Pan-cancer study of transcriptional responses to oncogenic somatic mutations

Arghavan Ashouri

Department of Medical Biochemistry and Cell Biology

Institute of Biomedicine

Sahlgrenska Academy, University of Gothenburg

UNIVERSITY OF GOTHENBURG

Gothenburg 2021

Cover illustration: Word-cloud illustrations referring to the three studies included in this thesis.

By Arghavan Ashouri. Figures generated in Wordclouds.com.

Pan-cancer study of transcriptional responses to oncogenic somatic mutations

arghavan.ashouri@gu.se

ISBN 978-91-8009-318-7 (PRINT)
ISBN 978-91-8009-319-4 (PDF)
http://hdl.handle.net/2077/68071

Printed in Borås, Sweden 2021
Printed by Stema Specialtryck AB

SVANENMÄRKET
Trycksak
3041 0234

To Liam and Lavin,

the real achievements of my PhD years…

If you do not change direction,

you may end up where you are heading.

- Lao Tzu

Pan-cancer study of transcriptional responses to oncogenic somatic mutations

Arghavan Ashouri

Department of Medical Biochemistry and Cell Biology,
Institute of Biomedicine,
Sahlgrenska Academy, University of Gothenburg
Gothenburg, Sweden

ABSTRACT

Cancer cells typically carry acquired somatic mutations in key cancer driver genes, which can be identified on the basis of recurrence in cancer cohorts. Such mutations may cause aberrant protein activity and altered gene expression in the nucleus, driving the cell toward a cancerous phenotype. Understanding the transcriptional consequences of cancer driver mutations can guide us towards better diagnosis and prognosis.

While the phenotypic effects of driver mutations are typically mediated by protein-coding genes, recent studies also describe roles for long non-coding RNAs as effector molecules in oncogenic pathways. LncRNAs are long transcripts that are transcribed and processed similarly to coding messenger RNAs but lack coding capacity. However, a systematic exploration of lncRNAs as potential mediators of oncogenic mutational events in cancer has been missing. In the first study, we established a methodology to uncover associations between key driver mutations and transcriptional alterations in lncRNAs, using publicly available genomic data from thousands of tumours and 19 different cancer types from The Cancer Genome Atlas (TCGA). Based on this, we found many putative lncRNA effectors, some of which could be validated in terms of transcriptional responsiveness using previously published data. We also designed experiments to investigate the function of NRF2 (also known as nuclear factor erythroid 2–related factor 2 (NFE2L2)) responsive lncRNAs in cancer cell lines. NRF2 is a master regulator of the cellular oxidative stress response, and it is frequently mutated in several cancer types. Using small interfering RNAs (siRNAs) targeting *NRF2*, followed by deep RNA sequencing we validated the expression changes of these lncRNAs downstream of NRF2. Furthermore, using ChIP-PCR (chromatin immunoprecipitation followed by PCR) we could confirm direct binding of NRF2 to the promoter of selected lncRNAs identified as NRF2-responsive in our screen. In summary, this study provides a comprehensive overview of lncRNA transcriptional alterations in relation to key driver mutational events in human cancers.

In the second study, we have focused further on the transcription factor NRF2, which is a key driver gene in lung cancer and several other cancer types. Despite this, the target repertoire of NRF2 remains incompletely characterized. To solve this, we performed NRF2 ChIP-seq (ChIP followed by high-throughput sequencing) using two different antibodies against NRF2 in two lung cancer cell lines, as well as *NRF2* knock-down with siRNA transfections plus control siRNAs followed by deep RNA sequencing, to discover genes with expression alteration upon blocking *NRF2*. Integrative analysis of the ChIP and RNA sequencing data has resulted in the most detailed characterization of the NRF2 targetome to date. This analysis confirmed most known targets and also revealed several new targets of NRF2. The resulting dataset constitutes an important resource that can facilitate our understanding of NRF2 and its complex role in cancer.

The third study describes transcriptional and phosphoproteomic responses following treatment with marketed ALK tyrosine kinase inhibitors (TKIs). *ALK* is a major driver gene in neuroblastoma and other cancers with frequent somatic mutations both in primary and relapsed tumours. Mapping of signaling and transcriptional events downstream of ALK revealed relevant biomarkers and signaling networks, such as ETS family transcription factors and the MAPK phosphatase *DUSP4*, some of which are potential therapeutic targets.

