂FC] 10.50, adj.*P*

= 3.39x10^{-10}) and *HNF1A* (log_2FC 8.33, adj.*P* = 3.39x10^{-10}). *TM4SF5* is a known tumorigenic factor in several cancer types, including liver, colon, pancreatic and esophageal cancers [317–322] whilst *HNF1A* encodes a putative master regulator of human pancreatic cancer stem cell properties[323]. The APOBEC1 complementation factor *(A1CF)* showed the highest log_2FC in expression as compared to Group 2, suggesting that A1CF protein expression could be used as a molecular marker. Group 1 therefore was labelled as L-CD-PanC.

Group 2 showed upregulation of various pathways involved in neuronal differentiation, peaking at serotonin neurotransmitter release cycle (NES -1,79, *P* = 0.004). In line with this pattern, significantly higher expression levels of various neuronal genes were observed in this group (Appendix, Supplementary Data 6.1). For example, *ASCL1* encodes a neuronal differentiation transcription factor and is a lineage-specific oncogene for high-grade neuroendocrine lung cancer[62,324]; whilst *SLIT1*, *ROBO1* and *SRGAP1* all represent members of the cell signalling protein complex slit/robo that latter being involved in axon guidance and angiogenesis. Moreover FAM3B/PANDER, a pleiotropic secreted cytokine that induces apoptosis in insulin-secreting beta-cells[325], was highly expressed in Group 2 (log_2FC 6.21, adj.*P* = 1.76x10^{-5}). PANDER is expressed ubiquitously, including lung[326] and some neurons of the brain[327], and it has recognised roles in invasiveness and tumorigenicity when overexpressed in prostate[328] and colon[329] cancers. Group 2 therefore was labelled as L-CD-NeU.

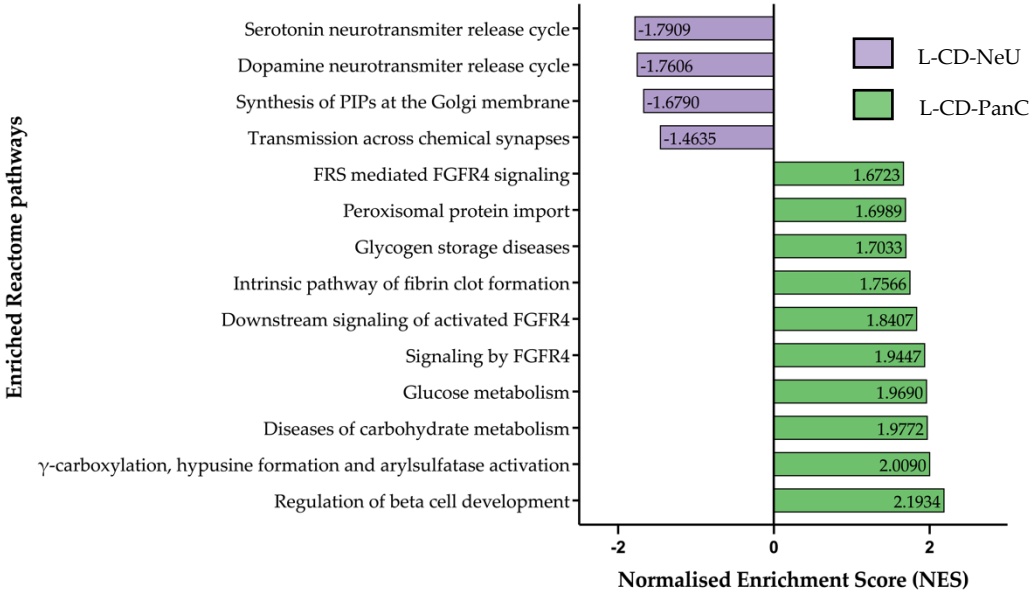


Figure 6. 3| Reactome pathways nominally enriched in L-CD molecular subtypes. Bar plot showing Reactome pathways achieving nominal significance for enrichment in a GSEA of RNA-sequencing data at a *P*-value threshold of 0.01. Using the MSigDB Reactome Canonical Pathways gene set[330], 10 enriched and 4 depleted pathways were identified in the L-CD-PanC subtype relative to the L-CD-NeU subtype. Each bar represents a pathway and its NES is given.

Further, immune cell content was explored in the L-CD dataset by scanning the transcript abundances of immune markers. A total of 25 transcripts were identified significantly differentially expressed between L-CD-PanC and L-CD-NeU tumours (*adj.P*≤0.05) (Supplementary Fig. 6.1). The expression of genes enriched in granulocytes, monocytes, dendritic cells and T-cells was significantly higher in L-CD-PanC tumours, whereas B-cell marker genes (3/5; 60%) were significantly enriched in L-CD-NeU tumours.

6.3 L-CD Subtypes are Associated with Histopathological Parameters

Integration of these findings with clinical parameters (Supplementary Table 6.1) revealed that L-CD-NeU tumours were characterised by a focal spindle cell morphology (two-sided Fisher's exact test: $P = 0.041$) and peripheral location (two-sided Fisher's exact test: $P = 0.007$) (Fig. 6.4) and included every tumour of an Atypical Lung Carcinoid histology (n=3). L-CD-PanCs were all located centrally within the bronchi. Whilst L-CD-NeU patients exhibited a trend towards an older age, this difference was not statistically significant (unpaired t-test, $P = 0.071$). No significant associations were seen with survival, smoking history, sex, presence of emphysema, nodal stage or either lymph or vasculature invasion.

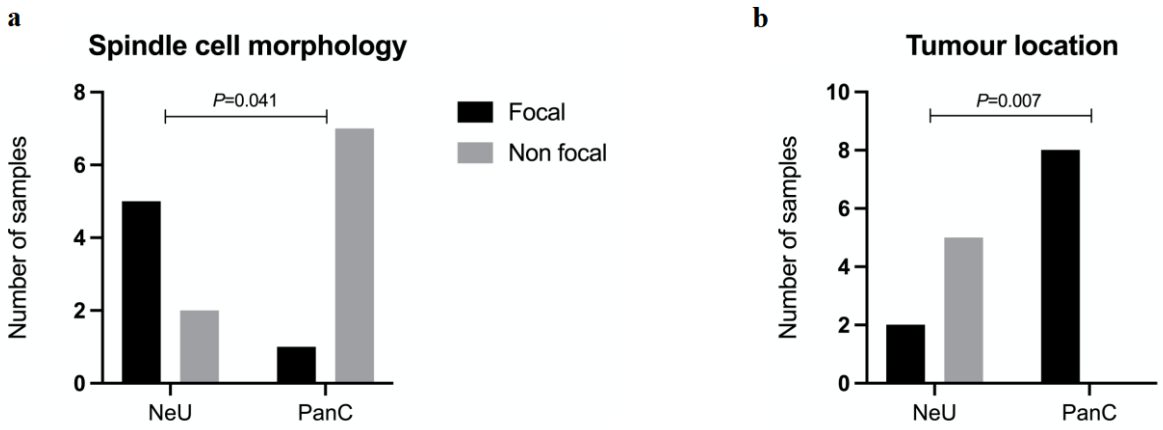


Figure 6. 4 | L-CD molecular subtypes have distinct histological characteristics. a) Focal spindle cell morphology in L-CD-NeU versus L-CD-PanC. b) Central or affecting the main lobar bronchi versus peripheral tumour location between L-CD-NeU and L-CD-PanC. *P*-values are for two-sided Fisher's statistical tests.

6.4 Mutational Signatures of L-CD Subtypes

Next, the molecular features associated with these two contrasting (in terms of expression) groups were investigated by analysing the spectrum of somatic base substitutions and their trinucleotide context. The mutational load was significantly higher in L-CD-NeU with 45.83 mutations on average compared with 27.87 for L-CD-PanC ($P = 5.53x10^{-4}$) (Fig. 6.5). This is consistent with a higher mean number of mutations detected in ACs and tumours from this histology all falling in the L-CD-NeU molecular group. Noteworthy, the number of mutations was also significantly higher in ACs as compared to TCs (two-sided *t*-test: t -2.41; estimate TC=31.63, estimate AC=45.66 [95% CI -26.71, -1.35]; df 12; *P*=0.03).

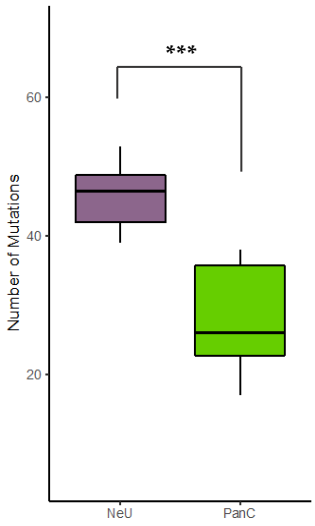


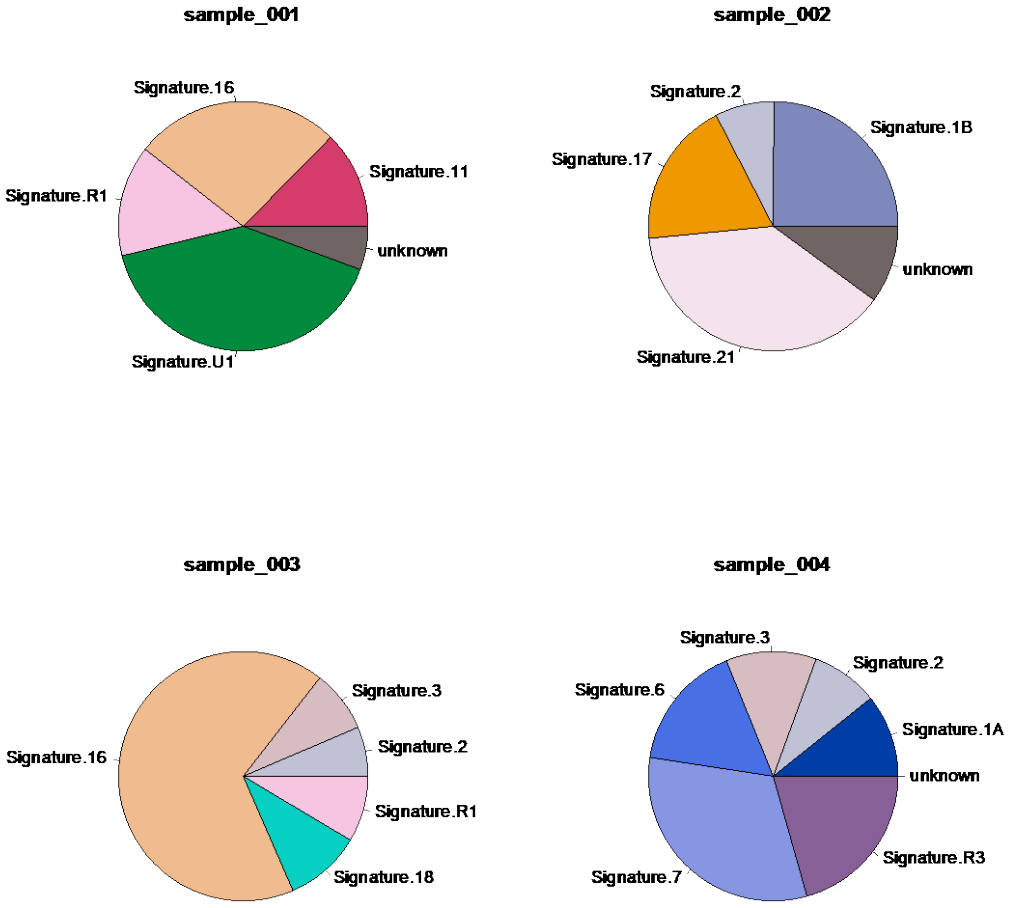
Figure 6. 5| Number of mutations in L-CD molecular subtypes was significantly higher in the L-CD-NeU subtype (two-sided *t*-test: $P = 5.53x10^{-4}$).

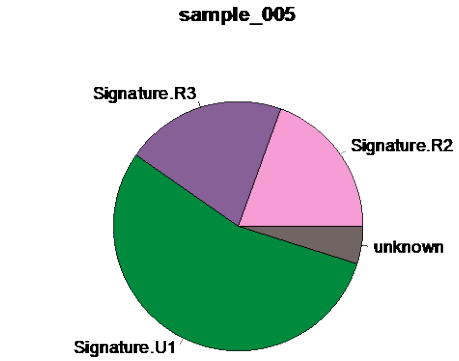
Considering COSMIC Mutational Signatures (CMS) (Chapter 3, Section 3.3.1) a different spectrum was consistently observed between the two L-CD subtypes with 50% of L-CD-PanC samples possessing an aflatoxin exposure signature. This is consistent with the differential gene expression data in which several genes (*GGT5*, *CYP34A*, *CYP3A5* and

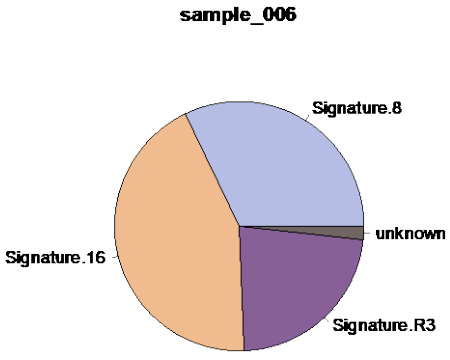
DPEP1) participating in aflatoxin activation and detoxification were expressed at significantly higher levels in PanC tumours.

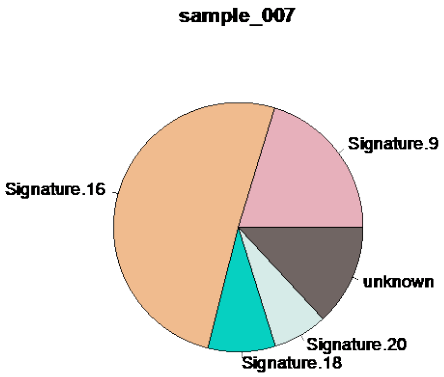
On the other hand, L-CD-NeU samples varied either showing a CMS 5 signature (57%), found in most cancer samples, and/or a CMS 8 signature (43%), which is associated with double strand break repair by homologous recombination. A summary of the mutational signatures is shown below in Figure 6.6.

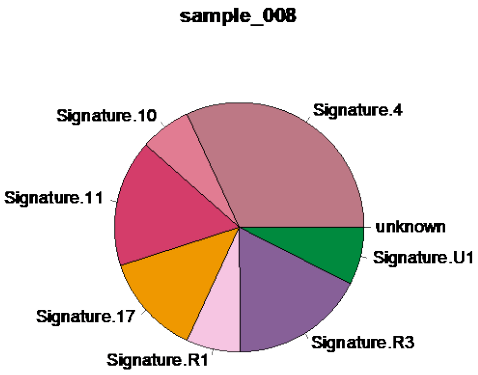
a) L-CD-PanC




sample_005
Signature.R3
Signature.R2
unknown
Signature.U1

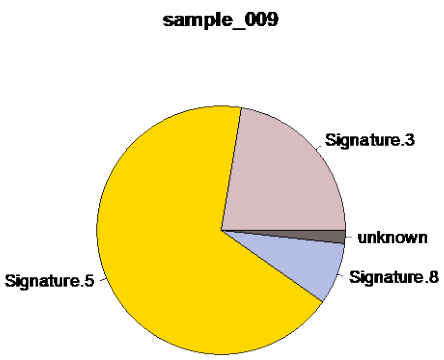

sample_006
Signature.8
unknown
Signature.16
Signature.R3

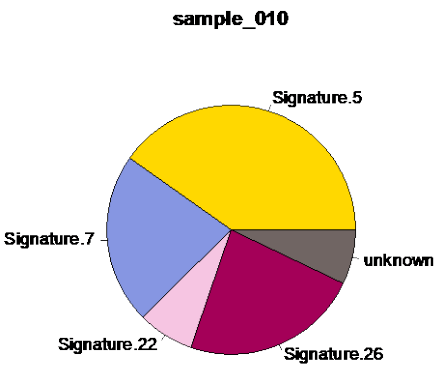

sample_007
Signature.9
Signature.16
unknown
Signature.20
Signature.18


sample_008
Signature.10
Signature.4
Signature.11
unknown
Signature.U1
Signature.17
Signature.R3
Signature.R1

---

b) L-CD-NeU


sample_009
Signature.3
unknown
Signature.5
Signature.8


sample_010
Signature.5
Signature.7
unknown
Signature.22
Signature.26

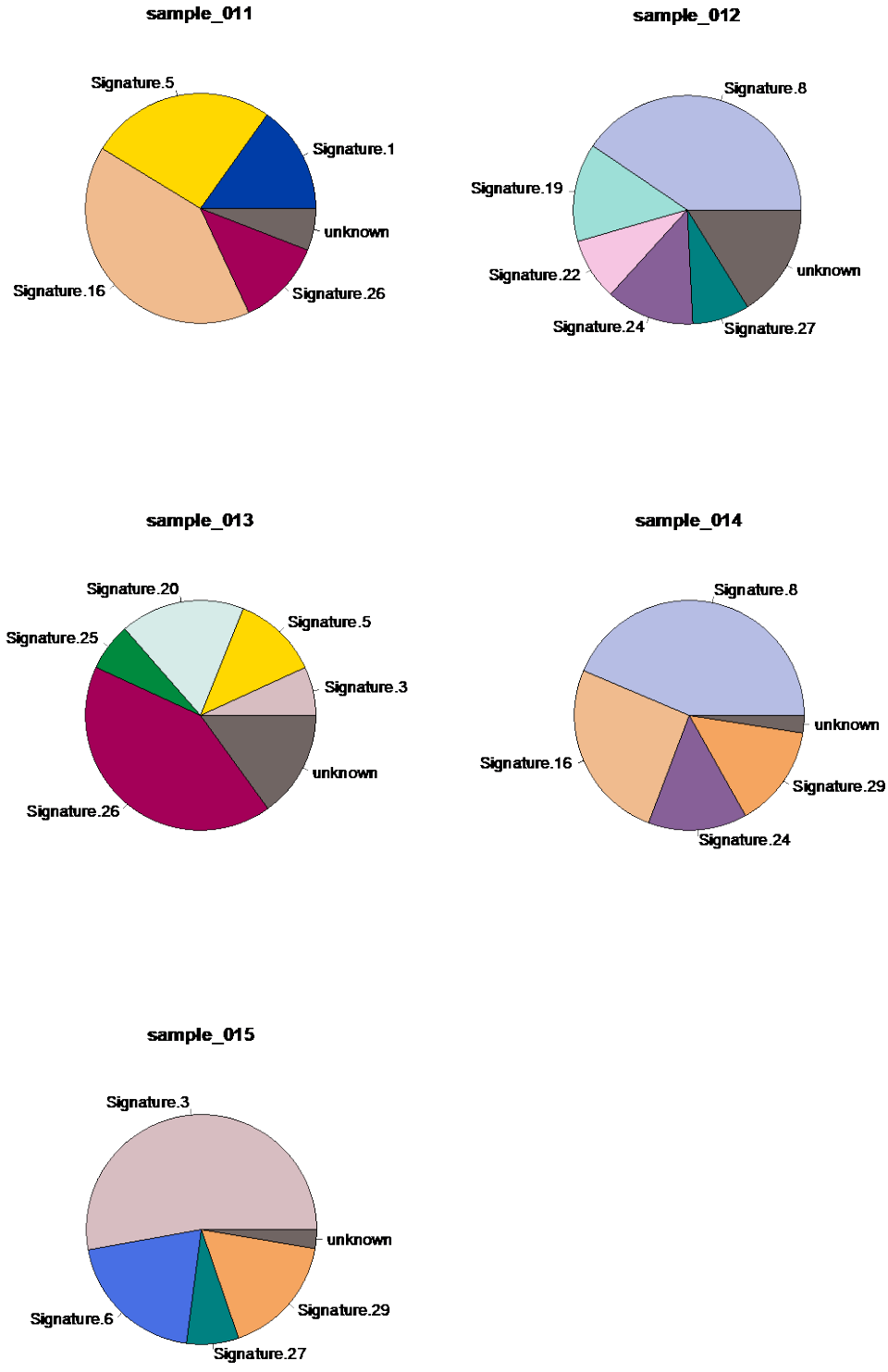


Figure 6. 6| Weights of each mutational signature operative in (a) L-CD-PanC and (b) L-CD-NeU tumours. Mutational Signatures identified with deconstructSigs in each molecular subtype with COSMIC mutational signatures version 2 by using the exome2genome normalization method.

Since most observed signatures were of unknown aetiology, *de novo* mutational signatures were then identified (Chapter 3, Section 3.3.2) and related to the catalogued (COSMIC) signatures[331]. Analysis with a cophenetic correlation metric confirmed a consistent presence of different spectra of mutational signatures in the two groups. L-CD-PanC showed an aflatoxin signature (CSM 24; cosine similarity of 0.603) while L-CD-NeU showed spontaneous deamination of 5-methylcytosine (CSM 1; cosine similarity of 0.801). Both subtypes were found to share a mutational signature associated with defective DNA mismatch repair (CSM 20; cosine similarities of 0.399 in PanC and 0.642 in NeU) (Fig. 6.7). These data confirmed a biological distinction between the two observed L-CD subtypes.

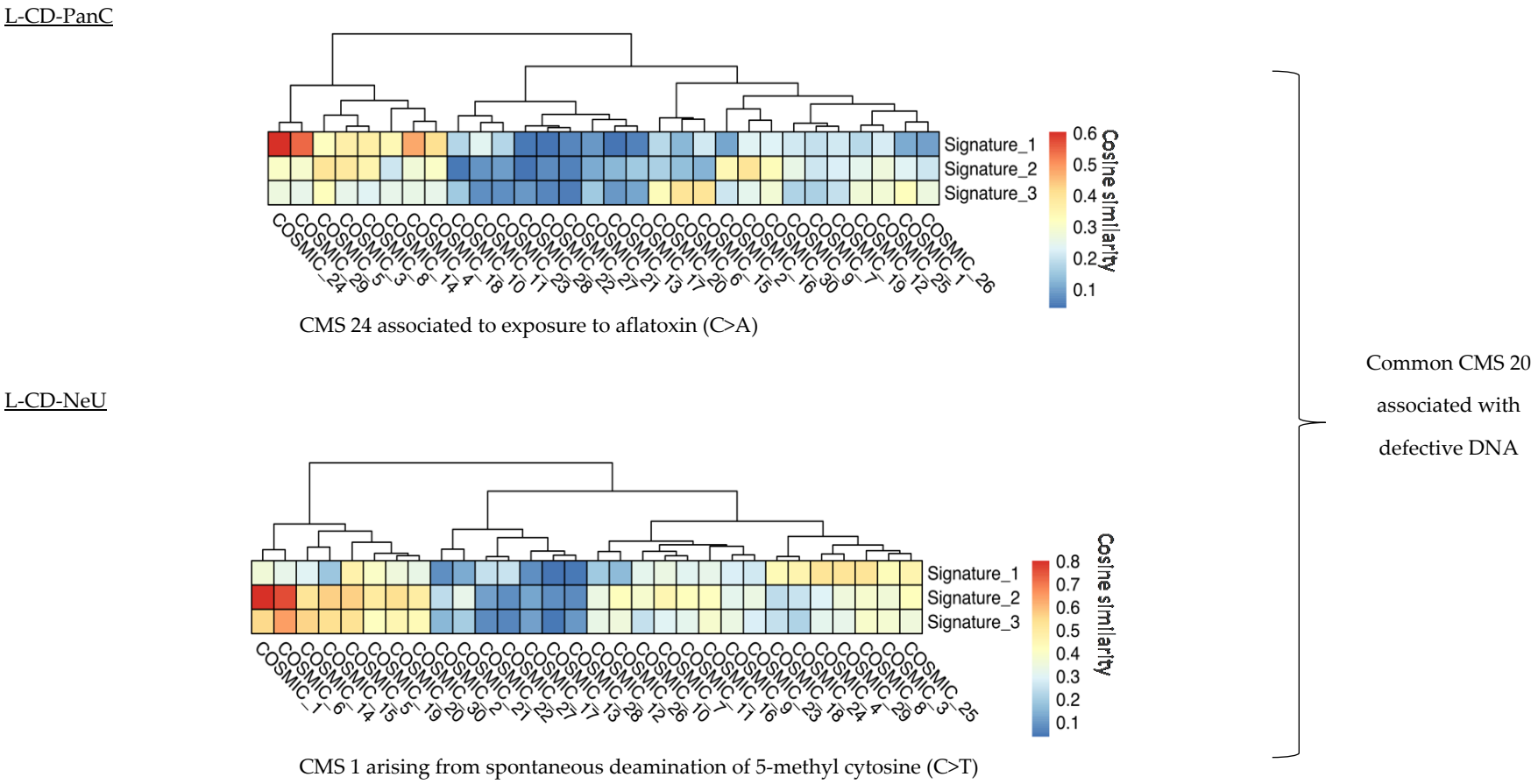


Figure 6. 7 | *De novo* Mutational signatures identified in L-CD subtypes using matrix factorization and compared to known mutagenic processes (COSMIC mutational signatures) identified by Alexandrov and colleagues[29]. Abbreviations: CMS, COSMIC Mutational Signature.