In conclusion, this thesis links key driver events in human cancer to specific transcriptional effects, providing insights into oncogenic signaling and clues to new treatments.

Keywords: Cancer, driver mutations, lncRNAs, NRF2, transcriptome, ALK.

ISBN 978-91-8009-318-7 (PRINT)

ISBN 978-91-8009-319-4 (PDF)

http://hdl.handle.net/2077/68071

SAMMANFATTNING PÅ SVENSKA

Cancerceller bär vanligtvis på förvärvade somatiska mutationer i viktiga cancerdrivande gener, som kan identifieras baserat på hur vanligt förekommande de är i cancerkohorter. Sådana mutationer kan orsaka avvikande proteinaktivitet och förändrat genuttryck i cellkärnan, vilket driver cellen mot en cancerfenotyp. Att förstå cancerdrivande mutationers påverkan på genuttryck kan vägleda oss mot bättre diagnostik och prognos.

Medan de fenotypiska effekterna av cancerdrivande mutationer vanligtvis medieras av proteinkodande gener, beskriver nya studier också roller för långa icke-kodande RNA (lncRNA) som effektormolekyler i onkogena signalvägar. LncRNA är långa transkript som transkriberas och bearbetas på samma sätt som kodande budbärar-RNA (mRNA) men saknar kodande kapacitet. En systematisk undersökning av lncRNA som potentiella förmedlare av onkogena mutationshändelser i cancer saknas dock. I den första studien etablerade vi en metod för att hitta samband mellan viktiga cancerdrivande mutationer och transkriptionsförändringar i lncRNA med hjälp av publikt tillgängliga genomdata från tusentals tumörer och 19 olika cancertyper från The Cancer Genome Atlas (TCGA). Baserat på detta hittade vi många möjliga lncRNA-effektormolekyler, varav några kunde valideras med avseende på transkriptionsrespons med hjälp av tidigare publicerade data. Vi designade också experiment för att undersöka funktionen av NRF2-responsiva lncRNAs i cancercellinjer. NRF2 (även känd som nuclear factor erythroid 2-related factor 2, NFE2L2) har en central roll i den cellulära oxidativa stressresponsen, och muteras ofta i flera cancertyper. Med hjälp av små interfererande RNA (siRNA) riktade mot *NRF2*, följt av RNA-sekvensering med högt läsdjup och bred täckning, validerade vi expressionsförändringarna av dessa lncRNAs nedströms NRF2. Vidare kunde vi med hjälp av ChIP-PCR (kromatinimmunprecipitation följt av PCR) bekräfta direkt bindning av NRF2 till promotorn för utvalda lncRNA som identifierats som NRF2-responsiva i vår screening. Sammanfattningsvis ger denna studie en omfattande översikt av transkriptionella förändringar av lncRNA i relation till centrala cancerdrivande mutationshändelser i cancer hos människan.

I den andra studien har vi fokuserat ytterligare på transkriptionsfaktorn NRF2, som är en viktig cancerdrivande gen i lungcancer och flera andra cancertyper. Trots dess viktiga roll är karaktäriseringen av NRF2:s målmolekyler ofullständig. För att undersöka detta utförde vi NRF2 ChIP-seq (ChIP följt av sekvensering) med två olika antikroppar mot NRF2 i två lungcancercellinjer, samt NRF2-hämning med siRNA följt av RNA sekvensering för att upptäcka gener med expressionsförändring vid blockering av NRF2. Denna integrerade analys av ChIP- och RNA-sekvenseringsdata har resulterat i den mest detaljerade karakteriseringen av NRF2s målmolekyler hittills. Denna analys bekräftade de flesta redan kända målen och avslöjade också flera nya mål för NRF2. Det resulterande datasetet utgör en viktig resurs som kan underlätta vår förståelse av NRF2 och dess komplexa roll i cancer.