6.5 Mutations and Copy Number Alterations (CNAs) in L-CD Subtypes

Next differences between the two subtypes (L-CD-PanC versus L-CD-NeU) in relation to mutations and CNAs (Chapter 2, Sections 2.2 and 2.4) were investigated. The most frequently mutated genes in L-CD-NeU were components of the cytoskeleton (57%), including *ITGA1*, *ITGA2*, *PLEC* and *TPJ1*, and histone covalent modifiers (43%) (*JMJD1C*, *KDM1B* and *KDM4E*) (Fig. 6.8). By contrast, L-CD-PanC tumours showed mutations in *ARID1A* and *ARID5B* (25%), both members of the SWI/SNF complex; and the Notch signalling genes *PLXND1* and/or *WWC1* (25%). Notch signalling controls cell fate decisions and has previously been found to be activated in neuroendocrine cells undergoing reprogramming after injury[332]. No significant difference was observed in the rate of gene mutations between the two L-CD groups.

No significant difference in somatic copy number burden (CNB) could be detected between the two L-CD classifications. Just 4.8 and 5.5% of the genome showed evidence of CNV in L-CD-NeU and L-CD-PanC respectively (inter-sample range L-CD-NeU: 0.3% - 14.5%; L-CD-PanC: 0.3%-19.3%). These values lie within the range previously defined in healthy human populations[333].

Individual CNAs were found in genes related to the innate immune system and neutrophil degranulation, and included deletions affecting C1orf127 (1p36.22), *TXK* and *TEC* (4q12), *NDUFS2* (1q23.3), *KIF20B* (10q23), *INMT*, *ROCK1* and zinc finger proteins (*ZNF180*, *ZNF846*, *ZNF283*, *ZNF404*). One significant amplification was found in *SCYP2*, a major component of the synaptonemal complex the latter being key in meiotic division.

A summary of the genes harbouring mutations and CNAs is given in Figure 6.8, with information regarding genes in significant cytobands and the percentages for CNAs in each subtype detailed in Table 6.3.

Similar to mutations, L-CD-NeUs also showed higher frequencies of CNAs compared to the L-CD-PanC group, with 71.43% of tumours harbouring *KIF20B* deletions, a kinesin involved in neuron polarization[334]. In contrast with earlier studies[49], no significant loss in *RB1*, *TP53* and *MEN1* or any significant gains in *TERT*, *SDHA* or *RICTOR* were detected, however this may be a function of low frequency and a small sample size in this present study[49,62].

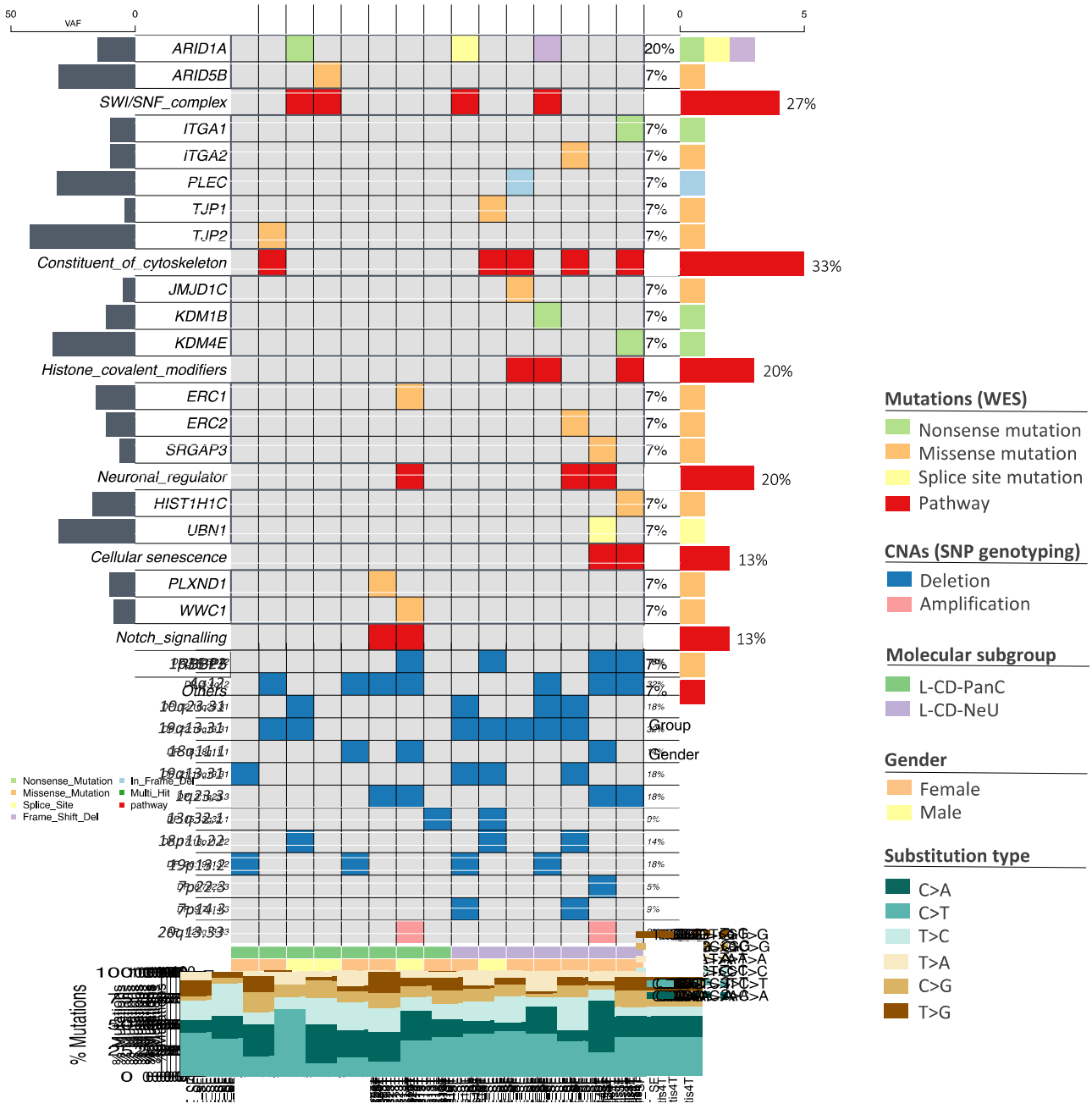


Figure 6. 8| Oncoplot of the most recurrent mutations, InDels and significant Copy Number Alterations (CNAs) in L-CD molecular subtypes. Columns represent each patient and in rows are listed the genetic alterations. Recurrent somatic mutations (coloured by type of mutation or InDel) and their related biological pathway, with the total frequency of mutations belonging to each pathway is shown in red. Left bar plot represents median Variant Allele Frequencies for each gene with somatic mutations. Cytobands with significant deletions (in blue) and amplifications (in red) appear ranked from top to bottom based on significance (q-value). Below CNAs are shown L-CD subgroup membership together with gender for each patient. On the bottom, summary of the types of substitutions as a stacked barplot showing the fraction of substitutions in each L-CD sample. Abbreviations: VAF (Variant Allele Frequency); WES (Whole-Exome Sequencing).

Cytoband	*Wide Peak Limits*	*Genes*	*q values*	*Residual q values*	*PanC (%)*	*NeU (%)*
1p36.22	chr1:11005496-11008843	*C1orf127*	4.28E-12	4.28E-12	12.50	42.86
4q12	chr4:48068784-52743948	*TXK, TEC*	2.01E-09	2.01E-09	50.00	42.86
1q23.3	chr1:161167699-161176136	*ADAMTS4, NDUFS2*	1.15E-07	1.15E-07	12.50	42.86
10q23.31	chr10:91497196-91511197	*KIF20B*	1.07E-05	1.07E-05	25.00	71.43
19q13.31	chr19:44978299-44988638	*ZNF180*	1.60E-05	0.00035745	25.00	14.29
7p14.3	chr7:30793498-30795447	*INMT*	0.00042248	0.0038469	12.50	42.86
18q11.1	chr18:14585295-18904233	*ROCK1, ANKRD30B*	0.00099715	0.00099715	25.00	28.57
20q13.33	chr20:58416429-58467096	*SCYP2*	0.0073662	0.0073662	12.50	14.29
19p13.2	chr19:9868450-9872906	*ZNF846*	0.0084492	0.0084492	12.50	28.57
19q13.31	chr19:44350719-44377800	*ZNF283, ZNF404*	0.0084492	0.22424	25.00	28.57
13q32.1	chr13:96496859-96546879	*UGGT2*	0.011732	0.011732	0.00	14.29
7p22.3	chr7:2748718-2752382	*AMZ1*	0.017476	0.12272	0.00	28.57
18p11.22	chr18:9254407-9256298	*ANKRD12*	0.017476	0.017476	12.50	14.29

Table 6. 3| Genes in significant cytobands identified with Gistic22 in L-CD molecular subtypes. Q-values for each called peak and the associated residual q-values after removing segments shared with higher peaks are shown. Percentage of samples harbouring each copy number alteration is given for each molecular subtype.

6.6 Focal and Widespread DNA Methylation Changes Distinguish L-CD Subtypes

Aberrant DNA methylation is a common feature of cancer and provides a mechanism of gene expression dysregulation. Whole Genome Bisulfite Sequencing (WGBS) was performed (Chapter 2, Section 2.6) to generate a snapshot of the DNA methylation profile in tumour and matched histologically normal tissue. Mirroring the observations from transcriptomic data (Fig. 6.1), PCA of genome-wide DNA methylation data differentiated L-CD-NeU from L-CD-PanC tumours (Fig. 6.9a).

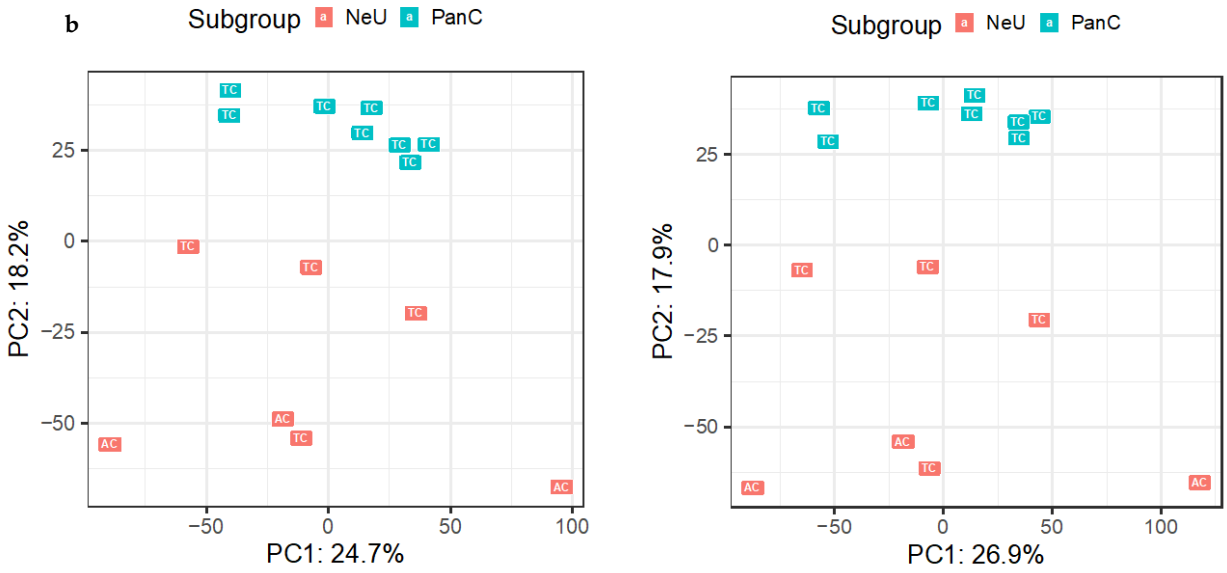


Figure 6. 9 | Principal components analysis of whole genome (a) and repeat element (b) DNA methylation data differentiates L-CD molecular subtypes. The figure shows principal components analysis of WGBS data, a) whole genome and b) repeat elements only. NeU tumours are shown in red, PanC tumours are shown in blue. Histology is shown as TC and AC for Typical Carcinoid and Atypical Carcinoid respectively.

A total number of 4,304 significant Differentially Methylated Regions (DMRs) were identified between L-CD groups (See Appendix, Supplementary Data 6.2 for full listing) and, consistent with gene expression data, promoters of genes expressed in beta cells and related to insulin secretion were identified to be hypomethylated in the L-CD-PanC group (Fig. 6.10). Other hypomethylated genes included *SMAD7*, *NRG1* and *PMS1*.

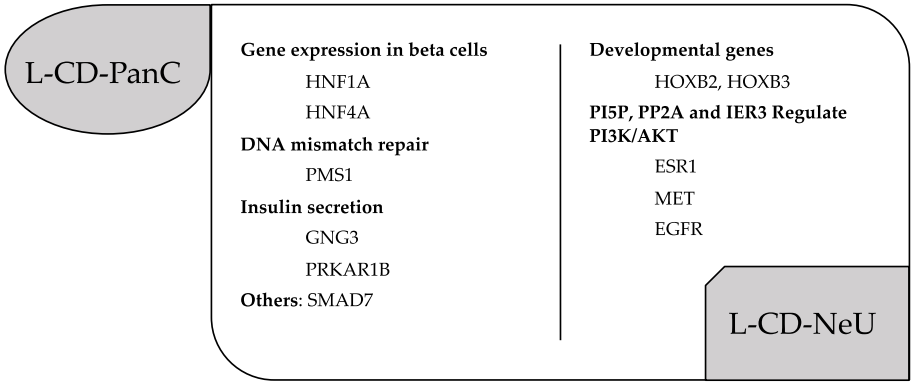


Figure 6. 10| Genes with promoters showing significant differential methylation of > 20% between L-CD subtypes. Statistical analysis was performed using a Wald test and P <1x10^{-6} was considered statistically significant.

On the other hand, L-CD-NeU tumours showed hypomethylation of the PI3K/AKT/mTOR signalling pathway, a pathway known to be involved in pluripotency and cell fate determination. *HOXB2* and *HOXB* developmental genes were also found to be hypomethylated. *HOX* genes are major transcriptional regulators with key roles in development, and frequent epigenetic and/or transcriptional deregulation in cancer [335,336]. Notably the majority of DMRs identified mapped to intergenic and intronic regions (Table 6.4).

Next the properties of DNA methylation alterations in the non-coding genome of the two L-CD groups was investigated. PCA analysis of different genomic regions' DNA methylation levels revealed that repeat elements explained most of the variance (26.89%) and alone distinguished L-CD-NeU from L-CD-PanC (Fig. 6.9b).

Annotation type	***DMRs (n)***
hg19_cpg_inter	3869
hg19_cpg_islands	60
hg19_cpg_shelves	194
hg19_cpg_shores	291
hg19_enhancers_fantom	174
hg19_genes_1to5kb	292
hg19_genes_3UTRs	89
hg19_genes_5UTRs	86
hg19_genes_exons	390
hg19_genes_introns	2265
hg19_genes_promoters	129
hg19_lncrna_gencode	497

Table 6. 4| Number of DMRs between L-CD-PanC and L-CD-NeU per annotation type identified with Annotatr R package.

Reactivation of transposable elements (TE) through epigenetic mechanisms is an established feature of some cancers with roles in tumour immunity[118]. In this present study, a significant enrichment of TEs in hypomethylated DMRs (relative enrichment = 1.31, 95% CI 1.16-1.48, $P = 8.03 \times 10^{-06}$) was found. Specifically, DMRs enriched in TEs with a 30% fraction

overlap were significantly lowly methylated (two-tailed Wilcoxon matched-pairs signed rank test, $P < 2.2 \times 10^{-16}$) in PanC tumours (Supplementary Fig. 6). Repeat elements and non-genic CpG sites are known to lose methylation during aging[337], nevertheless, no significant difference in age between L-CD subtypes was found in this present study.

In addition to *MEN1* mutations, alterations in *ATRX* or *DAXX* genes and activation of the Alternative Lengthening of Telomeres (ALT) pathway have previously been reported in Pancreatic Neuroendocrine Tumours (PanNETs)[338,339,340,341] and have been proposed as markers for sensitivity to ATR inhibitors (ATRi)[342]. Genetic alterations (by mutation or CNA) were not detected in *ATRX* and/or *DAXX* genes in the data set of this study. Similarly, whilst expression and methylation levels of these genes were concordant neither differed significantly between L-CD-PanC and L-CD-NeU tumours (Fig. 6.11).

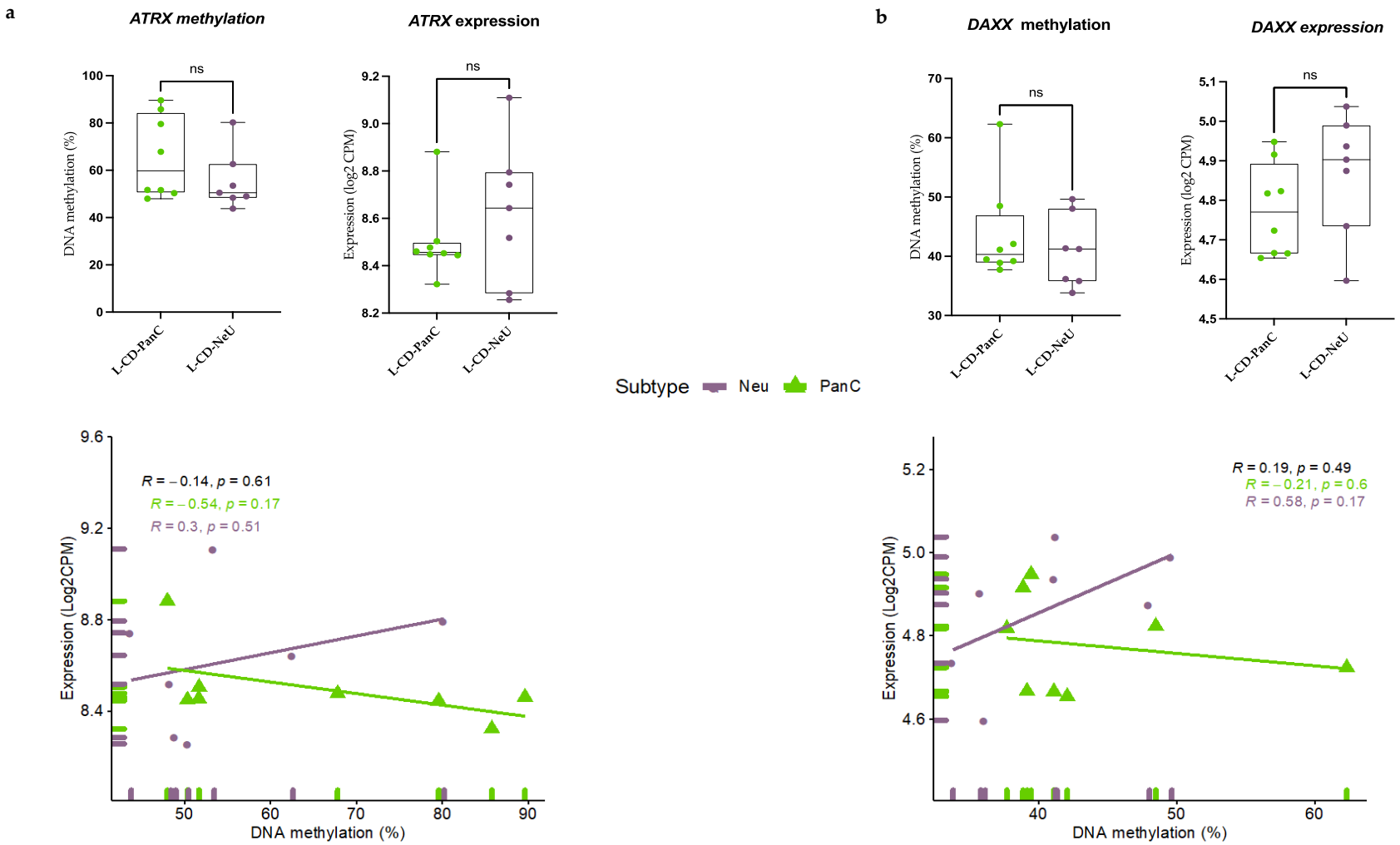


Figure 6. 11| Box plots of average DNA methylation percentage and expression levels for (a) *ATRX* and (b) *DAXX*, and their relationships. DNA methylation percentage was calculated for regions 2 Kb upstream of the first exon and the first exon itself. Spearman's correlation coefficients and associated *P*-values are shown in black for the whole L-CD data set (*n*=15), as well as the LCD-PanC (*n*=8) and L-CD-NeU (*n*=7) molecular subtypes.

6.7 Transposable Element (TE) Expression Analysis

Next the differential expression (DE) of TEs between the two L-CD groups were explored. From a total of 1,128 TEs detected, 29 were identified significantly as DE ($P < 0.01$). The Long Terminal Repeat (LTR) class of retrotransposons was found to be the most abundant (79.3%) among DE TEs. Interestingly, the majority of DE TEs (89%) were upregulated in the L-CD-NeU group as compared to the L-CD-PanC group, with the LTR retrotransposon *MER52* showing the highest expression ($\log_2$ FC 3.91, adj.$P = 1.7\text{x } 10^{-5}$) (Fig.6.12).

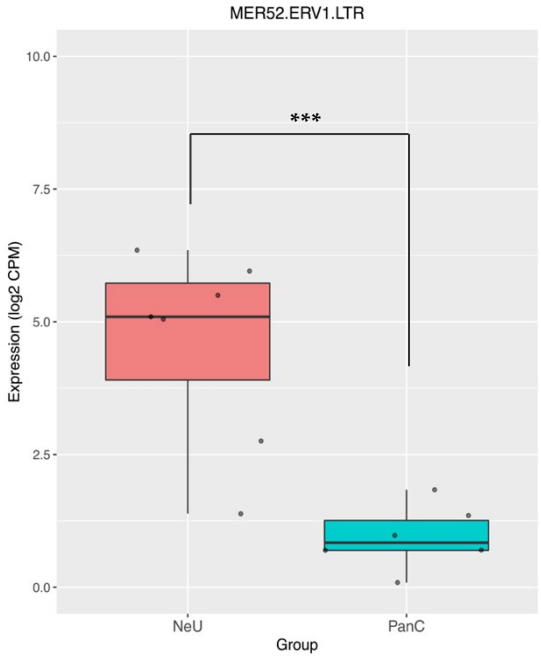


Figure 6. 12| Significant Differential Expression of the LTR *MER52* between L-CD subtypes. Normality was assessed with a Shapiro Wilk test. Differential expression was assessed using a Welch *t*-test ($P = 8.23 \text{ x } 10^{-4}$). Magnitude of significance is denoted with an asterisk (*).

MER52 is a TE from the Endogenous Retrovirus (ERV) superfamily within the LTR subclass of retrotransposons[343]. It is physically located in an intergenic region in the 4q13.2 cytoband (chr4:70021864-70023292 in hg19/Human). Proximal loss of DNA methylation has been found to result in reactivation of TEs in some cancers[118]. Therefore, for the present data set DNA methylation was assessed 1 Kb upstream of *MER52* and within the "gene body" of the Long Terminal Repeat (LTR). Interestingly, a positive correlation was observed in the L-CD-NeU group and the correlation between expression and DNA methylation was higher 1 Kb upstream of *MER52* (Fig. 6.13a-b), although neither of these correlations was significant.

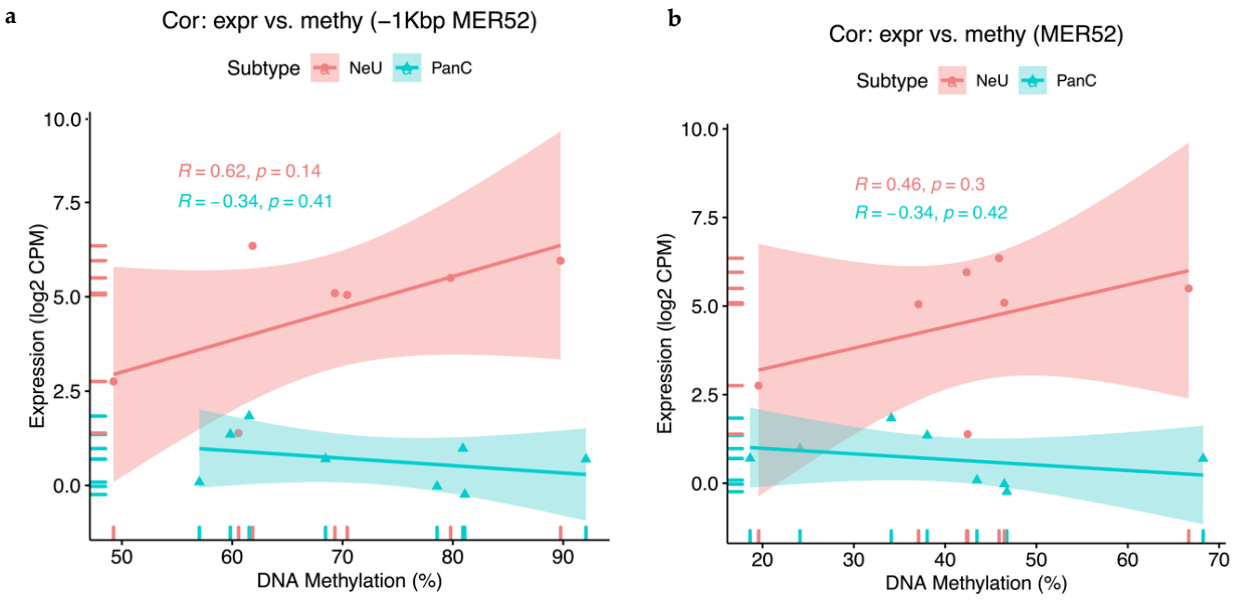


Figure 6. 13| *MER52* DNA methylation-expression relationship. *MER52* DNA methylation and gene expression were positively associated in the L-CD-NeU subtype, both 1 Kb upstream (R = 0.62, *P* = 0.14) and within the LTR sequence (R = 0.46, *P* = 0.3), although associations were not statistically significant. Correlation coefficients (R) are given for each genomic region and L-CD group, with test statistics based on Pearson's correlation coefficients cor(x, y).

TE sequences have been documented to regulate gene expression by acting as *cis*-regulatory elements[344,345] and are capable of controlling gene expression networks in a coordinated fashion[251]. Thus, a correlation matrix was computed to investigate genes whose expression co-varied with *MER52* expression. A total of 259 genes were found to be significantly correlated with *MER52* (adj.*P* < 0.05) and importantly 7 genes were amongst the top 20 genes DE (Table 6.2) between L-CD subtypes (Table 6.5).

Gene	*cor*	*P*	*adj.P*
A1CF	-0.884651743	1.19E-05	0.024431917
TM4SF5	-0.856803559	4.51E-05	0.039285805
GLYCTK	-0.854543163	4.97E-05	0.040579121
RDH12	-0.845541865	7.17E-05	0.045993567
DOK7	0.864826647	3.16E-05	0.034735741
FAM3B	0.887354754	1.02E-05	0.023153912
DPYSL3	0.913402472	1.98E-06	0.012433011

Table 6. 5| Genes that significantly correlated with *MER52* expression that were also identified as the top-most significantly DE between L-CD subtypes. Abbreviations: cor (Pearson correlation coefficients); adj.*P* (Benjamini-Hochberg FDR adjusted *P*-value).

Pathway enrichment analysis with genes showing a significant negative correlation identified IL-37 signalling ($P = 9.71 \times 10^{-4}$), protein localization ($P = 1.07 \times 10^{-3}$), peroxisomal protein import ($P = 1.33 \times 10^{-3}$), activation of gene expression by *SREBF* ($P = 2.14 \times 10^{-3}$) and *TP53* regulation of metabolic genes ($P = 5.4 \times 10^{-3}$) signalling pathways, amongst others. Genes related to cytosolic sulfonation of small molecules ($P = 9.9 \times 10^{-4}$) and ion homeostasis ($P = 8.51 \times 10^{-3}$) signalling pathways were enriched in genes that positively correlated with *MER52*.

6.8 Discussion

In this study, two distinct molecular groups of L-CDs were identified by analysis of RNA-sequencing data; L-CD-NeU and L-CD-PanC, and their molecular profiles were characterised using WES, SNP genotyping and WGBS. These two groups differed significantly in their transcriptional, mutational and epigenetic profiles, as well as their physical characteristics.

Differential analysis of gene expression data showed upregulation of metabolic pathways and hallmarks of pancreatic beta cells in PanC tumours, whereas pathways related to the neuronal system, neurotransmitter synthesis and release were found enriched in NeU tumours. Differential expression between these two groups was most marked with *TM4SF5*, notable since TM4SF5-targeted monoclonal antibody and peptide vaccination has previously established preventive and/or therapeutic effects in hepatocellular carcinoma, colon cancer

and pancreatic cancer models[322]. On the basis of these data, assessment of anti-hTM4SF5 antibody treatment efficacy in TM4SF5-expressing PanC L-CDs may be warranted.

The gene *A1CF* showed the highest $\log_2$ fold-change expression between the two L-CD groups suggesting that A1CF protein expression may have utility as a molecular marker for anti-hTM4SF5 antibody therapy. Furthermore, the NeU L-CD classification was significantly associated with a focal spindle cell morphology and peripheral tumour location, meaning that these physical features may provide a minimally invasive, accessible proxy, in agreement with the distinct characteristics of central and peripheral carcinoids reported by George *et al*.[346]. Additionally FAM3B/PANDER expression has been detected at the protein level and its inhibition has shown anti-tumour effects *in vitro* in prostate and several human cancer lines[326], suggesting therapeutic potential and a marker for the differential diagnosis of L-CD subtypes in combination with TM4SF5 and A1CF protein expression.

Solid tumours are commonly infiltrated by different types of immune cells. To compare the infiltration of immune cells between L-CD groups, we compared the expression levels of immune cell enriched genes as defined in the Human Protein Atlas, and identified significant differences for 25 transcripts ($adj.P \leq 0.05$). Thus, histological quantification of immune cell infiltrating cells in these tumours could be investigated to explore if L-CD-PanC tumours may be more likely to respond to checkpoint inhibitors.

Molecular profiling of L-CDs have previously shown chromatin remodelling genes, such as *MEN1*, *ARID1A*, *PSIP1*, *KMT2C* and *KMT2A*, to be recurrently mutated in L-CDs while *TP53*, *RB1* and *STK11* mutations have been found frequently altered in non-carcinoid NETs[347]. Other studies have emphasized the distinction between TCs and ACs, and molecular events distinguishing these subtypes are reported to affect the genes *MEN1, TP53*, *KMT2C, TERT*, *SDHA, RICTOR* and *RB1*[49]. In this present study, no mutation or CNA in any of these genes was detected although this may be a function of low frequency and a small sample size[49,62].

L-CD-NeU showed a higher tumour mutational load affecting cytoskeletal genes and histone covalent modifiers. Conversely, L-CD-PanC tumours showed a lower mutational load, mostly affecting members of the SWI/SNF complex and Notch signalling pathways. L-CD-NeU tumours also showed more recurrent CNAs than L-CD-PanCs. Although L-CD-NeUs encompassed all the ACs, most members of this group had typical histology

highlighting the potential importance of molecular screening to help in the therapy decision process.

To gain insights into the biological mechanisms involved in L-CD carcinogenesis, looking for the most frequent combinations of somatic mutations resulted in the identification of a combination of known and *de novo* mutational signatures. Known mutational signatures from the COSMIC database (https://cancer.sanger.ac.uk/signatures/) revealed a mixed repertoire of signatures that have been found in other cancer types. *De novo* signatures identified were compared with known catalogued COSMIC signatures. Both approaches linked an aflatoxin signature with L-CD-PanC, whereas signatures previously found in all cancer types were identified in the L-CD-NeU group.

Aflatoxin B1 (AFB1) is a potent genotoxin produced by *Aspergillus* fungus. It can bind to double stranded DNA[348] and induce hepatocellular carcinoma leaving a C→A mutational signature[349,350,351]. Pancreatic tumours have previously shown dominance of this signature potentially due to the mutational properties of AFB1-DNA adducts[352,353,82,354]. Extrahepatic tissues, such as the nasal olfactory and respiratory mucosa, and mucosa of the trachea and oesophagus have a high capacity to activate AFB1 which when inhaled may cause lung cancer[355,356]. In this present study a predominance of C→A mutations in the L-CD-PanC group was also observed (data not shown). High levels of AFB1 can be present in respiratory grain-dust particles[357,358] and as such could partially contribute to L-CD-PanC carcinogenesis.

The mutational landscapes of L-CDs alone could not explain the transcriptomic differences detected (Section 6.5 above) and therefore their DNA methylomes were investigated by analysis of WGBS data. DMRs in promoters of pancreatic beta cells and genes related to insulin secretion in L-CD-PanCs were identified, as well as mismatch repair genes pinpointing DNA methylation changes as a key event in this cancer group.

TE-enriched regions showed significant hypomethylation in L-CD-PanCs and alone were sufficient to differentiate from L-CD-Neu tumours by PCA analysis. This suggests that epigenetic dysregulation in the non-coding genome may be a major contributor to the PanC subtype. Significantly higher expression levels of *A1CF* in L-CD-PanC tumour genomes may contribute to this generalised hypomethylation. A1CF serves as a docking site to recruit APOBEC1 deaminase. The latter has a role in active DNA demethylation followed by T:G

mismatch repair[359,360]. Altogether these findings highlight the value of whole-genome data to better understand and refine the molecular alterations in these malignancies.

TEs of the LTR class were the most abundantly significantly DE between L-CD groups, in line with what has been observed in 13 TCGA cancer types[118]. L-CD-NeU tumours showed significant higher expression of TEs relative to L-CD-PanCs with the endogenous retrovirus, *MER52*, showing the highest log_2FC expression. *MER52* is a LTR retrotransposon of the Endogenous Retrovirus 1 family (ERV1). DNA methylation levels were then assessed within *MER52* LTR body sequence and its putative regulatory region 1 Kb upstream. In contrast to what other investigations have reported[361,362], a positive correlation between DNA methylation and TE expression was observed in the L-CD-NeU subtype (although not significant), with a higher Pearson correlation metric for the region 1 Kb upstream of *MER52*. This could be explained by the complex interplay between genetics and epigenetics, since TEs can be epigenetically silenced by DNA methylation and/or repressive histone modifications.

Furthermore, many TEs encode functional regulatory elements that can control gene regulatory networks. In this present study, a significant association between *MER52* expression and several genes found dysregulated between L-CD groups was observed. This suggests that TE expression could be associated with the disrupted gene expression programmes and phenotypic features of this group of tumours.

The present study does have several limitations notably that due to the relative rarity of L-CD tumours the study sample size is fairly small. The study would benefit by follow up of the findings by conducting histopathological analyses for the markers identified. This has the potential to not only validate the observed carcinoid classifications but also enable their translation into a clinical setting.

Chapter 7: Final Remarks

The overall aim of the work presented within this thesis was the integration of different sets of omics data to better understand the molecular processes contributing to lung cancer. In this project, it has been possible to integrate molecular with clinical data, enabling the detection of commonalities and differences at the genetic and epigenetic level as well as the pathway level, between different lung cancer histological and molecular groups.

In Chapter 3, recurrent somatic mutations and InDels were detected in known and novel genes in LUAD, LUSC and LNET tumours, whilst L-CDs showed a low number of somatic mutations, potentially due to their indolent behaviour and differentiation status (Section 3.4). Nevertheless, analysis of copy number data and its integration with mutational data revealed a relative high number of deletions and amplifications in L-CDs as compared to the other LC histotypes (Chapter 3, Section 3.5.6). Moreover, functional annotation of the genetically altered genes allowed the identification of perturbations in ten hallmark oncogenic pathways in nearly all LC histotypes except for L-CDs (Chapter 3, Section 3.6.3). The latter histotype showed alterations in six out of the ten most common altered pathways in cancer.

Next in Chapter 4, epigenetic alterations were investigated through the analysis of DNA methylation data at the whole-genome level in a subset of LC tumour and normal pairs. Unsupervised clustering analysis revealed that DNA methylation profiles clearly differentiated between LC histotypes, namely L-CDs and NSCLCs, as well as between tumour and normal matched tissue. A shared characteristic of L-CD and NSCLC histotypes was global DNA methylation loss (Chapter 4, Section 4.7.1). This genome-wide hypomethylation signature has been associated with the aberrant activation of repetitive elements and endogenous retroviruses and with the ectopic activation of non-lineage-appropriate enhancers, among others.

These distinct DNA methylation landscapes were subsequently explored by the annotation of differentially methylated regions into genic and CpG annotations, thus disentangling the putative cis-regulatory mechanisms altered in tumours relative to normal samples, as well as between L-CD and NSCLC histotypes (Chapter 4, Section 4.7.2). After intergenic regions, intronic regions and lncRNAs were the annotations that harboured the highest number of DMRs (Chapter 4, Section 4.7.2). This however was not necessarily surprising considering only ~2% of the human genome is composed of protein-coding regions.

Nevertheless, a striking difference in the number of DMRs between L-CDs and NSCLCs was found, with DMRs in L-CDs at intergenic regions being double in number compared to those detected in NSCLCs (Chapter 4, Figure 4.9). Worthy of note is that DNA methylation at the promoter level revealed not such a marked difference between LC histotypes (Chapter 4, Section 4.9). This is in line with the well documented hyper- and hypomethylation at CpG islands associated with gene promoters. These observations, together with a considerable amount of copy number alterations detected in L-CDs, suggest that L-CDs have a more unstable genome compared to NSCLCs, and may be an indication of different molecular mechanisms being involved in L-CD carcinogenesis. In other words, NSCLCs harbour more genetic and epigenetic aberrations centred at the (protein-coding) gene level, whilst L-CD molecular alterations are located outside of genes. The fact that non-coding regions have hardly been prioritized in research studies could explain the lack of genetic drivers and molecular biomarkers in this rarer LC histotype. Not only have non-coding alterations not been extensively explored but also the small number of L-CD cases seen at single centres make this cancer hard to investigate. Multinational multi genomic studies are therefore needed.

Despite that it is highly likely that the majority of DNA methylation changes in cancer may not have a direct functional effect, this study has revealed new potential mechanisms driving altered genetic programmes. Although analysis of non-coding regions was outside of the scope of this study, the potential of whole-genome DNA methylation data allowed some meaningful and novel insights on their dysregulation across LC types to be determined. Significant hypomethylation was detected at transposable elements in both L-CDs and NSCLCs (Chapter 4, Section 4.10.4), whilst hypermethylation at TEs was not prominent. Transposable elements are conducive to genome instability and can lead to abnormal cellular differentiation and organismal development[363].

The *cis*-regulatory activity of enhancers was also explored and was found to be related to developmental and differentiation programmes in both L-CDs and NSCLCs (Chapter 4, Section 4.10.3). Although some differences in DNA methylation between normal tissues and specific cancer types appear to be tissue-specific, these observations are consistent with significant similarities seen across cancer types[364,365]. DNA methylation changes occur during normal differentiation and are inherent to the cellular lineage and differentiation stage. Thus further studies should focus on trying to distinguish which changes are cancer specific. For example, ensuring that both the cancer and the reference samples are of high cell purity, as

well as by the inclusion of other layers of genome-wide data sets such as histone modifications, epigenetic states, genomic accessibility and three-dimensional chromatin structure.

Employing the knowledge acquired in Chapters 3 and 4, two sets of LC samples were investigated through the integration of different types of omics data together with clinical data (Chapter 5). As such, this investigation focused first on a group of NSCLC tumours that showed reduced Y gene expression. This group of patients featured a network of co-expressed genes related to a reduction of male-specific gene expression, together with recurrent *TP53* mutations and copy number alterations in DNA damage repair genes (Chapter 5, Section 5.3.5). Moreover, consistent with the central loss of *KDM5D* expression, this group of males also showed a DNA methylation loss signature together with APOBEC-related mutations (Chapter 5, Section 5.3.6). APOBEC deaminases target DNA and RNA substrates that can lead to both genetic and epigenetic changes. This hints a potential link between environmental exposures and the genomic instability within a more permissive environment detected in these male NSCLC tumours.

Finally, Chapter 6 focused on L-CDs because of their distinct genetic and epigenetic landscapes as compared the other LC types. Integration of genetic, epigenetic and gene expression data with clinical parameters in this final study uncovered the potential of omics-data integration for molecular subtyping. Two distinct L-CD groups were identified from transcriptomic (Chapter 6, Section 6.2) and DNA methylation data (Chapter 6, Section 6.6), with each group individually featuring pancreatic and neuronal gene-related pathways. Some of the differentially expressed genes were detected altered at the DNA methylation level in promoters revealing DNA methylation as a key mechanism leading to distinct gene expression programmes in L-CDs (Chapter 6, Section 6.6 and Table 6.10). Moreover, a different spectrum of mutational signatures and copy number alterations were identified (Section 6.5), together with significant associations with histological parameters (Chapter 6, Section 6.3) and therapeutic stratification.

It is becoming increasingly clear that gene expression is determined by a complex interplay among different genetic and epigenetic layers as well as external exposures. Further research should therefore aim to integrate different data sets from large cohorts of samples and applying recently developed single cell techniques and digital histopathology in order to

achieve a deeper understanding of DNA methylation in cancer. Here, within this thesis, the integration of genomics and epigenomics has allowed a more comprehensive understanding of the expression programmes associated with current LC histological classifications, and in the molecular groups identified in Chapters 5 and 6. These results thus emphasize the importance of integrating different data types to identify molecular groups relevant in the clinical setting.

For instance, the findings of this work could likely be translatable and would likely improve the detection, monitoring and stratification for targeted therapies in LC patients.

Bibliography

1. Willis-Owen, S. A. G. *et al.* Y disruption, autosomal hypomethylation and poor male lung cancer survival. *Sci. Rep.* **11**, 12453 (2021).
2. Hoang, L. T. *et al.* Metabolomic, transcriptomic and genetic integrative analysis reveals important roles of adenosine diphosphate in haemostasis and platelet activation in non-small-cell lung cancer. *Mol. Oncol.* **13**, 2406–2421 (2019).
3. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA. Cancer J. Clin.* **71**, 209–249 (2021).
4. Tumours, W. C. of. *Thoracic Tumours: WHO Classification of Tumours*. (2021).
5. Travis, W. D. *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.* **10**, 1243–1260 (2015).
6. Travis, W. D., Brambilla, E. & Riely, G. J. New Pathologic Classification of Lung Cancer: Relevance for Clinical Practice and Clinical Trials. *J. Clin. Oncol.* **31**, 992–1001 (2013).
7. Islami, F. *et al.* Proportion and number of cancer cases and deaths attributable to potentially modifiable risk factors in the United States. *CA. Cancer J. Clin.* **68**, 31–54 (2018).
8. Thun, M., Peto, R., Boreham, J. & Lopez, A. D. Stages of the cigarette epidemic on entering its second century. *Tob. Control* **21**, 96–101 (2012).
9. Lortet-Tieulent, J. *et al.* Convergence of decreasing male and increasing female incidence rates in major tobacco-related cancers in Europe in 1988-2010. *Eur. J. Cancer* **51**, 1144–1163 (2015).
10. Jemal, A. *et al.* Higher Lung Cancer Incidence in Young Women Than Young Men in the United States. *N. Engl. J. Med.* **378**, 1999–2009 (2018).
11. Alavanja, M. *et al.* IARC Monographs on the Evaluation of Carcinogenic Risks to Humans. Tobacco Smoke. *IARC Monogr. Eval. Carcinog. Risks to Humans* **83**, 1–1413 (2004).
12. Hung, R. J. *et al.* A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* **452**, 633–637 (2008).
13. Amos, C. I. *et al.* Genome-wide association scan of tag SNPs identifies a susceptibility

locus for lung cancer at 15q25. 1. *Nat. Genet.* **40**, 616–622 (2008).

14. Jenkins, R. A. *et al.* Exposure to environmental tobacco smoke in sixteen cities in the United States as determined by personal breathing zone air sampling. *J. Expo. Anal. Environ. Epidemiol.* **6**, 473–502 (1996).
15. Stanaway, J. D. *et al.* Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks for 195 countries and territories, 1990-2017: A systematic analysis for the Global Burden of Disease Stu. *Lancet* **392**, 1923–1994 (2018).
16. Mu, L. *et al.* Indoor air pollution and risk of lung cancer among Chinese female non-smokers. *Cancer Causes Control* **24**, 439–450 (2013).
17. Turner, M. C. *et al.* Outdoor air pollution and cancer: An overview of the current evidence and public health recommendations. *CA. Cancer J. Clin.* (2020) doi:10.3322/caac.21632.
18. Budisan, L. *et al.* Links between Infections, Lung Cancer, and the Immune System. *International Journal of Molecular Sciences* vol. 22 (2021).
19. Mao, Q. *et al.* Interplay between the lung microbiome and lung cancer. *Cancer Lett.* **415**, 40–48 (2018).
20. Akhtar, N. & Bansal, J. G. Risk factors of Lung Cancer in nonsmoker. *Curr. Probl. Cancer* **41**, 328–339 (2017).
21. Doria-Rose, V. P. & Szabo, E. Screening and prevention of lung cancer. *Lung cancer a Multidiscip. approach to diagnosis Manag.* **2**, (2010).
22. Mehta, S. R. et al. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. *N. Engl. J. Med.* **365**, 687–696 (2015).
23. de Koning, H. J. *et al.* Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. *N. Engl. J. Med.* **382**, 503–513 (2020).
24. Chen, Z., Fillmore, C. M., Hammerman, P. S., Kim, C. F. & Wong, K.-K. Non-small-cell lung cancers: a heterogeneous set of diseases. *Nat. Rev. Cancer* **15**, 247–247 (2015).
25. Scagliotti, G. V. *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.* **26**, 3543–3551 (2008).
26. Screening, T. *et al.* Activating Mutations in the Epidermal Growth Factor Receptor

Underlying Responsiveness of Non–Small-Cell Lung Cancer to Gefitinib. *N. Engl. J. Med.* **350**, 1757–1765 (2014).

27. Soda, M. *et al.* Identification of the transforming EML4–ALK fusion gene in non-small-cell lung cancer. *Nature* **448**, 561 (2007).
28. Shaw, A. T. *et al.* Crizotinib in ROS1-Rearranged Non–Small-Cell Lung Cancer. *N. Engl. J. Med.* **371**, 1963–1971 (2014).
29. Moro-Sibilot, D. *et al.* Crizotinib in c-MET- or ROS1-positive NSCLC: results of the AcSé phase II trial. *Ann. Oncol.* **30**, 1985–1991 (2019).
30. Wakelee, H. Evaluating the Role of Targeted Therapy in Lung Cancer. *Oncology (Williston Park, N.Y.)* vol. 33 (2019).
31. Malhotra, J., Jabbour, S. K. & Aisner, J. Current state of immunotherapy for non-small cell lung cancer. *Transl. Lung Cancer Res.* **6**, 196–211 (2007).
32. Society, A. C. *Cancer facts & figures 2015*. (American Cancer Society, 2015).
33. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2019. *CA. Cancer J. Clin.* **69**, 7–34 (2019).
34. Martincorena, I. *et al.* High burden and pervasive positive selection of somatic mutations in normal human skin. *Science (80-.).* **348**, 880–886 (2015).
35. Loh, P. R. *et al.* Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* **559**, 350–355 (2018).
36. Martincorena, I. *et al.* Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
37. Neelakantan, D., Drasin, D. J. & Ford, H. L. Intratumoral heterogeneity: Clonal cooperation in epithelial-to-mesenchymal transition and metastasis. *Cell Adh. Migr.* **9**, 265–276 (2015).
38. Chalmers, Z. R. *et al.* Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med.* **9**, 1–14 (2017).
39. Kan, Z. *et al.* Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature* **466**, 869–873 (2010).
40. Rizvi, N. A. *et al.* Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* **348**, 124–128 (2015).
41. Devarakonda, S., Morgensztern, D. & Govindan, R. Genomic alterations in lung

adenocarcinoma. *Lancet Oncol.* **16**, e342–e351 (2015).

42. Tonon, G. *et al.* High-resolution genomic profiles of human lung cancer. *Proc. Natl. Acad. Sci.* **102**, 9625–9630 (2005).
43. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).
44. Rossi, G., Bertero, L., Marchiò, C. & Papotti, M. Molecular alterations of neuroendocrine tumours of the lung. *Histopathology* **72**, 142–152 (2018).
45. Di Domenico, A., Wiedmer, T., Marinoni, I. & Perren, A. Genetic and epigenetic drivers of neuroendocrine tumours (NET). *Endocr. Relat. Cancer* **24**, R315–R334 (2017).
46. Zhou, Z. *et al.* Comparison of genomic landscapes of large cell neuroendocrine carcinoma, small cell lung carcinoma, and large cell carcinoma. *Thorac. cancer* **10**, 839–847 (2019).
47. George, J. *et al.* Integrative genomic profiling of large-cell neuroendocrine carcinomas reveals distinct subtypes of high-grade neuroendocrine lung tumors. *Nat. Commun.* **9**, 1048 (2018).
48. George, J. *et al.* Comprehensive genomic profiles of small cell lung cancer. *Nature* **524**, 47–53 (2015).
49. Simbolo, M. *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.* **241**, 488–500 (2017).
50. Wakelee, H. A. *et al.* Lung Cancer Incidence in Never Smokers. *J. Clin. Oncol.* **25**, 472–478 (2007).
51. Sun, S., Schiller, J. H. & Gazdar, A. F. Lung cancer in never smokers — a different disease. *Nat. Rev. Cancer* **7**, 778–790 (2007).
52. Shi, Y. X., Sheng, D. Q., Cheng, L. & Song, X. Y. Current Landscape of Epigenetics in Lung Cancer: Focus on the Mechanism and Application. *J. Oncol.* **2019**, (2019).
53. Lee, K.-S. *et al.* Runx3 is required for the differentiation of lung epithelial cells and suppression of lung cancer. *Oncogene* **29**, 3349–3361 (2010).
54. Pulling, L. C. *et al.* Promoter hypermethylation of the O6-methylguanine-DNA methyltransferase gene: More common in lung adenocarcinomas from never-smokers than smokers and associated with tumor progression. *Cancer Res.* **63**, 4842–4848 (2003).

55. Dubois, F. *et al.* RASSF1A Suppresses the Invasion and Metastatic Potential of Human Non-Small Cell Lung Cancer Cells by Inhibiting YAP Activation through the GEF-H1/RhoB Pathway. *Cancer Res.* **76**, 1627–1640 (2016).

56. Tang, X. *et al.* Hypermethylation of the Death-Associated Protein (DAP) Kinase Promoter and Aggressiveness in Stage I Non-Small-Cell Lung Cancer. **92**, (2017).

57. Shi, Y.-X. *et al.* Genome-wide DNA methylation profiling reveals novel epigenetic signatures in squamous cell lung cancer. *BMC Genomics* **18**, 901 (2017).

58. Kalari, S., Jung, M., Kernstine, K. H., Takahashi, T. & Pfeifer, G. P. The DNA methylation landscape of small cell lung cancer suggests a differentiation defect of neuroendocrine cells. *Oncogene* **32**, 3559–3568 (2013).

59. Poirier, J. T. *et al.* DNA methylation in small cell lung cancer defines distinct disease subtypes and correlates with high expression of EZH2. *Oncogene* **34**, 5869–5878 (2015).

60. Pelosi, G. *et al.* Dual role of RASSF1 as a tumor suppressor and an oncogene in neuroendocrine tumors of the lung. *Anticancer Res.* **30**, 4269–4281 (2010).

61. Malpeli, G. *et al.* Methylation Dynamics of RASSF1A and Its Impact on Cancer. *Cancers* vol. 11 (2019).

62. Laddha, S. V. *et al.* Integrative genomic characterization identifies molecular subtypes of lung carcinoids. *Cancer Res.* **79**, 4339–4347 (2019).

63. Alcala, N. *et al.* Integrative and comparative genomic analyses identify clinically relevant pulmonary carcinoid groups and unveil the supra-carcinoids. *Nat. Commun.* **10**, (2019).

64. Epigenetics, Energy Balance, and Cancer. in *Energy Balance and Cancer* (ed. Berger, N. A.) 167–189 (Springer US, 2016). doi:10.1007/978-3-319-41610-6_7.

65. Haber, D. A. & Settleman, J. Drivers and passengers. *Nature* **446**, 145–146 (2007).

66. Bashyam, M. D., Animireddy, S., Bala, P., Naz, A. & George, S. A. The Yin and Yang of cancer genes. *Gene* **704**, 121–133 (2019).

67. Reeves, M. E., Firek, M., Chen, S.-T. & Amaar, Y. The RASSF1 Gene and the Opposing Effects of the RASSF1A and RASSF1C Isoforms on Cell Proliferation and Apoptosis. *Mol. Biol. Int.* **2013**, 145096 (2013).

68. Estrabaud, E. *et al.* RASSF1C, an isoform of the tumor suppressor RASSF1A, promotes the accumulation of beta-catenin by interacting with betaTrCP. *Cancer Res.* **67**, 1054–

1061 (2007).

69. Reeves, M. E., Firek, M., Chen, S. T. & Amaar, Y. G. Evidence that RASSF1C stimulation of lung cancer cell proliferation depends on IGFBP-5 and PIWIL1 expression levels. *PLoS One* **9**, 3–12 (2014).

70. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).

71. Negrini, S., Gorgoulis, V. G. & Halazonetis, T. D. Genomic instability--an evolving hallmark of cancer. *Nat. Rev. Mol. Cell Biol.* **11**, 220–228 (2010).

72. Colotta, F., Allavena, P., Sica, A., Garlanda, C. & Mantovani, A. Cancer-related inflammation, the seventh hallmark of cancer: links to genetic instability. *Carcinogenesis* **30**, 1073–1081 (2009).

73. Sansone, S.-A. *et al.* Toward interoperable bioscience data. *Nat. Genet.* **44**, 121–126 (2012).

74. Livingston, K. M., Bada, M., Baumgartner, W. A. J. & Hunter, L. E. KaBOB: ontology-based semantic integration of biomedical databases. *BMC Bioinformatics* **16**, 126 (2015).

75. Jeggo, P. A., Pearl, L. H. & Carr, A. M. DNA repair, genome stability and cancer: A historical perspective. *Nat. Rev. Cancer* **16**, 35–42 (2016).

76. Loeb, L. A. A mutator phenotype in cancer. *Cancer Res.* **61**, 3230–3239 (2001).

77. Kunkel, T. A. DNA-mismatch repair. The intricacies of eukaryotic spell-checking. *Curr. Biol.* **5**, 1091–1094 (1995).

78. Genschel, J., Littman, S. J., Drummond, J. T. & Modrich, P. Isolation of MutSbeta from human cells and comparison of the mismatch repair specificities of MutSbeta and MutSalpha. *J. Biol. Chem.* **273**, 19895–19901 (1998).

79. Thompson, S. L., Bakhoum, S. F. & Compton, D. A. Mechanisms of chromosomal instability. *Curr. Biol.* **20**, R285–R295 (2010).

80. Jackson, S. P. & Bartek, J. The DNA-damage response in human biology and disease. *Nature* **461**, 1071–1078 (2009).

81. Gonzalez-Perez, A., Sabarinathan, R. & Lopez-Bigas, N. Local Determinants of the Mutational Landscape of the Human Genome. *Cell* **177**, 101–114 (2019).

82. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).

83. Singer, B. & Hang, B. Nucleic acid sequence and repair: role of adduct, neighbor bases and enzyme specificity. *Carcinogenesis* **21**, 1071–1078 (2000).

84. Alexandrov, L. B. *et al.* The Repertoire of Mutational Signatures in Human Cancer. *bioRxiv* (2018) doi:10.1101/322859.

85. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).

86. D'Arcy, M. S. Cell death: a review of the major forms of apoptosis, necrosis and autophagy. *Cell Biol. Int.* **43**, 582–592 (2019).

87. Seoane, J. & Gomis, R. R. TGF-β Family Signaling in Tumor Suppression and Cancer Progression. *Cold Spring Harb. Perspect. Biol.* **9**, (2017).

88. Zhao, M., Mishra, L. & Deng, C.-X. The role of TGF-β/SMAD4 signaling in cancer. *Int. J. Biol. Sci.* **14**, 111–123 (2018).

89. Hayden, M. S. & Ghosh, S. Shared principles in NF-κB signaling. *Cell* **132**, 344–362 (2008).

90. Vallabhapurapu, S. & Karin, M. Regulation and function of NF-κB transcription factors in the immune system. *Annu. Rev. Immunol.* **27**, 693–733 (2009).

91. Challa, S. *et al.* Targeting the IκB Kinase Enhancer and Its Feedback Circuit in Pancreatic Cancer. *Transl. Oncol.* **13**, 481–489 (2020).

92. Katoh, M. & Katoh, M. WNT signaling pathway and stem cell signaling network. *Clin. Cancer Res.* **13**, 4042–4045 (2007).

93. Ramakrishnan, A. B., Chen, L., Burby, P. E. & Cadigan, K. M. Wnt target enhancer regulation by a CDX/TCF transcription factor collective and a novel DNA motif. *Nucleic Acids Res.* **49**, 8625–8641 (2021).

94. Dogan, S. *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. an Off. J. Am. Assoc. Cancer Res.* **18**, 6169–6177 (2012).

95. Porta, C., Paglino, C. & Mosca, A. Targeting PI3K/Akt/mTOR signaling in cancer. *Front. Oncol.* **4**, 64 (2014).

96. Piccolo, S., Dupont, S. & Cordenonsi, M. The biology of YAP/TAZ: hippo signaling and beyond. *Physiol. Rev.* (2014).

97. Dhanasekaran, R. *et al.* The MYC oncogene — the grand orchestrator of cancer growth

and immune evasion. *Nat. Rev. Clin. Oncol.* (2021) doi:10.1038/s41571-021-00549-2.

98. Shang, S., Hua, F. & Hu, Z.-W. The regulation of β-catenin activity and function in cancer: therapeutic opportunities. *Oncotarget* **8**, 33972–33989 (2017).
99. Mumm, J. S. & Kopan, R. Notch signaling: from the outside in. *Dev. Biol.* **228**, 151–165 (2000).
100. Borggrefe, T. & Oswald, F. The Notch signaling pathway: Transcriptional regulation at Notch target genes. *Cell. Mol. Life Sci.* **66**, 1631–1646 (2009).
101. Kansanen, E., Kuosmanen, S. M., Leinonen, H. & Levonen, A.-L. The Keap1-Nrf2 pathway: Mechanisms of activation and dysregulation in cancer. *Redox Biol.* **1**, 45–49 (2013).
102. DeNicola, G. M. *et al.* Oncogene-induced Nrf2 transcription promotes ROS detoxification and tumorigenesis. *Nature* **475**, 106–109 (2011).
103. Rojo de la Vega, M., Chapman, E. & Zhang, D. D. NRF2 and the Hallmarks of Cancer. *Cancer Cell* **34**, 21–43 (2018).
104. Esteller, M. Epigenetic gene silencing in cancer: the DNA hypermethylome. *Hum. Mol. Genet.* **16 Spec No**, R50-9 (2007).
105. Stadler, M. B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**, 490–495 (2011).
106. Feldmann, A. *et al.* Transcription factor occupancy can mediate active turnover of DNA methylation at regulatory regions. *PLoS Genet.* **9**, e1003994 (2013).
107. Fraga, M. F. *et al.* Epigenetic differences arise during the lifetime of monozygotic twins. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 10604 LP – 10609 (2005).
108. Shoemaker, R., Deng, J., Wang, W. & Zhang, K. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Res.* **20**, 883–889 (2010).
109. Kerkel, K. *et al.* Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. *Nat. Genet.* **40**, 904–908 (2008).
110. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–329 (2015).
111. Bhutani, N., Burns, D. M. & Blau, H. M. DNA demethylation dynamics. *Cell* **146**, 866–872 (2011).
112. Jones, P. L. *et al.* Methylated DNA and MeCP2 recruit histone deacetylase to repress

transcription . *Nat. Genet.* **19**, 187–191 (1998).

113. Becker, P. B. & Workman, J. L. Nucleosome remodeling and epigenetics. *Cold Spring Harb. Perspect. Biol.* **5**, a017905 (2013).

114. Bernstein, B. E. *et al.* Methylation of histone H3 Lys 4 in coding regions of active genes. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 8695–8700 (2002).

115. Santos-Rosa, H. *et al.* Active genes are tri-methylated at K4 of histone H3. *Nature* **419**, 407–411 (2002).

116. Guil, S. & Esteller, M. Cis-acting noncoding RNAs: friends and foes. *Nat. Struct. Mol. Biol.* **19**, 1068–1075 (2012).

117. McClintock, B. The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci.* **36**, 344 LP – 355 (1950).

118. Kong, Y. *et al.* Transposable element expression in tumors is associated with immune infiltration and increased antigenicity. *Nat. Commun.* **10**, (2019).

119. Sharma, S., Kelly, T. K. & Jones, P. A. Epigenetics in cancer. *Carcinogenesis* **31**, 27–36 (2009).

120. Peter, A. J. & Baylin, S. B. The Epigenomic of Cancer. *Cell* **128**, 683–692 (2007).

121. Andrew P. Feinberg & Bert Vogelstein. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature* **301**,.

122. Gama-Sosa, M. A. *et al.* The 5-methylcytosine content of DNA from human tumors. *Nucleic Acids Res.* **11**, 6883–6894 (1983).

123. Ehrlich, M. DNA methylation and cancer-associated genetic instability. *Adv. Exp. Med. Biol.* **570**, 363–392 (2005).

124. Kulis, M. *et al.* Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nat. Genet.* **44**, 1236–1242 (2012).

125. Ziller, M. J. *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**, 477–481 (2013).

126. Bergman, Y. & Cedar, H. DNA methylation dynamics in health and disease. *Nat. Struct. Mol. Biol.* **20**, 274–281 (2013).

127. Cao, R. *et al.* Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science* **298**, 1039–1043 (2002).

128. Kimura, K. *et al.* Diversification of transcriptional modulation: Large-scale

identification and characterization of putative alternative promoters of human genes. *Genome Res.* **16**, 55–65 (2006).

129. Keshet, I. *et al.* Evidence for an instructive mechanism of de novo methylation in cancer cells. *Nat. Genet.* **38**, 149–153 (2006).
130. Martín-Subero, J. I. *et al.* New insights into the biology and origin of mature aggressive B-cell lymphomas by combined epigenomic, genomic, and transcriptional profiling. *Blood* **113**, 2488–2497 (2009).
131. Ohm, J. E. *et al.* A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nat. Genet.* **39**, 237–242 (2007).
132. Schlesinger, Y. *et al.* Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. *Nat. Genet.* **39**, 232–236 (2007).
133. Widschwendter, M. *et al.* Epigenetic stem cell signature in cancer. *Nat. Genet.* **39**, 157–158 (2007).
134. Brown, S. J., Stoilov, P. & Xing, Y. Chromatin and epigenetic regulation of pre-mRNA processing. *Hum. Mol. Genet.* **21**, R90-6 (2012).
135. Jjingo, D., Conley, A. B., Yi, S. V, Lunyak, V. V & Jordan, I. K. On the presence and role of human gene-body DNA methylation. *Oncotarget* **3**, 462–474 (2012).
136. Hansen, K. D. *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nat. Genet.* **43**, 768–775 (2011).
137. Hon, G. C. *et al.* Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome Res.* **22**, 246–258 (2012).
138. Baylin, S. B. & Ohm, J. E. Epigenetic gene silencing in cancer - a mechanism for early oncogenic pathway addiction? *Nat. Rev. Cancer* **6**, 107–116 (2006).
139. You, J. S. & Jones, P. A. Cancer genetics and epigenetics: two sides of the same coin? *Cancer Cell* **22**, 9–20 (2012).
140. Baylin, S. B. & Jones, P. A. A decade of exploring the cancer epigenome — biological and translational implications. *Nat. Rev. Cancer 11, 726-734* (2011).
141. Feinberg, A. P. & Irizarry, R. A. Evolution in health and medicine Sackler colloquium: Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proc. Natl. Acad. Sci. U. S. A.* **107 Suppl**, 1757–1764 (2010).
142. Gutierrez-Arcelus, M. *et al.* Passive and active DNA methylation and the interplay with

genetic variation in gene regulation. *Elife* **2**, e00523 (2013).

143. Kishimoto, M. *et al.* Mutations and deletions of the CBP gene in human lung cancer. *Clin. cancer Res. an Off. J. Am. Assoc. Cancer Res.* **11**, 512–519 (2005).

144. Rekhtman, N. *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.* **22**, 3618–3629 (2016).

145. Fernandez-Cuesta, L. & McKay, J. D. Genomic architecture of lung cancers. *Curr. Opin. Oncol.* **28**, 52–57 (2016).

146. Peifer, M. *et al.* Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. *Nat. Genet.* **44**, 1104–1110 (2012).

147. Pleasance, E. D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184–190 (2010).

148. Rudin, C. M. *et al.* Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. *Nat. Genet.* **44**, 1111–1116 (2012).

149. Shen, H. *et al.* Alteration in Mir-21/PTEN Expression Modulates Gefitinib Resistance in Non-Small Cell Lung Cancer. *PLoS One* **9**, e103305 (2014).

150. Flores-Pérez, J. A., De La, F., Oliva, R., Argenes, Y. & Meneses-Garcia, A. *Translational Research and Onco-Omics Applications in the Era of Cancer Personal Genomics*. *Advances in Experimental Medicine and Biology* vol. 1168 (2019).

151. Lyko, F. & Brown, R. DNA methyltransferase inhibitors and the development of epigenetic cancer therapies. *J. Natl. Cancer Inst.* **97**, 1498–1506 (2005).

152. GNYSZKA, A., JASTRZĘBSKI, Z. & FLIS, S. DNA Methyltransferase Inhibitors and Their Emerging Role in Epigenetic Therapy of Cancer. *Anticancer Res.* **33**, 2989 LP – 2996 (2013).

153. Oh, Y.-K. & Park, T. G. siRNA delivery systems for cancer treatment. *Adv. Drug Deliv. Rev.* **61**, 850–862 (2009).

154. Hoang, L. T. *et al.* Metabolomic, transcriptomic and genetic integrative analysis reveals important roles of adenosine diphosphate in haemostasis and platelet activation in non-small-cell lung cancer. *Mol. Oncol.* **13**, 2406–2421 (2019).

155. Zhao, S. *et al.* Strategies for processing and quality control of Illumina genotyping arrays. *Brief. Bioinform.* 1–11 (2017) doi:10.1093/bib/bbx012.

156. Wickham, H. *et al.* Welcome to the Tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).

157. Diskin, S. J. *et al.* Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res.* **36**, 1–12 (2008).

158. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci.* **107**, 16910–16915 (2010).

159. Seshan, V. & Olshen, A. DNAcopy: A Package for Analyzing DNA Copy Data. in (2015).

160. Church, D. M. *et al.* Modernizing reference genome assemblies. *PLoS Biol.* **9**, e1001091 (2011).

161. Genome Reference Consortium. https://www.ncbi.nlm.nih.gov/grc/human.

162. Strachan T, Goodship J, C. P. *Genetics and genomics in medicine*. (Taylor & Francis).

163. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).

164. Wang, S., Zhang, J., He, Z., Wu, K. & Liu, X.-S. The predictive power of tumor mutational burden in lung cancer immunotherapy response is influenced by patients' sex. *Int. J. cancer* **145**, 2840–2849 (2019).

165. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41–R41 (2011).

166. Swanton, C. & Govindan, R. Clinical Implications of Genomic Discoveries in Lung Cancer. *N. Engl. J. Med.* **374**, 1864–1873 (2016).

167. Berger, A. H. *et al.* High-throughput Phenotyping of Lung Cancer Somatic Mutations. *Cancer Cell* **30**, 214–228 (2016).

168. Campbell, J. D. *et al.* Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat. Genet.* **48**, 607–616 (2016).

169. SureSelectXT Target Enrichment System for the Illumina Platform Protocol. 1–102 https://www.agilent.com/cs/library/usermanuals/Public/G7530-90000.pdf (2021).

170. Mayakonda, A., Lin, D., Assenov, Y., Plass, C. & Koeffler, H. P. Maftools : efficient and comprehensive analysis of somatic variants in cancer. 1–10 (2018) doi:10.1101/gr.239244.118.

171. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids*

Res. **47**, D941–D947 (2019).

172. Catalogue of Somatic Mutations in Cancer. cancer.sanger.ac.uk.
173. Bourgey, M. *et al.* GenPipes: An open-source framework for distributed and scalable genomic analyses. *Gigascience* **8**, 1–11 (2019).
174. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
175. Liu, Y., Siegmund, K. D., Laird, P. W. & Berman, B. P. Bis-SNP: Combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biol.* **13**, R61 (2012).
176. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).
177. Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).
178. Feng, H., Conneely, K. N. & Wu, H. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Res.* **42**, e69 (2014).
179. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**, 1165–1188 (2001).
180. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).
181. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
182. Alexandrov, L. B. Understanding the origins of human cancer. *Science (80-.).* **350**, 1175–1177 (2015).
183. Patel, K. *et al.* Whole-Exome Sequencing Analysis of Oral Squamous Cell Carcinoma Delineated by Tobacco Usage Habits. *Front. Oncol.* **11**, 660696 (2021).
184. Pansare, K. *et al.* Establishment and genomic characterization of gingivobuccal carcinoma cell lines with smokeless tobacco associated genetic alterations and oncogenic PIK3CA mutation. *Sci. Rep.* **9**, 8272 (2019).
185. Harris, R. S., Petersen-Mahrt, S. K. & Neuberger, M. S. RNA editing enzyme APOBEC1 and some of its homologs can act as DNA mutators. *Mol. Cell* **10**, 1247–1253 (2002).

186. Di Noia, J. M. & Neuberger, M. S. Molecular mechanisms of antibody somatic hypermutation. *Annu. Rev. Biochem.* **76**, 1–22 (2007).

187. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).

188. Taylor, B. J. *et al.* DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. *Elife* **2**, e00534 (2013).

189. Cervantes-Gracia, K., Gramalla-Schmitz, A., Weischedel, J. & Chahwan, R. APOBECs orchestrate genomic and epigenomic editing across health and disease. *Trends Genet.* **37**, 1028–1043 (2021).

190. Letouzé, E. *et al.* Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nat. Commun.* **8**, (2017).

191. Haupt, S. *et al.* Identification of cancer sex-disparity in the functional integrity of p53 and its X chromosome network. *Nat. Commun.* **10**, 5385 (2019).

192. Wood, R. D., Mitchell, M., Sgouros, J. & Lindahl, T. Human DNA repair genes. *Science (80-.).* **291**, 1284–1289 (2001).

193. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).

194. Dewhurst, S. M. *et al.* Tolerance of Whole-Genome Doubling Propagates Chromosomal Instability and Accelerates Cancer Genome Evolution. *Cancer Discov.* **4**, 175 LP – 185 (2014).

195. Kuznetsova, A. Y. *et al.* Chromosomal instability, tolerance of mitotic errors and multidrug resistance are promoted by tetraploidization in human cells. *Cell cycle* **14**, 2810–2820 (2015).

196. Dong, J. *et al.* c-Myc regulates self-renewal in bronchoalveolar stem cells. *PLoS One* **6**, e23707 (2011).

197. Motooka, Y. *et al.* Pathobiology of Notch2 in lung cancer. *Pathology* **49**, 486–493 (2017).

198. Bazellières, E., Aksenova, V., Barthélémy-Requin, M., Massey-Harroche, D. & Le Bivic, A. Role of the Crumbs proteins in ciliogenesis, cell migration and actin organization. *Semin. Cell Dev. Biol.* **81**, 13–20 (2018).

199. Liu, P. *et al.* Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing. *Carcinogenesis* **33**, 1270–1276 (2012).

200. Centonze, G. *et al.* Beyond Traditional Morphological Characterization of Lung Neuroendocrine Neoplasms: In Silico Study of Next-Generation Sequencing Mutations Analysis across the Four World Health Organization Defined Groups. *Cancers* vol. 12 (2020).

201. Volkova, N. V. *et al.* Mutational signatures are jointly shaped by DNA damage and repair. *Nat. Commun.* **11**, (2020).

202. Shrivastav, M., De Haro, L. P. & Nickoloff, J. A. Regulation of DNA double-strand break repair pathway choice. *Cell Res.* **18**, 134–147 (2008).

203. Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).

204. Cooper GM. *The Development and Causes of Cancer*. *The Cell: A Molecular Approach* (2000).

205. Di Domenico, A., Wiedmer, T., Marinoni, I. & Perren, A. Genetic and epigenetic drivers of neuroendocrine tumours (NET). *Endocr. Relat. Cancer* **24**, R315–R334.

206. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

207. Brass, N. *et al.* Amplification of the genes BCHE and SLC2A2 in 40% of squamous cell carcinoma of the lung. *Cancer Res.* **57**, 2290–2294 (1997).

208. Zakut, H., Lapidot-Lifson, Y., Beeri, R., Ballin, A. & Soreq, H. In vivo gene amplification in non-cancerous cells: cholinesterase genes and oncogenes amplify in thrombocytopenia associated with lupus erythematosus. *Mutat. Res.* **276**, 275–284 (1992).

209. Zakut, H. *et al.* Acetylcholinesterase and butyrylcholinesterase genes coamplify in primary ovarian carcinomas. *J. Clin. Invest.* **86**, 900–908 (1990).

210. Heyn, H. *et al.* DNA methylation contributes to natural human variation. *Genome Res.* **23**, 1363–1372 (2013).

211. Waddington, C. H. The Epigenotype. *Int. J. Epidemiol.* **41**, 10–13 (2012).

212. Baubec, T. & Schübeler, D. Genomic patterns and context specific interpretation of DNA methylation. *Curr. Opin. Genet. Dev.* **25**, 85–92 (2014).

213. Jaenisch, R. & Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.* **33 Suppl**, 245–254 (2003).

214. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).

215. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).

216. Bell, J. T. *et al.* DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.* **12**, R10 (2011).

217. Timp, W. & Feinberg, A. P. Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. *Nat. Publ. Gr.* **13**, 497–510 (2013).

218. Madakashira, B. P. & Sadler, K. C. DNA Methylation, Nuclear Organization, and Cancer. *Front. Genet.* **8**, 76 (2017).

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

220. Holmes, E. E. *et al.* Performance evaluation of kits for bisulfite-conversion of DNA from tissues, cell lines, FFPE tissues, aspirates, lavages, effusions, plasma, serum, and urine. *PLoS One* **9**, e93933 (2014).

221. Wen, J., Fu, J., Zhang, W. & Guo, M. Genetic and epigenetic changes in lung carcinoma and their clinical implications. *Mod. Pathol.* **24**, 932–943 (2011).

222. Abaffy, T. Human Olfactory Receptors Expression and Their Role in Non-Olfactory Tissues Ã¢Â©Â© A Mini-Review. *J. Pharmacogenomics Pharmacoproteomics* **06**, 1–7 (2015).

223. Flegel, C., Manteniotis, S., Osthold, S., Hatt, H. & Gisselmann, G. Expression profile of ectopic olfactory receptors determined by deep sequencing. *PLoS One* **8**, e55368 (2013).

224. Pluznick, J. L. Gut microbiota in renal physiology: focus on short-chain fatty acids and their receptors. *Kidney Int.* **90**, 1191–1198 (2016).

225. Munakata, Y. *et al.* Olfactory receptors are expressed in pancreatic β-cells and promote glucose-stimulated insulin secretion. *Sci. Rep.* **8**, 1499 (2018).

226. Neuhaus, E. M. *et al.* Activation of an olfactory receptor inhibits proliferation of prostate cancer cells. *J. Biol. Chem.* **284**, 16218–16225 (2009).

227. Giandomenico, V. *et al.* Olfactory receptor 51E1 as a novel target for diagnosis in somatostatin receptor-negative lung carcinoids. *J. Mol. Endocrinol.* **51**, 277–286 (2013).

228. Ramos-Lopez, O. *et al.* Associations between olfactory pathway gene methylation marks, obesity features and dietary intakes. *Genes Nutr.* **14**, 11 (2019).

229. Ranzani, M. *et al.* Revisiting olfactory receptors as putative drivers of cancer. *Wellcome open Res.* **2**, 9 (2017).

230. Szklarczyk, D. *et al.* STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).

231. Jeong, B. Y. *et al.* Rab25 augments cancer cell invasiveness through a β1 integrin/EGFR/VEGF-A/Snail signaling axis and expression of fascin. *Exp. Mol. Med.* **50**, e435–e435 (2018).

232. Zhang, L. *et al.* Rab25-Mediated EGFR Recycling Causes Tumor Acquired Radioresistance. *iScience* **23**, 100997 (2020).

233. Wang, J. *et al.* Rab25 promotes erlotinib resistance by activating the β1 integrin/AKT/β-catenin pathway in NSCLC. *Cell Prolif.* **52**, e12592 (2019).

234. Suh, Y. & Lee, C. Genome-wide association study for genetic variants related with maximal voluntary ventilation reveals two novel genomic signals associated with lung function. *Medicine (Baltimore).* **96**, e8530 (2017).

235. Zhao, E. *et al.* Identification of a Six-lncRNA Signature With Prognostic Value for Breast Cancer Patients. *Front. Genet.* **11**, 673 (2020).

236. Zhang, X., Hu, Y., Gong, C. & Zhang, C. Overexpression of miR-518b in non-small cell lung cancer serves as a biomarker and facilitates tumor cell proliferation, migration and invasion. *Oncol. Lett.* **20**, 1213–1220 (2020).

237. Kucuksayan, H. *et al.* TGF-β-SMAD-miR-520e axis regulates NSCLC metastasis through a TGFBR2-mediated negative-feedback loop. *Carcinogenesis* **40**, 695–705 (2019).

238. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, pl1 (2013).

239. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).

240. Liu, L. *et al.* TERT Promoter Hypermethylation in Gastrointestinal Cancer: A Potential Stool Biomarker. *Oncologist* **22**, 1178–1188 (2017).

241. Barthel, F. P. *et al.* Systematic analysis of telomere length and somatic alterations in 31 cancer types. *Nat. Genet.* **49**, 349–357 (2017).

242. Lee, D. D. *et al.* DNA hypermethylation within TERT promoter upregulates TERT expression in cancer. *J. Clin. Invest.* **129**, 223–229 (2019).

243. Cui, S. *et al.* Prediction of MiR-21-5p in Promoting the Development of Lung

Adenocarcinoma via PDZD2 Regulation. *Med. Sci. Monit. Int. Med. J. Exp. Clin. Res.* **26**, e923366 (2020).

244. Li, H.-P. *et al.* Inactivation of the tight junction gene CLDN11 by aberrant hypermethylation modulates tubulins polymerization and promotes cell migration in nasopharyngeal carcinoma. *J. Exp. Clin. Cancer Res.* **37**, 102 (2018).

245. Aspenström, P. Activated Rho GTPases in Cancer-The Beginning of a New Paradigm. *Int. J. Mol. Sci.* **19**, 3949 (2018).

246. 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.* **5**, 2892–2911 (2015).

247. D'Angelo, S. P. *et al.* Incidence of EGFR exon 19 deletions and L858R in tumor specimens from men and cigarette smokers with lung adenocarcinomas. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **29**, 2066–2070 (2011).

248. Chen, Y.-F. *et al.* Motor protein-dependent membrane trafficking of KCl cotransporter-4 is important for cancer cell invasion. *Cancer Res.* **69**, 8585–8593 (2009).

249. Shen, M.-R. *et al.* KCl cotransport is an important modulator of human cervical cancer growth and invasion. *J. Biol. Chem.* **278**, 39941–39950 (2003).

250. Hsu, Y.-M. *et al.* IGF-1 upregulates electroneutral K-Cl cotransporter KCC3 and KCC4 which are differentially required for breast cancer cell proliferation and invasiveness. *J. Cell. Physiol.* **210**, 626–636 (2007).

251. Feschotte, C. The contribution of transposable elements to the evolution of regulatory networks. *Nat. Rev. Genet.* **9**, 397–405 (2008).

252. Kulis, M., Queirós, A. C., Beekman, R. & Martín-Subero, J. I. Intragenic DNA methylation in transcriptional regulation, normal differentiation and cancer. *Biochim. Biophys. Acta - Gene Regul. Mech.* **1829**, 1161–1174 (2013).

253. Montero, A. J. *et al.* Epigenetic inactivation of EGFR by CpG island hypermethylation in cancer. *Cancer Biol. Ther.* **5**, 1494–1501 (2006).

254. Gery, S. *et al.* Epigenetic silencing of the candidate tumor suppressor gene Per1 in non-small cell lung cancer. *Clin. cancer Res. an Off. J. Am. Assoc. Cancer Res.* **13**, 1399–1404 (2007).

255. Riscuta, G. *et al. Cancer Epigenetics for Precision Medicine*. *Methods in Molecular Biology*

vol. 1856 (2018).

256. Aguilera, O., Fernández, A. F., Muñoz, A. & Fraga, M. F. Epigenetics and environment: a complex relationship. *J. Appl. Physiol.* **109**, 243–251 (2010).

257. Stankovic, B. *et al.* Immune Cell Composition in Human Non-small Cell Lung Cancer. *Front. Immunol.* **9**, 3101 (2018).

258. Singmann, P. *et al.* Characterization of whole-genome autosomal differences of DNA methylation between men and women. *Epigenetics Chromatin* **8**, 43 (2015).

259. Yuan, Y. *et al.* Comprehensive Characterization of Molecular Differences in Cancer between Male and Female Patients. *Cancer Cell* **29**, 711–722 (2016).

260. Yusipov, I. *et al.* Age-related DNA methylation changes are sex-specific: a comprehensive assessment. *Aging (Albany. NY).* **12**, 24057–24080 (2020).

261. Gatev, E. *et al.* Autosomal sex-associated co-methylated regions predict biological sex from DNA methylation. *Nucleic Acids Res.* (2021) doi:10.1093/nar/gkab682.

262. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

263. Jansen, R. *et al.* Sex differences in the human peripheral blood transcriptome. *BMC Genomics* **15**, 33 (2014).

264. Gershoni, M. & Pietrokovski, S. The landscape of sex-differential transcriptome and its consequent selection in human adults. *BMC Biol.* **15**, 7 (2017).

265. Shen, E. Y. *et al.* Epigenetics and sex differences in the brain: A genome-wide comparison of histone-3 lysine-4 trimethylation (H3K4me3) in male and female mice. *Exp. Neurol.* **268**, 21–29 (2015).

266. Cedar, H. & Bergman, Y. Linking DNA methylation and histone modification: patterns and paradigms. *Nat. Rev. Genet.* **10**, 295–304 (2009).

267. Balasubramanian, D. *et al.* H3K4me3 inversely correlates with DNA methylation at a large class of non-CpG-island-containing start sites. *Genome Med.* **4**, 47 (2012).

268. Grafodatskaya, D. *et al.* Multilocus loss of DNA methylation in individuals with mutations in the histone H3 lysine 4 demethylase KDM5C. *BMC Med. Genomics* **6**, 1 (2013).

269. Rauch, T. A. *et al.* High-resolution mapping of DNA hypermethylation and hypomethylation in lung cancer. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 252–257 (2008).

270. Ozanne, B. W., Spence, H. J., McGarry, L. C. & Hennigan, R. F. Transcription factors control invasion: AP-1 the first among equals. *Oncogene* **26**, 1–10 (2007).

271. Sugiaman-Trapman, D. *et al.* Characterization of the human RFX transcription factor family by regulatory and target gene analysis. *BMC Genomics* **19**, 181 (2018).

272. Sun, W. *et al.* The association between copy number aberration, DNA methylation and gene expression in tumor samples. *Nucleic Acids Res.* **46**, 3009–3018 (2018).

273. Kim, N., Hong, Y., Kwon, D. & Yoon, S. Somatic Mutaome Profile in Human Cancer Tissues. *Genomics Inform.* **11**, 239 (2013).

274. Liu, P. *et al.* Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing. *Carcinogenesis* **33**, 1270–1276 (2012).

275. Pongor, L. *et al.* A genome-wide approach to link genotype to clinical outcome by utilizing next generation sequencing and gene chip data of 6,697 breast cancer patients. *Genome Med.* **7**, 1–11 (2015).

276. Thompson, D. J. *et al.* Genetic predisposition to mosaic Y chromosome loss in blood. *Nature* **575**, 652–657 (2019).

277. Zhou, W. *et al.* Mosaic loss of chromosome Y is associated with common variation near TCL1A. *Nat. Genet.* **48**, 563–568 (2016).

278. Wright, D. J. *et al.* Genetic variants associated with mosaic Y chromosome loss highlight cell cycle genes and overlap with cancer susceptibility. *Nat. Genet.* **49**, 674–679 (2017).

279. Komura, K. *et al.* ATR inhibition controls aggressive prostate tumors deficient in Y-linked histone demethylase KDM5D. *J. Clin. Invest.* **128**, 2979–2995 (2018).

280. Komura, K. *et al.* Resistance to docetaxel in prostate cancer is associated with androgen receptor activation and loss of KDM5D expression. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 6259–6264 (2016).

281. Wood, R. D., Mitchell, M. & Lindahl, T. Human DNA repair genes, 2005. *Mutat. Res.* **577**, 275–283 (2005).

282. Shen, W. H. *et al.* Essential role for nuclear PTEN in maintaining chromosomal integrity. *Cell* **128**, 157–170 (2007).

283. Paez, J. & Sellers, W. R. PI3K/PTEN/AKT pathway. A critical mediator of oncogenic signaling. *Cancer Treat. Res.* **115**, 145–167 (2003).

284. Michael R. Bronsert, William G. Henderson, Robert Valuck, Patrick Hosokawa, and K.

H. Oxidative Damage Targets Complexes Containing DNA Methyltransferases, SIRT1 and Polycomb Members to Promoter CpG Islands. *Bone* **23**, 1–7 (2008).

285. Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non–Small-Cell Lung Cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).
286. Roper, N. *et al.* APOBEC Mutagenesis and Copy-Number Alterations Are Drivers of Proteogenomic Tumor Evolution and Heterogeneity in Metastatic Thoracic Tumors. *Cell Rep.* **26**, 2651-2666.e6 (2019).
287. Forsberg, L. A. *et al.* Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. *Nat. Genet.* **46**, 624–628 (2014).
288. Forsberg, L. A. Loss of chromosome Y (LOY) in blood cells is associated with increased risk for disease and mortality in aging men. *Hum. Genet.* **136**, 657–663 (2017).
289. Xue, D. *et al.* TTN/TP53 mutation might act as the predictor for chemotherapy response in lung adenocarcinoma and lung squamous carcinoma patients. *Transl. Cancer Res.* **10**, 1284–1294 (2021).
290. Wang, Z., Wang, C., Lin, S. & Yu, X. Effect of TTN Mutations on Immune Microenvironment and Efficacy of Immunotherapy in Lung Adenocarcinoma Patients. *Front. Oncol.* **11**, 1–11 (2021).
291. Wilhelm, T., Said, M. & Naim, V. DNA Replication Stress and Chromosomal Instability: Dangerous Liaisons. *Genes (Basel).* **11**, 642 (2020).
292. Aitken, R. J. & Krausz, C. Oxidative stress, DNA damage and the Y chromosome. *Reproduction* **122**, 497–506 (2001).
293. Kato, T., Nagasawa, H., Warner, C., Okayasu, R. & Bedford, J. S. Cytotoxicity of cigarette smoke condensate is not due to DNA double strand breaks: Comparative studies using radiosensitive mutant and wild-type CHO cells. *Int. J. Radiat. Biol.* **83**, 583–591 (2007).
294. Albino, A. P. *et al.* Induction of DNA double-strand breaks in A549 and normal human pulmonary epithelial cells by cigarette smoke is mediated by free radicals. *Int. J. Oncol.* **28**, 1491–1505 (2006).
295. Roberts, S. A. *et al.* An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat. Genet.* **45**, 970–976 (2013).
296. Landry, S., Narvaiza, I., Linfesty, D. C. & Weitzman, M. D. APOBEC3A can activate the

DNA damage response and cause cell-cycle arrest. *EMBO Rep.* **12**, 444–450 (2011).

297. Nikkilä, J. *et al.* Elevated APOBEC3B expression drives a kataegic-like mutation signature and replication stress-related therapeutic vulnerabilities in p53-defective cells. *Br. J. Cancer* **117**, 113–123 (2017).

298. Green, A. M. *et al.* Cytosine Deaminase APOBEC3A Sensitizes Leukemia Cells to Inhibition of the DNA Replication Checkpoint. *Cancer Res.* **77**, 4579–4588 (2017).

299. Buisson, R., Lawrence, M. S., Benes, C. H. & Zou, L. APOBEC3A and APOBEC3B Activities Render Cancer Cells Susceptible to ATR Inhibition. *Cancer Res.* **77**, 4567–4578 (2017).

300. Domingo-Sabugo, C. *et al.* Distinct pancreatic and neuronal Lung Carcinoid molecular subtypes revealed by integrative omic analysis. *medRxiv* 2021.07.27.21260865 (2021).

301. Ullmann, R. *et al.* Unbalanced chromosomal aberrations in neuroendocrine lung tumors as detected by comparative genomic hybridization. *Hum. Pathol.* **29**, 1145–1149 (1998).

302. Derks, J. L. *et al.* New Insights into the Molecular Characteristics of Pulmonary Carcinoids and Large Cell Neuroendocrine Carcinomas, and the Impact on Their Clinical Management. *J. Thorac. Oncol. Off. Publ. Int. Assoc. Study Lung Cancer* **13**, 752–766 (2018).

303. Pelosi, G. *et al.* Most high-grade neuroendocrine tumours of the lung are likely to secondarily develop from pre-existing carcinoids: innovative findings skipping the current pathogenesis paradigm. *Virchows Arch.* **472**, 567–577 (2018).

304. Swarts, D. R. A., Ramaekers, F. C. S. & Speel, E. J. M. Molecular and cellular biology of neuroendocrine lung tumors: Evidence for separate biological entities. *Biochim. Biophys. Acta - Rev. Cancer* **1826**, 255–271 (2012).

305. Fernandez-Cuesta, L. *et al.* Frequent mutations in chromatin-remodelling genes in pulmonary carcinoids. *Nat. Commun.* **5**, 3518 (2014).

306. Caplin, M. E. *et al.* Pulmonary neuroendocrine (carcinoid) tumors: European Neuroendocrine Tumor Society expert consensus and recommendations for best practice for typical and atypical pulmonary carcinoids. *Ann. Oncol.* **26**, 1604–1620 (2015).

307. Swarts, D. R. A. *et al.* Interobserver variability for the WHO classification of pulmonary carcinoids. *Am. J. Surg. Pathol.* **38**, 1429–1436 (2014).

308. Feinberg, Y., Law, C., Singh, S. & Wright, F. C. Patient experiences of having a neuroendocrine tumour: A qualitative study. *Eur. J. Oncol. Nurs.* **17**, 541–545 (2013).

309. Randimbison, L., Rindi, G. & Vecchia, C. La. Epidemiology of carcinoid neoplasms in Vaud, Switzerland, 1974–97. *Br. J. Cancer* **83**, 952–955 (2000).

310. Skuladottir, H., Hirsch, F. R., Hansen, H. H. & Olsen, J. H. E-74. Pulmonary neuroendocrine tumors: Incidence and prognosis of histological subtypes. A population-based study in Denmark. *Lung Cancer* **41**, S92 (2003).

311. Govindan, R. *et al.* Changing epidemiology of small-cell lung cancer in the United States over the last 30 years: Analysis of the surveillance, epidemiologic, and end results database. *J. Clin. Oncol.* **24**, 4539–4544 (2006).

312. Yao, J. C. *et al.* One hundred years after 'carcinoid': Epidemiology of and prognostic factors for neuroendocrine tumors in 35,825 cases in the United States. *J. Clin. Oncol.* **26**, 3063–3072 (2008).

313. Chang, J. S. & Chen, L. The epidemiologic trends of neuroendocrine tumors in Taiwan : an updated analysis of a nation-wide population-based study. 1–16 (2020) doi:10.21203/rs.3.rs-69891/v1.

314. Hallet, J. *et al.* Exploring the rising incidence of neuroendocrine tumors: A population-based analysis of epidemiology, metastatic presentation, and outcomes. *Cancer* **121**, 589–597 (2015).

315. Ellis, L., Shale, M. J. & Coleman, M. P. Carcinoid tumors of the gastrointestinal tract: trends in incidence in England since 1971. *Am. J. Gastroenterol.* **105**, 2563–2569 (2010).

316. Murtagh, F. & Legendre, P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *J. Classif.* **31**, 274–295 (2014).

317. Müller-Pillasch, F. *et al.* Identification of a new tumour-associated antigen TM4SF5 and its expression in human cancer. *Gene* **208**, 25–30 (1998).

318. Lee, S. A. *et al.* Tetraspanin TM4SF5 mediates loss of contact inhibition through epithelial-mesenchymal transition in human hepatocarcinoma. *J. Clin. Invest.* **118**, 1354–1366 (2008).

319. Kwon, S. *et al.* Prophylactic effect of a peptide vaccine targeting TM4SF5 against colon cancer in a mouse model. *Biochem. Biophys. Res. Commun.* **435**, 134–139 (2013).

320. Wu, Y. B. *et al.* A high level of TM4SF5 is associated with human esophageal cancer

progression and poor patient survival. *Dig. Dis. Sci.* **58**, 2623–2633 (2013).

321. Lee, J. W. TM4SF5-mediated protein-protein networks and tumorigenic roles. *BMB Rep.* **47**, 483–487 (2014).
322. Park, S., Kim, D., Park, J. A., Kwon, H. J. & Lee, Y. Targeting TM4SF5 with anti-TM4SF5 monoclonal antibody suppresses the growth and motility of human pancreatic cancer cells. *Oncol. Lett.* **19**, 641–650 (2020).
323. Abel, E. V. *et al.* HNF1A is a novel oncogene that regulates human pancreatic cancer stem cell properties. *Elife* **7**, 1–35 (2018).
324. Richards et al. ASCL1 and NEUROD1 reveal heterogeneity in pulmonary neuroendocrine tumors and regulate distinct genetic programs. *Physiol. Behav.* **176**, 139–148 (2018).
325. Cao, X. *et al.* Pancreatic-derived factor (FAM3B), a novel islet cytokine, induces apoptosis of insulin-secreting β-cells. *Diabetes* **52**, 2296–2303 (2003).
326. Mou, H. *et al.* Knockdown of FAM3B triggers cell apoptosis through p53-dependent pathway. *Int. J. Biochem. Cell Biol.* **45**, 684–691 (2013).
327. Zhu, Y. *et al.* Cloning, expression, and initial characterization of a novel cytokine-like gene family. *Genomics* **80**, 144–150 (2002).
328. Maciel-Silva, P. *et al.* FAM3B/PANDER inhibits cell death and increases prostate tumor growth by modulating the expression of Bcl-2 and Bcl-XL cell survival genes. *BMC Cancer* **18**, 1–15 (2018).
329. Li, Z. *et al.* A non-secretory form of FAM3B promotes invasion and metastasis of human colon cancer cells by upregulating Slug expression. *Cancer Lett.* **328**, 278–284 (2013).
330. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. **1**, 417–425 (2016).
331. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Rep.* **3**, 246–259 (2013).
332. Ouadah, Y. *et al.* Rare Pulmonary Neuroendocrine Cells Are Stem Cells Regulated by Rb, p53, and Notch. *Cell* **179**, 403-416.e23 (2019).
333. Zarrei, M., MacDonald, J. R., Merico, D. & Scherer, S. W. A copy number variation map of the human genome. *Nat. Rev. Genet.* **16**, 172–183 (2015).

334. McNeely, K. C. *et al.* Mutation of Kinesin-6 Kif20b causes defects in cortical neuron polarization and morphogenesis. *Neural Dev.* **12**, 1–18 (2017).

335. Golpon, H. A. *et al.* HOX Genes in Human Lung. *Am. J. Pathol.* **158**, 955–966 (2001).

336. Shah, N. & Sukumar, S. The Hox genes and their roles in oncogenesis. *Nat. Rev. Cancer* **10**, 361–371 (2010).

337. Bollati, V. *et al.* Decline in Genomic DNA Methylation through Aging in a Cohort of Elderly Subjects. *October* **130**, 234–239 (2010).

338. Marinoni, I. *et al.* Loss of DAXX and ATRX are associated with chromosome instability and reduced survival of patients with pancreatic neuroendocrine tumors. *Gastroenterology* **146**, 453–60.e5 (2014).

339. Kim, J. Y. *et al.* Alternative Lengthening of Telomeres in Primary Pancreatic Neuroendocrine Tumors Is Associated with Aggressive Clinical Behavior and Poor Survival. *Clin. Cancer Res.* **23**, 1598–1606 (2017).

340. Jiao, Y. *et al.* DAXX/ATRX, MEN1, and mTOR pathway genes are frequently altered in pancreatic neuroendocrine tumors. *Science* **331**, 1199–1203 (2011).

341. Lakis, V. *et al.* DNA methylation patterns identify subgroups of pancreatic neuroendocrine tumors with clinical association. *Commun. Biol.* **4**, 155 (2021).

342. Flynn, R. L. *et al.* Alternative lengthening of telomeres renders cancer cells hypersensitive to ATR inhibitors. *Science* **347**, 273–277 (2015).

343. Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007).

344. Sundaram, V. & Wysocka, J. Transposable elements as a potent source of diverse cis-regulatory sequences in mammalian genomes. *Philos. Trans. R. Soc. B Biol. Sci.* **375**, (2020).

345. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: From conflicts to benefits. *Nat. Rev. Genet.* **18**, 71–86 (2017).

346. Papaxoinis, G., Lamarca, A., Quinn, A. M., Mansoor, W. & Nonaka, D. Clinical and Pathologic Characteristics of Pulmonary Carcinoid Tumors in Central and Peripheral Locations. *Endocr. Pathol.* **29**, 259–268 (2018).

347. Asiedu, M. K. *et al.* Pathways impacted by genomic alterations in pulmonary carcinoid tumors. *Clin. Cancer Res.* **24**, 1691–1704 (2018).

348. Yu, F.-L., Bender, W. & Geronimo, I. H. Base and sequence specificities of aflatoxin B1 binding to single- and double-stranded DNAs. *Carcinogenesis* **11**, 475–478 (1990).

349. Bannasch, P. *et al.* Synergistic Hepatocarcinogenic Effect of Hepadnaviral Infection and Dietary Aflatoxin B1 in Woodchucks. *Cancer Res.* **55**, 3318–3330 (1995).

350. Chawanthayatham, S. *et al.* Mutational spectra of aflatoxin B1 in vivo establish biomarkers of exposure for human hepatocellular carcinoma. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E3101–E3109 (2017).

351. Schulze, K. *et al.* Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nat. Genet.* **47**, 505–511 (2015).

352. Smela, M. E. *et al.* The aflatoxin B1 formamidopyrimidine adduct plays a major role in causing the types of mutations observed in human hepatocellular carcinoma. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 6655–6660 (2002).

353. Woo, L. L. *et al.* Aflatoxin B1-DNA adduct formation and mutagenicity in livers of neonatal male and female B6C3F1 mice. *Toxicol. Sci.* **122**, 38–44 (2011).

354. Alekseyev, Y. O., Hamm, M. L. & Essigmann, J. M. Aflatoxin B1 formamidopyrimidine adducts are preferentially repaired by the nucleotide excision repair pathway in vivo. *Carcinogenesis* **25**, 1045–1051 (2004).

355. Coulombe, R., Eaton, D. & Groopman, J. Nonhepatic Disposition and Effects of Aflatoxin B1. in (1993).

356. Gursoy, N. *et al.* Changes in spontaneous contractions of rat ileum by aflatoxin in vitro. *Food Chem. Toxicol.* **46**, 2124–2127 (2008).

357. Sorenson, W. G., Simpson, J. P., III, M. J. P., Thedell, T. D. & Olenchock, S. A. Aflatoxin in respirable corn dust particles. *J. Toxicol. Environ. Health* **7**, 669–672 (1981).

358. Burg, W. R. & Shotwell, O. L. Aflatoxin Levels in Airborne Dust Generated from Contaminated Corn During Harvest and at an Elevator in 1980. *J. Assoc. Off. Anal. Chem.* **67**, 309–312 (1984).

359. Ramiro, A. R. & Barreto, V. M. Activation-induced cytidine deaminase and active DNA demethylation. *Trends Biochem. Sci.* **40**, 172–181 (2015).

360. Jonathan Posner and Bradley S. Peterson, J. A. R. Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. *Bone* **23**, 1–7 (2008).

361. Salta, E. & De Strooper, B. Non-coding RNAs with essential roles in neurodegenerative

disorders. *Lancet Neurol* **11**, 189–200 (2012).

362. Burns, K. H. Transposable elements in cancer. *Nat. Rev. Cancer* **17**, 415–424 (2017).

363. Aguilera, A. & García-Muse, T. Causes of genome instability. *Annu. Rev. Genet.* **47**, 1–32 (2013).

364. Chen, Y., Breeze, C. E., Zhen, S., Beck, S. & Teschendorff, A. E. Tissue-independent and tissue-specific patterns of DNA methylation alteration in cancer. *Epigenetics Chromatin* **9**, 10 (2016).

365. Pérez, R. F., Tejedor, J. R., Bayón, G. F., Fernández, A. F. & Fraga, M. F. Distinct chromatin signatures of DNA hypomethylation in aging and cancer. *Aging Cell* **17**, e12744 (2018).

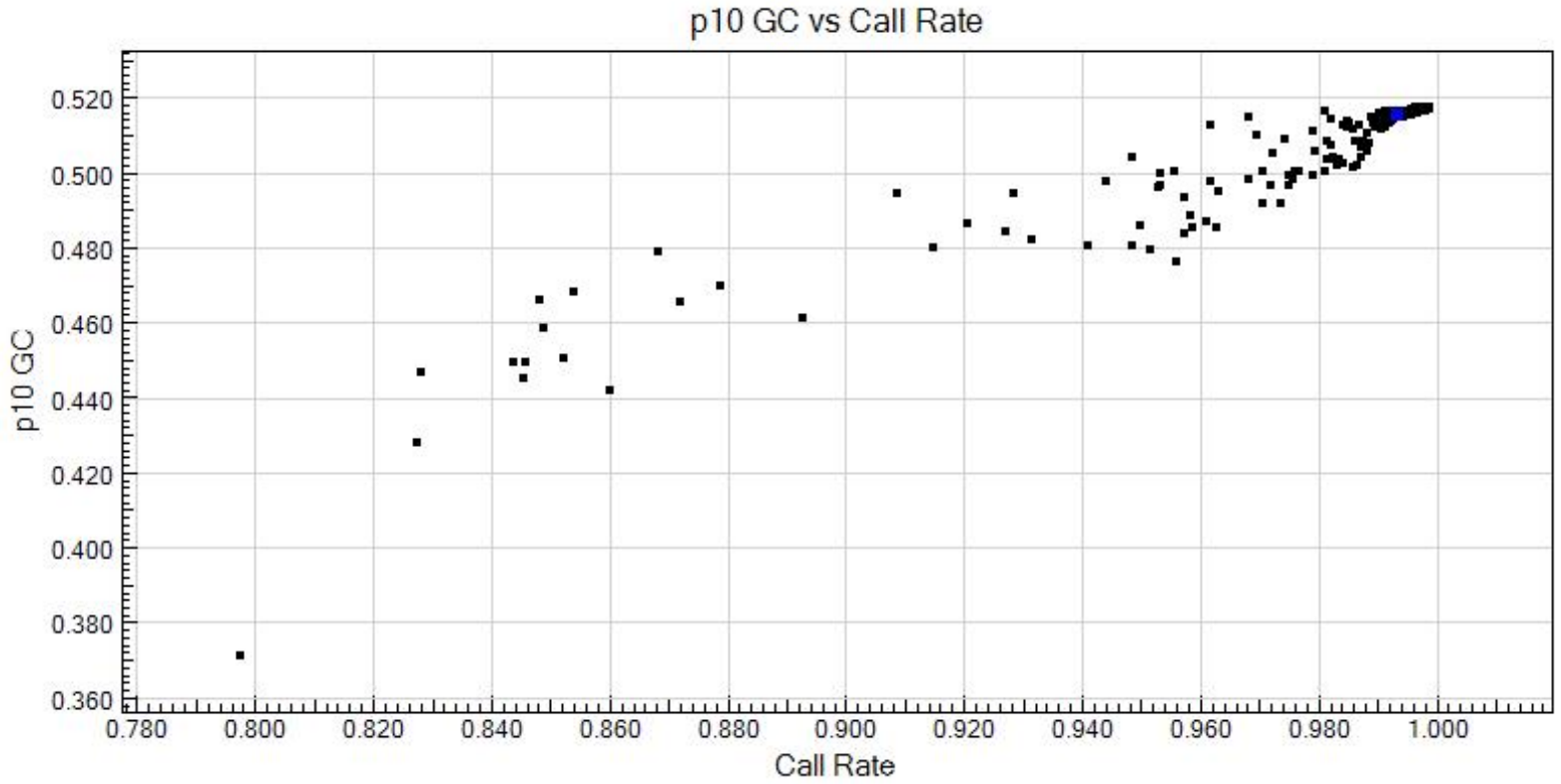
Supplementary Data

Chapter 3

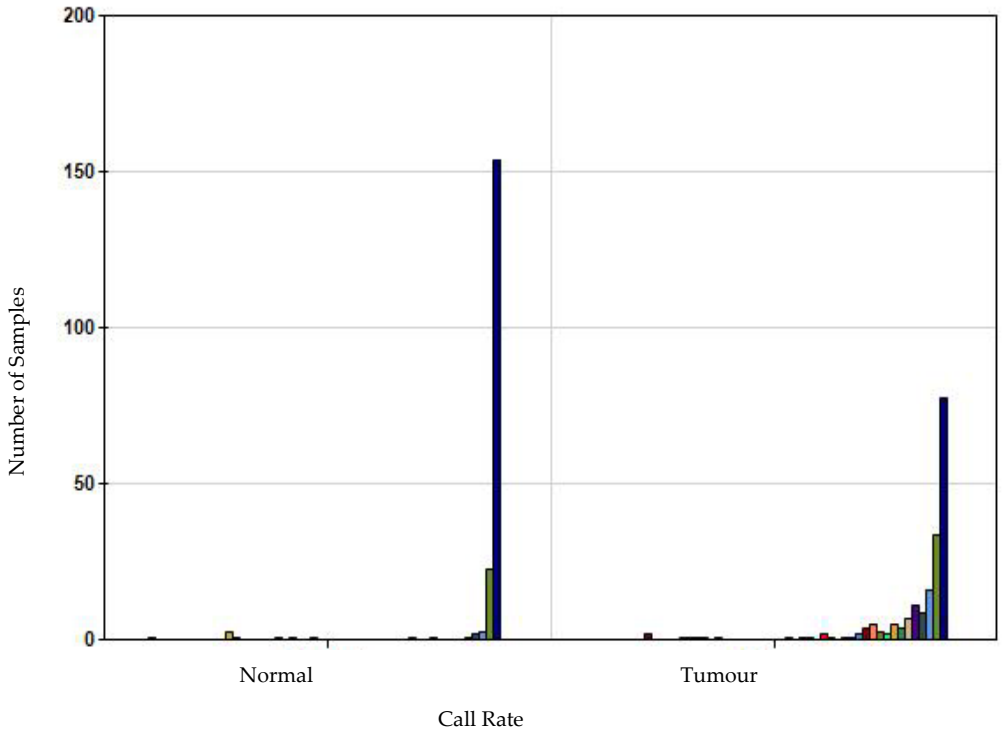
Supplementary Table 3.1| Recurrent COSMIC mutational signatures detected in L-CD tumours using deconstructSigs on WES data. Abbreviations: Whole exome sequencing (WES); lung carcinoids (L-CD).

CMS	*Cancer types*	*Proposed Aetiology*	*Comments*	*n samples with CMS (%)*
3	Breast, ovarian, and pancreatic cancers	Failure of DNA double-strand break-repair by homologous recombination	Associated with germline and somatic BRCA1 and BRCA2 mutations in breast, pancreatic, and ovarian cancers. In pancreatic cancer, responders to platinum therapy usually exhibit Signature 3 mutations	10 (45.5%)
16	Liver cancer	Unknown		8 (36.4%)
8	Breast cancer and medulloblastoma	Unknown		8 (36.4%)
29	Gingivo-buccal oral squamous cell carcinoma	Tobacco chewing habit		8 (36.4%)
5	All cancer types and most cancer samples	Unknown		7 (31.8%)
20	Stomach and breast cancers	Defective DNA mismatch repair	Often found in the same samples as Signatures 6, 15, and 26 associated with DNA mismatch repair	7 (31.8%)
24	Subset of liver cancers	Found in cancer samples with known exposures to aflatoxin		5 (22.7%)
26	Breast cancer, cervical cancer, stomach cancer and uterine carcinoma	Defective DNA mismatch repair		5 (22.7%)

Supplementary Figure 3.1| Scatter plot of 10% GC scores compared to call rates for Lung Cancer samples. p10 GC represents the 10th percentile of the GenCall score across all called genotypes and higher values indicate the reliability of the genotype called. Call rate and p10 GC are positively correlated.



Supplementary Figure 3.2| Histogram of Call Rate for SNP genotyped LC samples representing the frequency of call rate in tumour versus normal samples.



Supplementary Figure 3.3|Correlation of inferred Tumour purity and Tumour content determined by immunohistochemical analysis. Spearman correlation coefficient and *P*-value are shown for the whole lung cancer (LC) dataset. Abbreviations: LUAD, Lung Adenocarcinoma; LUSC, Lung Squamous Carcinoma; LNET, Lung Neuroendocrine Tumour; L-CD, Lung Carcinoid.

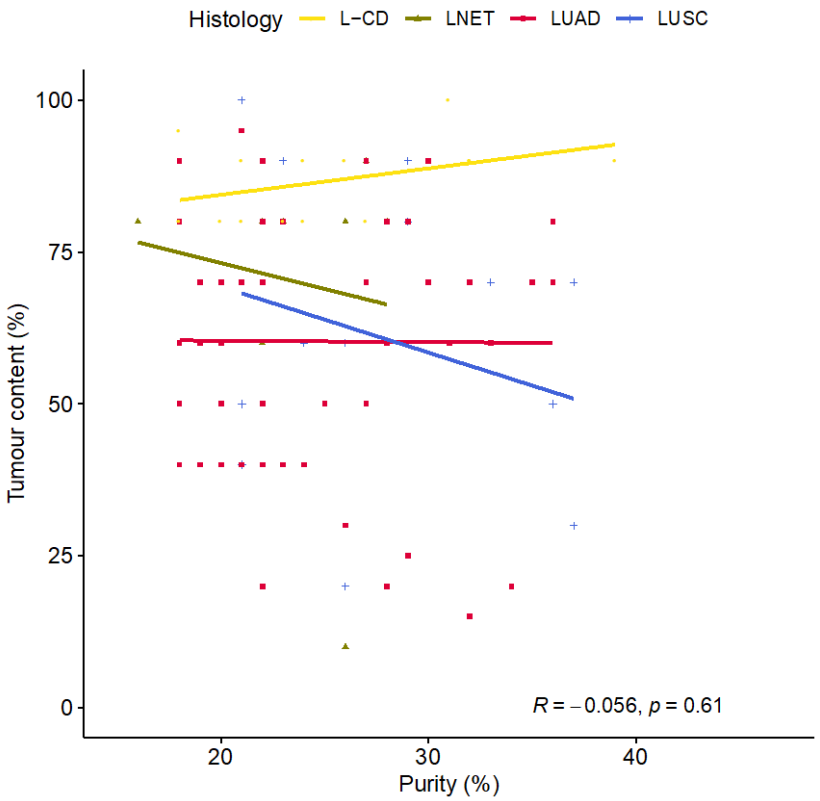
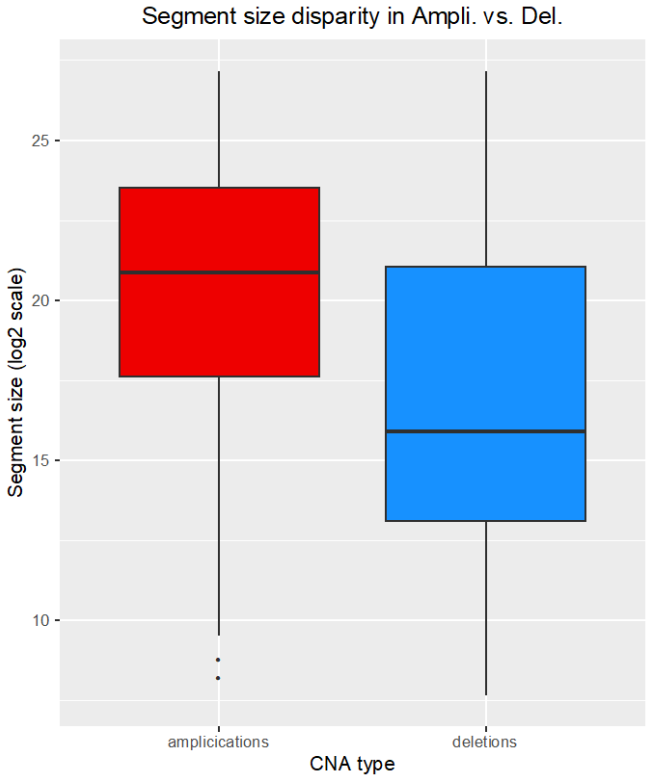


Table 3.2| Significant peaks detected in the normal lung cancer data with GISTIC that were subtracted from the Lung Cancer tumour data. Significant peaks were considered when a residual q-value of $<1x10^{-5}$ was obtained.

Unique Name	*Descriptor*	*Peak Limits*	*q values*	*Residual q values after removing segments shared with higher peaks*
Deletion Peak 14	10q22.2	chr10:76855413-76867752(probes 382025:382035)	2.21E-34	3.87E-33
Deletion Peak 16	12q13.12	chr12:49425197-49433598(probes 446233:446241)	2.66E-07	2.66E-07
Deletion Peak 17	12q21.31	chr12:85444295-85450290(probes 455465:455475)	4.03E-29	1.75E-28
Deletion Peak 19	14q32.11	chr14:91699419-91747826(probes 507689:507700)	1.56E-15	1.56E-15
Deletion Peak 20	15q15.1	chr15:42144040-42150800(probes 516377:516386)	2.32E-10	2.32E-10
Deletion Peak 21	16p13.3	chr16:3611762-3614135(probes 534372:534379)	2.66E-07	2.66E-07
Deletion Peak 25	19p13.3	chr19:2396594-2422071(probes 591512:591521)	6.74E-19	6.74E-19
Deletion Peak 26	19q13.2	chr19:40519936-40542425(probes 602359:602371)	5.39E-29	5.39E-29
Deletion Peak 27	20q12	chr20:39976285-39980461(probes 618409:618420)	1.52E-06	1.52E-06
Deletion Peak 28	22q11.23	chr22:24977287-24983109(probes 634910:634920)	9.90E-06	9.90E-06
Deletion Peak 3	2p13.1	chr2:73653611-73678538(probes 76994:77011)	2.45E-13	2.45E-13
Amplification Peak 1	5q31.3	chr5:139931385-139931690(probes 218698:218707)	2.26E-13	2.26E-13
Deletion Peak 5	5q31.3	chr5:139931385-139931660(probes 218698:218706)	2.45E-91	2.45E-91
Deletion Peak 6	6p21.31	chr6:36336804-36345976(probes 241042:241053)	1.85E-16	1.85E-16
Deletion Peak 7	6q22.1	chr6:116441647-116443022(probes 257545:257554)	2.86E-145	2.86E-145
Deletion Peak 10	8q24.3	chr8:144995957-145000981(probes 334753:334755)	1.52E-06	1.52E-06
Deletion Peak 12	9q22.33	chr9:99521191-99521889(probes 351735:351742)	5.98E-34	3.00E-33
Deletion Peak 13	9q34.11	chr9:131457126-131475467(probes 360506:360515)	1.30E-09	1.51E-07
Deletion Peak 30	Xp11.21	chrX:57475062-57875541(probes 645540:645546)	2.53E-12	2.53E-12
Deletion Peak 32	Xq21.1	chrX:83126544-83141578(probes 646412:646413)	1.42E-35	1.71E-25
Deletion Peak 33	Xq22.3	chrX:107816827-108023532(probes 647561:647570)	2.37E-100	2.91E-95

Figure 3.4| Median segment size of segments deemed to be amplifications (Ampli) and deletions (Del). Median sizes were larger for amplification events: amplifications had a median segment size of ~2,000 Kb, while deletions had a median size of ~61 Kb.

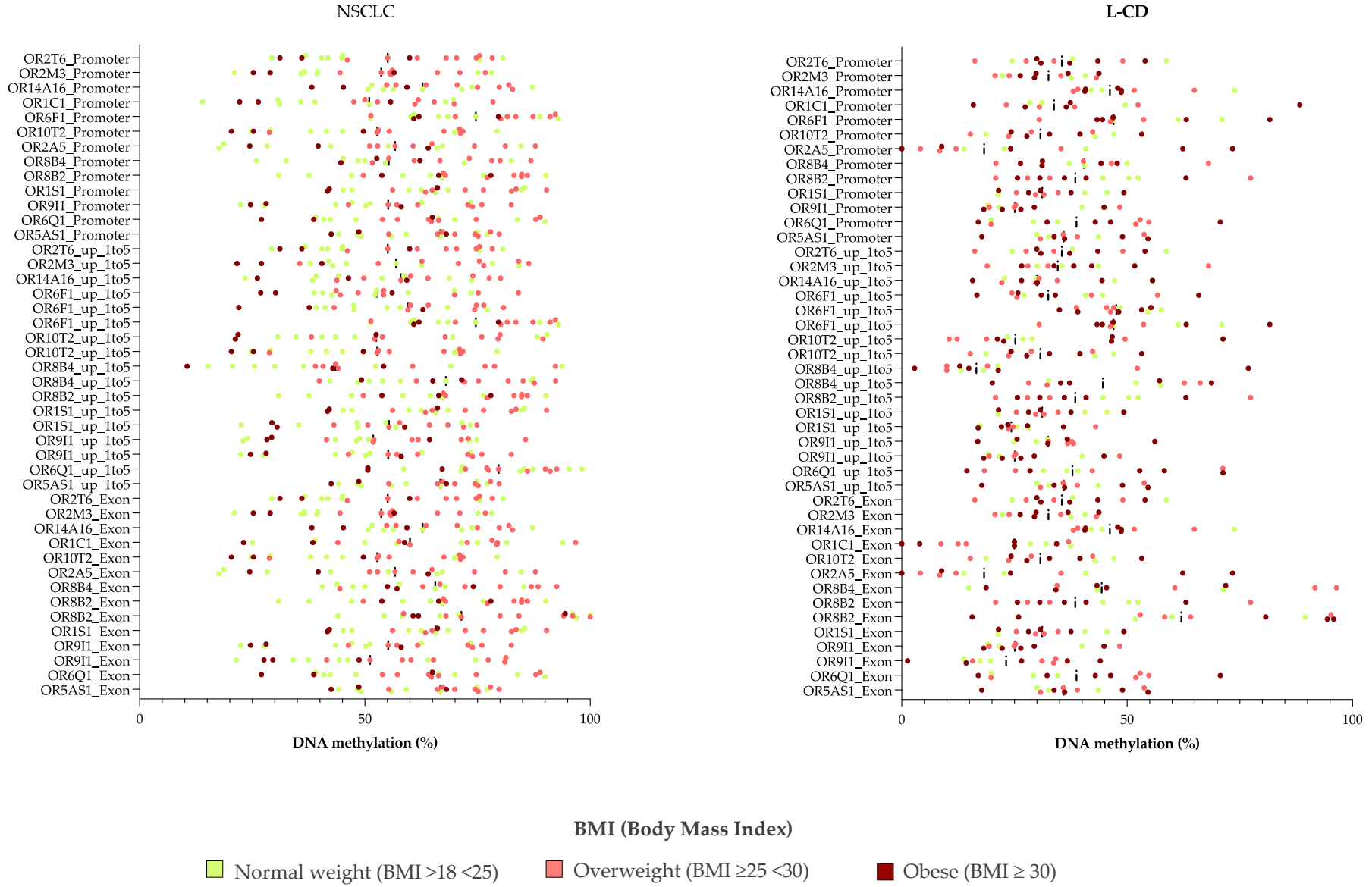


Chapter 4

Supplementary Table 4.1| Number of Differentially Methylated Regions (DMRs) hypomethylated and hypermethylated at each genomic annotation category. NSCLC vs L-CD refers to the tumour-tumour differential methylation comparison, whereas NSCLC and L-CD refers to the tumour-normal differential methylation comparison. Abbreviations: L-CD, Lung Carcinoids; NSCLC, Non-Small Cell Lung Cancer (includes Lung Squamous Carcinoma [LUSC] and Lung Adenocarcinoma [LUAD]); CGI, CpG Island; LncRNA, Long non-coding RNA; UTR, Untranslated region; Kb, Kilo base.

	NSCLC vs L-CD			NSCLC			L-CD		
	Hypo	Hyper	Total	Hypo	Hyper	Total	Hypo	Hyper	Total
1to5Kb	1470	2409	3879	684	202	886	4650	1157	5807
promoters	790	1091	1881	278	178	456	1926	767	2693
5'UTR	460	873	1333	130	140	270	1150	421	1571
exons	1160	3022	4182	539	387	926	4604	1107	5711
introns	3832	6246	10078	1839	690	2529	8062	3411	11473
3'UTR	303	862	1165	132	81	213	1457	303	1760
InterCGI	17132	41434	58566	32908	41434	74342	151562	41434	192996
CpG shelf	1078	2027	3105	587	2027	2614	4804	2027	6831
CpG shore	2286	2709	4995	577	2709	3286	4868	2709	7577
CGI	415	833	1248	38	833	871	568	833	1401
enhancers	2925	1235	4160	647	1235	1882	2580	1235	3815
lncRNAs	1284	2245	3529	1243	171	1414	4008	978	4986

Supplementary Figure 4.1| DNA methylation levels at 1 to 5 Kb of promoter regions, within promoters and in exonic regions of olfactory receptors (OR) detected hypomethylated in both NSCLC and L-CD tumours. Body Mass Index (BMI) of patients is shown in green (normal weight; BMI >18-<25), overweight (salmon; BMI ≥25 - <30) and obese (garnet; BMI ≥ 30).



Supplementary Table 4.2| Significant pathways enriched in genes whose promoters of regions up to 5 Kb were identified hypomethylated and hypermethylated. a) between L-CD tumours and their normal matched tissue **b)** between NSCLC tumours and their normal matched tissue and **c)** between NSCLC tumours and L-CD tumours.

a

Promoters	Pathway name	Entities found	Entities total	Entities ratio	Entities pValue	Entities FDR
Hypo	Olfactory Signaling Pathway	181	432	2.97E-02	1.11E-16	9.26E-14
	Sensory Perception	196	681	4.68E-02	1.11E-16	9.26E-14
Hyper	Olfactory Signaling Pathway	181	432	2.97E-02	1.11E-16	9.26E-14
	Sensory Perception	196	681	4.68E-02	1.11E-16	9.26E-14

1to5Kb	Pathway name	Entities found	Entities total	Entities ratio	Entities pValue	Entities FDR
Hypo	Olfactory Signaling Pathway	257	432	2.97E-02	1.44E-15	3.11E-12
	Sensory Perception	287	681	4.68E-02	1.09E-03	9.85E-01
Hyper	Attenuation phase	14	47	3.23E-03	1.44E-04	2.27E-01
	Apoptotic cleavage of cell adhesion proteins	6	11	7.56E-04	5.87E-04	4.63E-01
	RUNX1 regulates expression of components of tight junctions	5	8	5.50E-04	9.22E-04	4.85E-01
	HSF1-dependent transactivation	14	59	4.06E-03	1.30E-03	5.11E-01
	Other semaphorin interactions	7	19	1.31E-03	2.02E-03	6.14E-01
	HSF1 activation	11	43	2.96E-03	2.34E-03	6.14E-01
	EGR2 and SOX10-mediated initiation of Schwann cell myelination	10	39	2.68E-03	3.55E-03	7.50E-01
	Signaling by FGFR2 IIIa TM	7	24	1.65E-03	7.07E-03	7.50E-01
	Sema3A PAK dependent Axon repulsion	6	19	1.31E-03	8.53E-03	7.50E-01
	Signaling by FGFR in disease	15	82	5.64E-03	9.50E-03	7.50E-01
	Regulation of gene expression in endocrine-committed (NEUROG3+) progenitor cells	4	9	6.19E-04	9.74E-03	7.50E-01

b

Promoters

	Pathway name	#Entities found	#Entities total	Entities ratio	Entities pValue	Entities FDR
Hypo	Olfactory Signaling Pathway	26	432	2.97E-02	1.82E-06	1.72E-03
	Sensory Perception	28	681	4.68E-02	5.57E-04	2.63E-01
	RUNX3 regulates RUNX1-mediated transcription	2	4	2.75E-04	3.28E-03	5.08E-01
	Activation of Matrix Metalloproteinases	4	35	2.41E-03	6.34E-03	5.08E-01
	Downregulation of ERBB2 signaling	4	36	2.48E-03	7.20E-03	5.08E-01
	transition)	3	19	1.31E-03	7.65E-03	5.08E-01
	Fatty Acids bound to GPR40 (FFAR1) regulate insulin secretion	3	19	1.31E-03	7.65E-03	5.08E-01
	Acetylcholine regulates insulin secretion	3	19	1.31E-03	7.65E-03	5.08E-01
	Formation of the cornified envelope	8	138	9.49E-03	9.08E-03	5.08E-01
	signaling	2	7	4.81E-04	9.63E-03	5.08E-01
	RUNX1 regulates transcription of genes involved in BCR signaling	2	7	4.81E-04	9.63E-03	5.08E-01
Hyper	Activation of HOX genes during differentiation	12	116	7.98E-03	1.02E-07	2.64E-05
	Activation of anterior HOX genes in hindbrain development during early embryogenesis	12	116	7.98E-03	1.02E-07	2.64E-05
	Hyaluronan biosynthesis and export	2	6	4.13E-04	3.21E-03	4.82E-01
	POU5F1 (OCT4), SOX2, NANOG repress genes related to differentiation	2	10	6.88E-04	8.60E-03	4.82E-01

1to5Kb

	Pathway name	#Entities found	#Entities total	Entities ratio	Entities pValue	Entities FDR
Hypo	Olfactory Signaling Pathway	51	432	2.97E-02	8.03E-08	1.00E-04
	MECP2 regulates transcription factors	6	10	6.88E-04	1.68E-05	1.05E-02
	Neurexins and neuroligins	12	60	4.13E-03	9.28E-05	3.86E-02
	Sensory Perception	58	681	4.68E-02	1.85E-04	5.76E-02
	TP53 Regulates Transcription of Death Receptors and Ligands	5	18	1.24E-03	2.68E-03	6.70E-01
	Attenuation phase	8	47	3.23E-03	3.53E-03	6.98E-01
	Protein-protein interactions at synapses	12	93	6.40E-03	3.92E-03	6.98E-01
	MECP2 regulates neuronal receptors and channels	6	32	2.20E-03	6.99E-03	7.16E-01
	TRAIL signaling	3	8	5.50E-04	8.66E-03	7.16E-01
Hyper	Activation of anterior HOX genes in hindbrain development during early embryogenesis	20	116	7.98E-03	9.21E-15	2.99E-12
	Activation of HOX genes during differentiation	20	116	7.98E-03	9.21E-15	2.99E-12
	factors	5	11	7.56E-04	1.20E-06	2.61E-04
	Transcriptional regulation of pluripotent stem cells	7	45	3.09E-03	1.03E-05	1.65E-03
	Developmental Biology	41	1262	8.68E-02	1.27E-05	1.65E-03
	Regulation of gene expression in endocrine-committed (NEUROG3+) progenitor cells	4	9	6.19E-04	1.39E-05	1.72E-03
	Activation of the TFAP2 (AP-2) family of transcription factors	4	12	8.25E-04	4.85E-05	4.51E-03
	Regulation of beta-cell development	7	67	4.61E-03	1.24E-04	9.93E-03
	RUNX3 regulates BCL2L11 (BIM) transcription	3	6	4.13E-04	1.38E-04	9.93E-03
	POU5F1 (OCT4), SOX2, NANOG activate genes related to proliferation	4	21	1.44E-03	4.06E-04	2.64E-02
	PKA activation in glucagon signalling	4	23	1.58E-03	5.70E-04	3.37E-02
	Transcriptional regulation by the AP-2 (TFAP2) family of transcription factors	5	52	3.58E-03	1.66E-03	8.99E-02
	homeostasis	3	15	1.03E-03	1.94E-03	9.69E-02
	factors	2	5	3.44E-04	3.06E-03	1.39E-01
	G alpha (z) signalling events	5	62	4.26E-03	3.53E-03	1.39E-01
	Glucagon signaling in metabolic regulation	4	40	2.75E-03	4.22E-03	1.39E-01
	SUMOylation of transcription factors	3	20	1.38E-03	4.33E-03	1.39E-01
	TFAP2 (AP-2) family regulates transcription of cell cycle factors	2	6	4.13E-04	4.36E-03	1.39E-01
	MTF1 activates gene expression	2	6	4.13E-04	4.36E-03	1.39E-01
	Hyaluronan biosynthesis and export	2	6	4.13E-04	4.36E-03	1.39E-01
	Transcriptional regulation of testis differentiation	3	21	1.44E-03	4.95E-03	1.44E-01
	TFAP2 (AP-2) family regulates transcription of growth factors and their receptors	3	21	1.44E-03	4.95E-03	1.44E-01

C

Promoters	Pathway name	Entities found	Entities total	Entities ratio	Entities pValue	Entities FDR
Hypo	Regulated Necrosis	15	77	5.30E-03	1.42E-04	1.08E-01
	CLEC7A/inflammasome pathway	5	8	5.50E-04	1.70E-04	1.08E-01
	Neutrophil degranulation	51	480	3.30E-02	2.26E-04	1.08E-01
	RUNX1 regulates transcription of genes involved in differentiation of myeloid cells	5	11	7.56E-04	7.16E-04	2.56E-01
	TRAIL signaling	4	8	5.50E-04	1.74E-03	3.90E-01
	CASP8 activity is inhibited	5	14	9.63E-04	2.05E-03	3.90E-01
	RUNX3 regulates RUNX1-mediated transcription	3	4	2.75E-04	2.15E-03	3.90E-01
	Interleukin-4 and Interleukin-13 signaling	25	211	1.45E-02	2.19E-03	3.90E-01
	Regulation of necroptotic cell death	8	38	2.61E-03	3.07E-03	4.85E-01
	FasL/ CD95L signaling	3	5	3.44E-04	4.01E-03	5.69E-01
	Regulation by c-FLIP	4	11	7.56E-04	5.36E-03	6.08E-01
	Dimerization of procaspase-8	4	11	7.56E-04	5.36E-03	6.08E-01
	TP53 Regulates Transcription of Death Receptors and Ligands	5	18	1.24E-03	5.90E-03	6.08E-01
	Pyroptosis	7	34	2.34E-03	6.14E-03	6.08E-01
	RUNX2 regulates genes involved in differentiation of myeloid cells	3	6	4.13E-04	6.62E-03	6.08E-01
	TP53 Regulates Transcription of Cell Death Genes	12	83	5.71E-03	6.96E-03	6.08E-01
	Interleukin-1 processing	4	12	8.25E-04	7.24E-03	6.08E-01
	Caspase activation via Death Receptors in the presence of ligand	5	20	1.38E-03	9.03E-03	6.94E-01
	RIPK1-mediated regulated necrosis	8	46	3.16E-03	9.25E-03	6.94E-01
Hyper	Nuclear Receptor transcription pathway	18	86	5.91E-03	3.31E-04	7.75E-01
	MECP2 regulates transcription factors	4	10	6.88E-04	1.09E-02	8.60E-01
	FBXW7 Mutants and NOTCH1 in Cancer	3	6	4.13E-04	1.49E-02	8.60E-01
	Loss of Function of FBXW7 in Cancer and NOTCH1 Signaling	3	6	4.13E-04	1.49E-02	8.60E-01

1to5Kb	Pathway name	Entities found	Entities total	Entities ratio	Entities pValue	Entities FDR
Hypo	Attenuation phase	14	47	3.23E-03	1.22E-03	8.30E-01
	HSF1-dependent transactivation	15	59	4.06E-03	3.70E-03	8.30E-01
	HSF1 activation	12	43	2.96E-03	4.38E-03	8.30E-01
	EGR2 and SOX10-mediated initiation of Schwann cell myelination	11	39	2.68E-03	5.73E-03	8.30E-01
	BH3-only proteins associate with and inactivate anti-apoptotic BCL-2 members	5	11	7.56E-04	8.95E-03	8.30E-01
Hyper	Endosomal/Vacuolar pathway	38	82	5.64E-03	5.76E-07	1.11E-03
	Antigen Presentation: Folding, assembly and peptide loading of class I MHC	38	102	7.01E-03	6.82E-05	6.54E-02
	Activation of anterior HOX genes in hindbrain development during early embryogenesis	37	116	7.98E-03	1.42E-03	6.83E-01
	Activation of HOX genes during differentiation	37	116	7.98E-03	1.42E-03	6.83E-01
	Interferon alpha/beta signaling	51	186	1.28E-02	4.74E-03	9.47E-01
	ER-Phagosome pathway	44	173	1.19E-02	2.54E-02	9.47E-01
	Hyaluronan biosynthesis and export	4	6	4.13E-04	2.66E-02	9.47E-01
	Interferon gamma signaling	60	250	1.72E-02	2.94E-02	9.47E-01
	Apoptosis induced DNA fragmentation	6	13	8.94E-04	3.62E-02	9.47E-01
	The AIM2 inflammasome	3	4	2.75E-04	3.94E-02	9.47E-01
	RUNX3 Regulates Immune Response and Cell Migration	5	10	6.88E-04	4.04E-02	9.47E-01

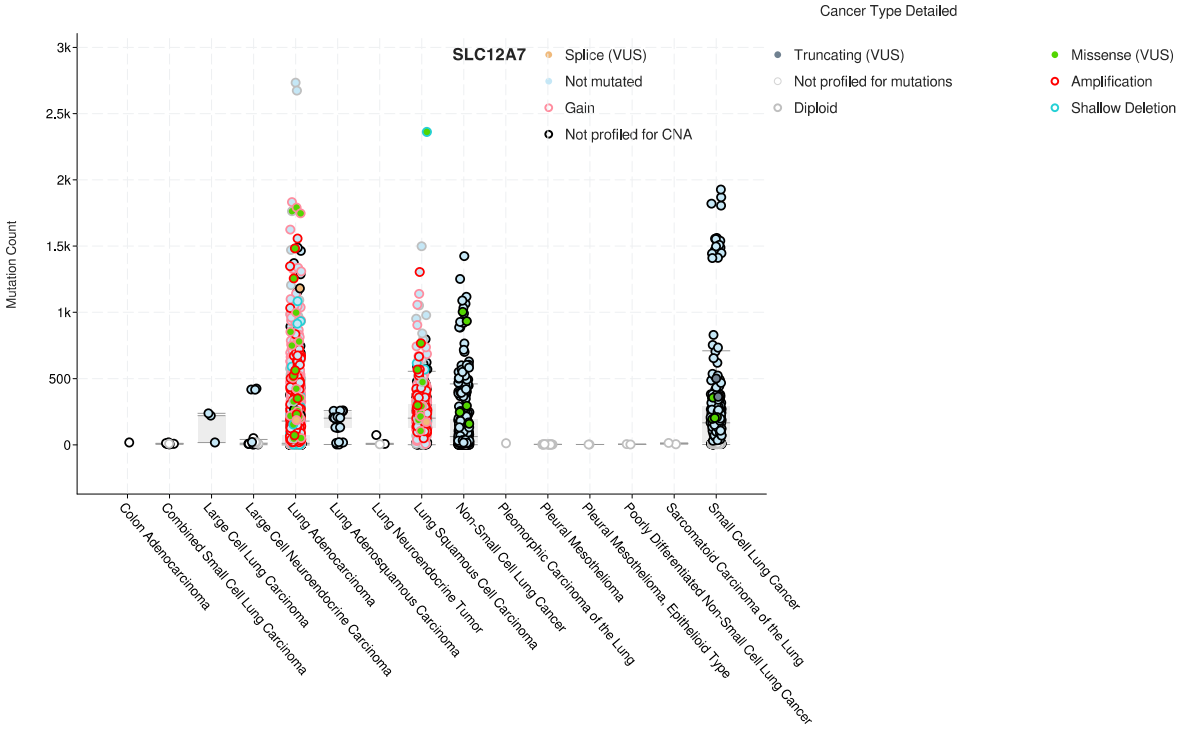
Supplementary Table 4.4| Genes identified mutated and differentially methylated at the promoter level in L-CDs.

38 common elements in "DMR" and "Mutated"	Status
ABHD10	Hypomethylated
ADAMTS18	Hypomethylated
ANKRD11	Hypomethylated
APBB1	Hypomethylated
APCDD1	Hypermethylated
ATP4A	Hypomethylated
BNC1	Hypermethylated
C6	Hypomethylated
CHSY1	Hypomethylated
CTNNA3	Hypomethylated
CTSW	Hypomethylated
CYFIP2	Hypomethylated
DRAM1	Hypermethylated
DSPP	Hypomethylated
ERC1	Hypomethylated
FAT4	Hypomethylated
FGF11	Hypermethylated
HIVEP3	Hypomethylated
HNF1A	Hypomethylated
HOXB3	Hypermethylated
INPPL1	Hypomethylated
INSRR	Hypermethylated
IQSEC1	Hypermethylated
LRRK2	Hypomethylated
NBEA	Hypermethylated
OTOF	Hypermethylated
PACRGL	Hypomethylated
PCDH8	Hypermethylated
PDZD2	Hypermethylated
PTMS	Hypomethylated
RGS7	Hypomethylated
RYR2	Hypomethylated
SVIL	Hypomethylated
TJP1	Hypermethylated
TMCC1	Hypomethylated
TMEM26	Hypomethylated
TRIB1	Hypomethylated
USP3	Hypermethylated

Supplementary Table 4.5| Genes identified mutated and differentially methylated at the prom oter level in NSCLCs.

6 common elements in "CNA" and "DMR":	32 common elements in "DMR" and "Mutation":	13 common elements in "CNA" and "Mutation":	2 common elements in "CNA", "DMR" and
MIR1204	*PLK4*	*MKRN3*	*EGFR*
IRX4	*CD1E*	*MYC*	*SLC12A7*
S100A11	*TENM2*	*BRD9*	
PIP5K1A	*GHR*	*CLPTM1L*	
TXNIP	*XIRP2*	*ARNT*	
NKX2-8	*DEPDC1*	*ITGA10*	
	TNN	*PIAS3*	
	RGS7	*CELF3*	
	RYR2	*ZNF229*	
	THSD7B	*ZNF467*	
	MET	*ANK1*	
	PLXNA4	*KAT6A*	
	HTR5A	*KDR*	
	HYAL4		
	SSPO		
	OR4D9		
	MMP13		
	PLCB1		
	NWD1		
	UBA2		
	SBNO2		
	ZNF217		
	DNAH9		
	POU4F2		
	RD3		
	PAX3		
	HOXD3		
	SLC18A3		
	SDK1		
	CDKN2A		
	PPP2R1B		
	DDHD1		

Supplementary Figure 4.6| Mutation and Copy Number count in *SLC12A7* detected in 4,767 Lung Cancer samples from TCGA obtained from cBioPortal for Cancer Genomics.



Chapter 5

Supplementary Table 5.1| Dataset demographics for sequence (WES) read depth analysis.

	n Males	Histology *n* LUAD/LUSC (NR)	Age μ(SD)	Tumour stage *n* IA/IB/II/IIA/IIB/ III/IIIA/IIIB/IV (NR)	Smoking *n* NS/EX/CS (NR)	Deceased *n* T/F (NR)
Low Y expression	6	2/4 (0)	66 (3.35)	0/2/0/2/1/0/1/0/0 (0)	0/2/4 (0)	4/2 (0)
Non-low Y expression	9	5/4 (0)	67.67 (8.70)	1/1/0/2/2/0/3/0/0 (0)	0/6/3 (0)	3/6 (0)
Overall	15	7/8 (0)	67 (6.93)	1/3/0/4/3/0/4/0/0 (0)	0/8/7 (0)	7/8 (0)

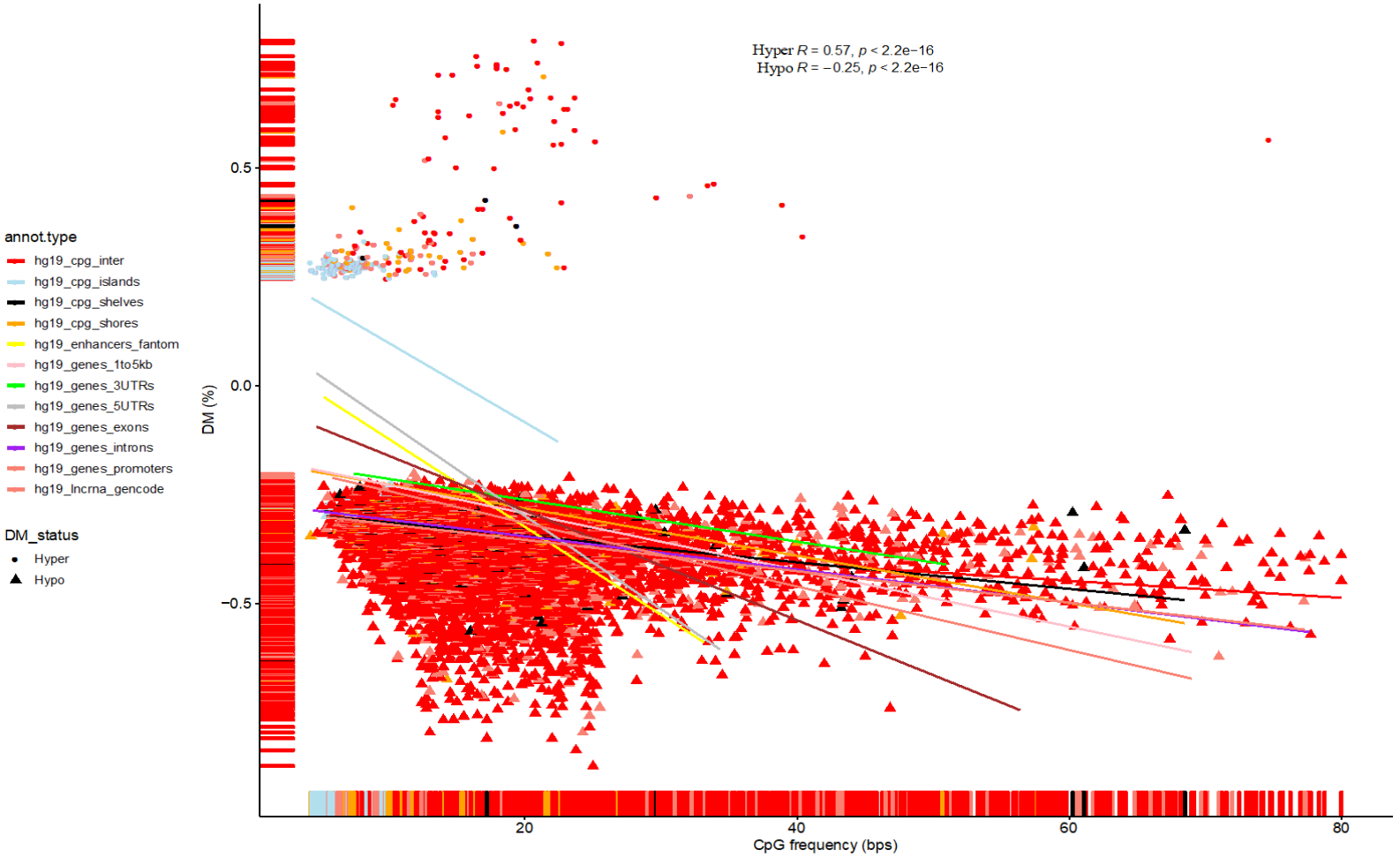
Abbreviations: LUAD (lung adenocarcinoma); LUSC (lung squamous cell carcinoma); NS (non-smoker); EX (ex-smoker); CR (current smoker); NR (not recorded); T (true); F (false); sd (standard deviation); WES (whole exome sequencing).

Supplementary Table 5.2| Dataset demographics for methylation (WGBS) analysis.

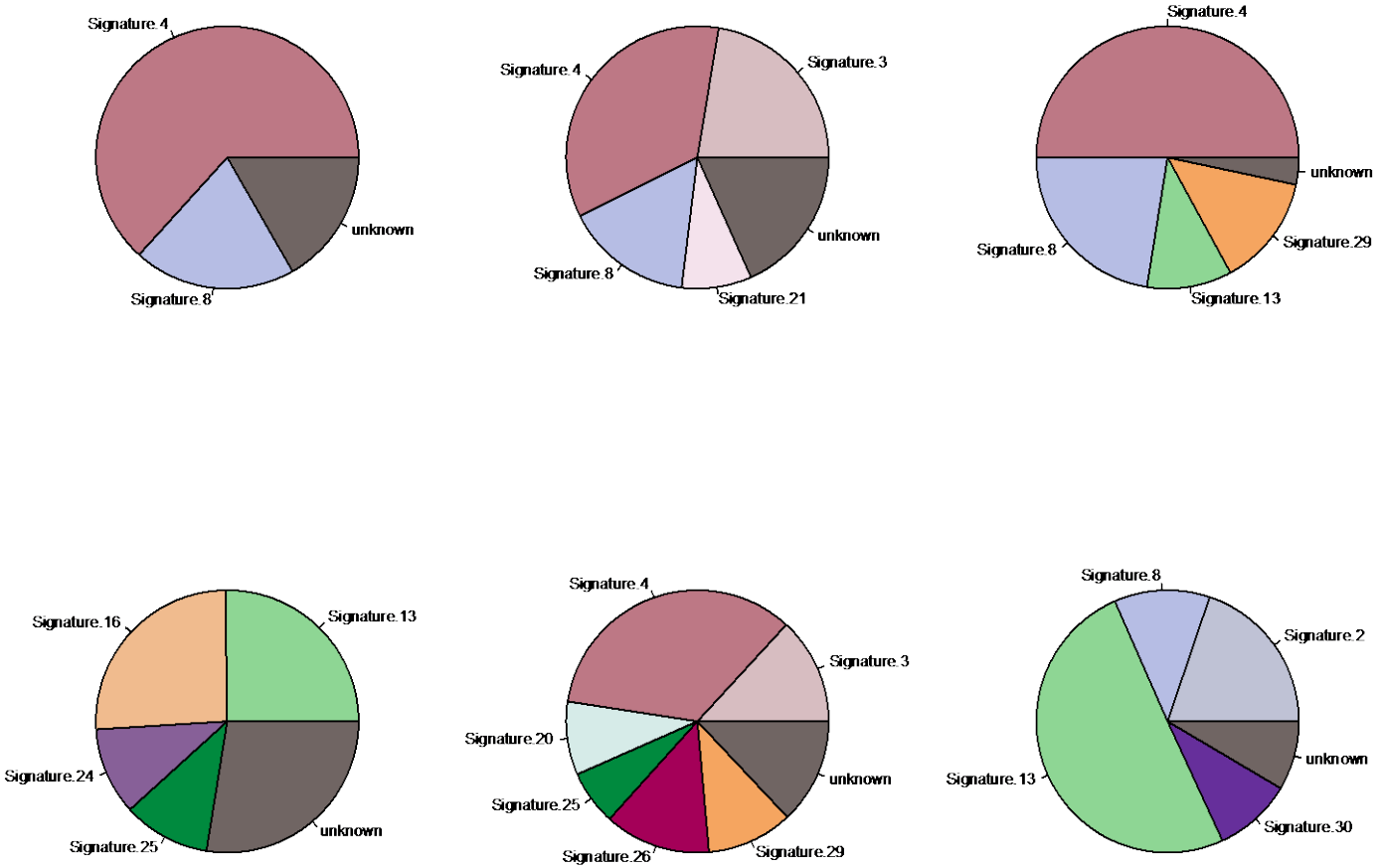
	***n* Males**	**Histology *n* LUAD/LUSC (NR)**	**Age μ (sd)**	**Tumour stage *n* IA/IB/II/IIA/IIB/ III/IIIA/IIIB/IV (NR)**	**Smoking *n* NS/EX/CS (NR)**	**Deceased n T/F (NR)**
Low Y expression	17	11/6 (0)	69.29 (6.43)	5/3/0/3/1/0/4/0/1 (0)	1/10/6 (0)	10/7 (0)
Non-low Y expression	5	5/0 (0)	71.20 (8.35)	2/1/0/0/0/0/1/0/1 (0)	0/3/2 (0)	1/4 (0)
Overall	22	16/6 (0)	69.73 (6.74)	7/4/0/3/1/0/5/0/2 (0)	1/13/8 (0)	11/11 (0)

Abbreviations: LUAD (lung adenocarcinoma); LUSC (lung squamous cell carcinoma); NS (non-smoker); EX (ex-smoker); CR (current smoker); NR (not recorded); T (true); F (false); sd (standard deviation).

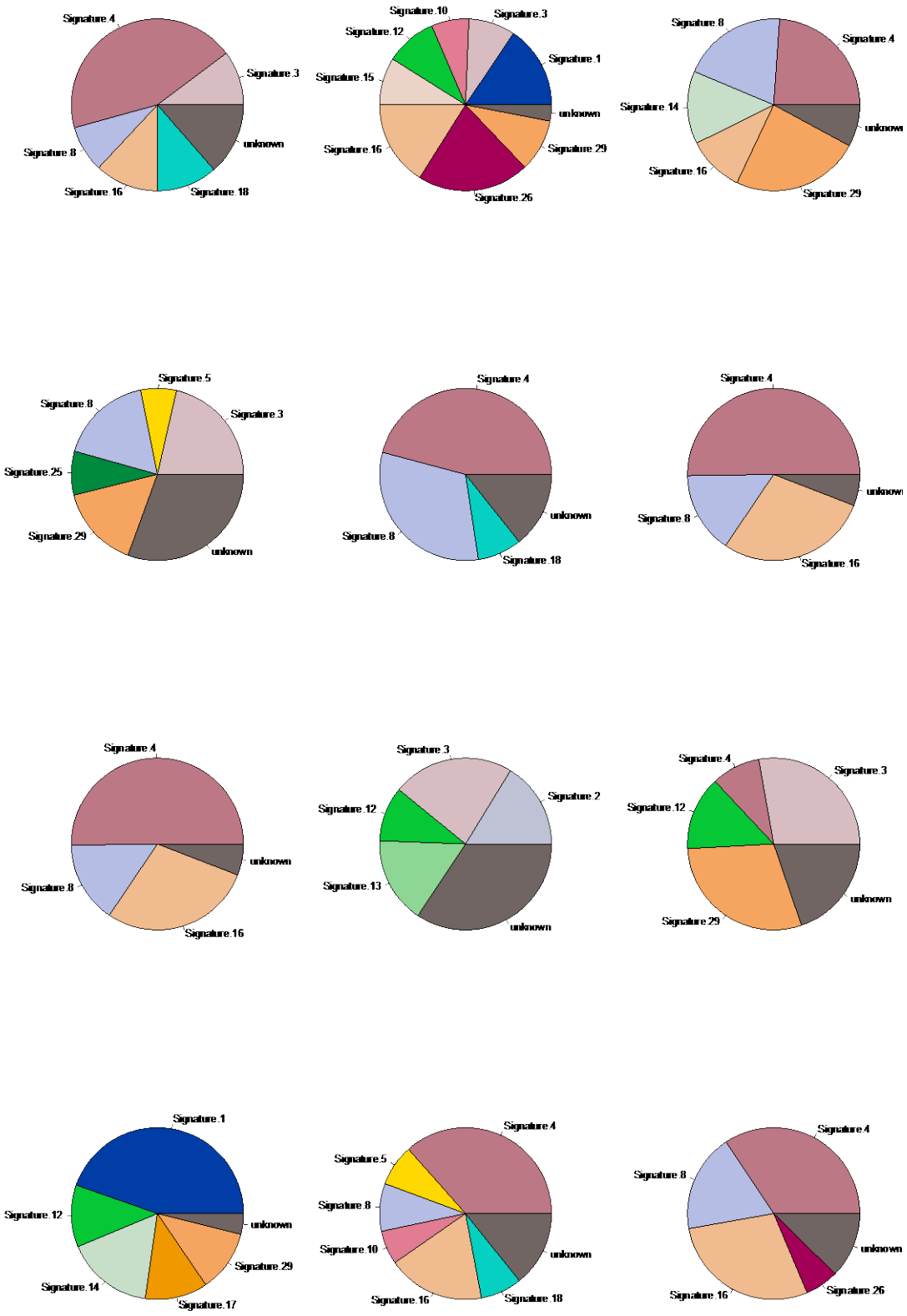
Supplementary Figure 5.1| Correlation between CpG frequency and Differential Methylation (DM) percentage detected in tumours with deficient Y-chromosome expression when compared against their matched normal tissues. Different genomic categories (see legend colour) are shown for hypermethylated (circles) and hypomethylated (in triangles) CpGs.

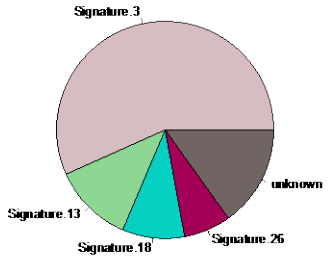


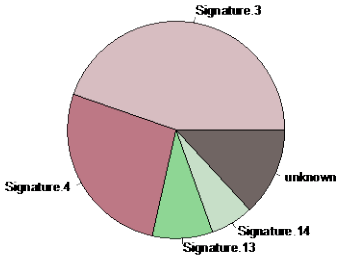
Supplementary Figure 5.2| Weights of the mutational signatures in LYE tumours. COSMIC Mutational Signatures (CMS) were identified with decosntructSigs R package with COSMIC mutational signatures version 2 by using the exome2genome normalization method.

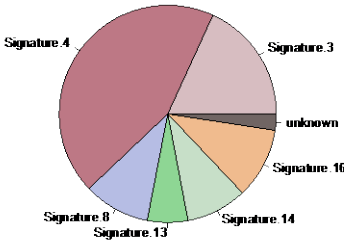


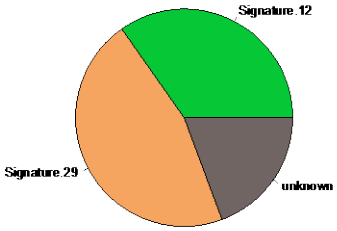
Supplementary Figure 5.3| Weights of the mutational signatures in non-LYE tumours. CMSs were identified with decosntructSigs R package with COSMIC mutational signatures version 2 by using the exome2genome normalization method.



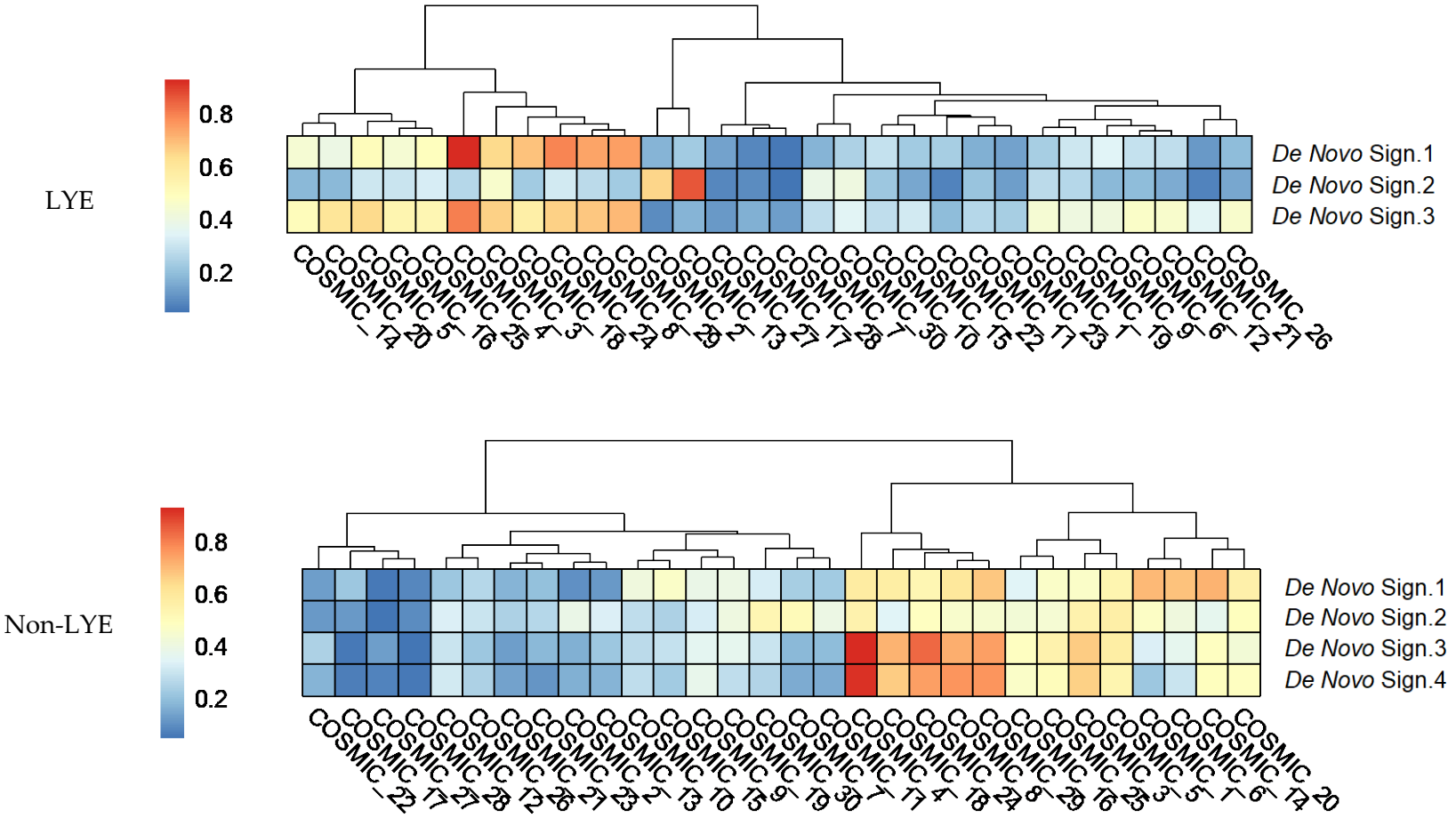

Signature.3
unknown
Signature.13
Signature.18
Signature.26


Signature.3
Signature.4
unknown
Signature.14
Signature.13


Signature.4
Signature.3
unknown
Signature.16
Signature.8
Signature.13
Signature.14


Signature.12
Signature.29
unknown

Supplementary Figure 5.4| *De novo* Mutational signatures identified in LYE and Non-LYE tumours. De novo signatures are shown on the *Y* axis and were compared against COSMIC mutational signatures (*X* axis) identified by Alexandrov and colleagues[49].



Chapter 6

Supplementary Table 6.1. Clinical data for L-CD patients

Sample ID	*Gender (f/m)*	*Age years*	*Stage*	*Deceased (T/F)*	*Survival (months)*	*Smoking category*	*Tumour percentage*	*BMI*	*Molecular group*	*Histology*	*Location*	*Spindle morphology*	*cellLymph invasion*	*Emphysema presence*	*Mitotic rate per mm2*
sample_001	f	76	IB	F	71	NS		28	PanC	TC	Central	No	No	Yes	1
sample_002	f	72	IIB	F	69	NS		33	PanC	TC	Central	Prevalent	Yes	Yes	1
sample_003	m	63	IA2	F	66	ES	80	26	PanC	TC	Central	No	No	No	1
sample_004	m	67	IB	F	61	ES	90	33	PanC	TC	Central	No	Yes	No	1
sample_005	f	39	IA	F	77	NS		25	PanC	TC	Central	Prevalent	No	Yes	1
sample_006	f	54	IIIB	F	29	CS	90	22	PanC	TC	Central	No	Yes	Yes	2
sample_007	m	28	IIIA				95		PanC	TC	Central	Focal	Yes	No	1
sample_008	f	30	IIA	F	45	NS	80	18	PanC	TC	Central	No	No	No	1
sample_009	f	82	IA	T	41	NS		26	NeU	AC	Peripheral	Prevalent	Yes	Yes	1
sample_010	m	54	IA	F	79	NS		31	NeU	TC	Peripheral	Prevalent	No	No	1
sample_011	f	73	IA	F	63	ES	90		NeU	TC	Peripheral	Focal	No	No	1
sample_012	f	71	IIIA	T	39	NS	80	30	NeU	AC	Central	Focal	Yes	No	7
sample_013	f	61	IIIA	F	77	ES		30	NeU	TC	Central	Focal	Yes	No	1
sample_014	f	66	IA3	F	43	NS	90	22	NeU	AC	Peripheral	Focal	Yes	Yes	7
sample_015	f	75	IIB	F	47	NS	90		NeU	TC	Peripheral	Focal	Yes	No	1

Supplementary Figure 6.1: Boxplots of the 25 immune markers significantly differentially expressed between L-CD groups (*adj.P*≤0.05).