Den tredje studien beskriver transkriptions- och fosfoproteomiska förändringar efter behandling med marknadsförda ALK-tyrosinkinashämmare. *ALK* är en viktig cancerdrivande gen i neuroblastom och andra cancerformer med frekventa somatiska mutationer både i primära och återkommande tumörer. Kartläggningen av signalering och transkriptionshändelser nedströms ALK avslöjade relevanta biomarkörer och signalnätverk, såsom ETS-familjetranskriptionsfaktorer och MAPK-fosfataset *DUSP4*, varav några är potentiella terapeutiska mål.

Sammanfattningsvis länkar denna avhandling viktiga cancerdrivande mutationshändelser i mänsklig cancer till specifika genuttrycksförändringar, vilket ger insikt i onkogena signalvägar och ledtrådar till nya behandlingar.

Nyckelord: Cancer, cancerdrivande mutationer, lncRNAs, NRF2, transkriptom, ALK.

LIST OF PAPERS

This thesis is based on the following studies, referred to in the text by their Roman numerals.

I. Pan-cancer transcriptomic analysis associates long non-coding RNAs with key mutational driver events.
Ashouri A, Sayin VI, Van den Eynden J, Singh SX, Papagiannakopoulos T, Larsson E.
Nat Commun. 2016 Oct 25;7:13197.

II. Genome-wide characterization of the NRF2 targetome.
Ashouri A, Larsson E.
Manuscript

III. Phosphoproteome and gene expression profiling of ALK inhibition in neuroblastoma cell lines reveals conserved oncogenic pathways.
Van den Eynden J, Umapathy G, **Ashouri A**, Cervantes-Madrid D, Szydzik J, Ruuth K, Koster J, Larsson E, Guan J, Palmer RH, Hallberg B.
Sci Signal. 2018 Nov 20;11(557)

PAPERS NOT INCLUDED IN THIS THESIS

1. Morud J, **Ashouri A**, Larsson E, Ericson M, Söderpalm B. Transcriptional profiling of the rat nucleus accumbens after modest or high alcohol exposure.
   PLoS One. 2017 Jul 17;12(7)

2. Jacobson T, Priya S, Sharma SK, Andersson S, Jakobsson S, Tanghe R, **Ashouri A**, Rauch S, Goloubinoff P, Christen P, Tamás MJ. Cadmium Causes Misfolding and Aggregation of Cytosolic Proteins in Yeast.
   Mol Cell Biol. 2017 Aug 11;37(17)

3. Schneider E, Staffas A, Röhner L, Malmberg ED, **Ashouri A**, Krowiorz K, Pochert N, Miller C, Wei SY, Arabanian L, Buske C, Döhner H, Bullinger L, Fogelstrand L, Heuser M, Döhner K, Xiang P, Ruschmann J, Petriv OI, Heravi-Moussavi A, Hansen CL, Hirst M, Humphries RK, Rouhi A, Palmqvist L, Kuchenbauer F. Micro-ribonucleic acid-155 is a direct target of Meis1, but not a driver in acute myeloid leukemia.
   Haematologica. 2018 Feb;103(2)

4. Schneider E, Pochert N, Ruess C, MacPhee L, Escano L, Miller C, Krowiorz K, Delsing Malmberg E, Heravi-Moussavi A, Lorzadeh A, **Ashouri A**, Grasedieck S, Sperb N, Kumar Kopparapu P, Iben S, Staffas A, Xiang P, Rösler R, Kanduri M, Larsson E, Fogelstrand L, Döhner H, Döhner K, Wiese S, Hirst M, Keith Humphries R, Palmqvist L, Kuchenbauer F, Rouhi A. MicroRNA-708 is a novel regulator of the Hoxa9 program in myeloid cells.
   Leukemia. 2020 May;34(5)

TABLE OF CONTENT

ABBREVIATIONS

A	Adenine
ACC	Adrenocortical carcinoma
AHR	Aryl hydrocarbon receptor
Ala	Alanine
ALCL	Anaplastic large-cell lymphoma
ALK	Anaplastic lymphoma kinase
APC	Adenomatous polyposis coli
AREs	Antioxidant response elements
Arg	Arginine
ASO	Antisense oligonucleotide
ATP	Adenosine triphosphate
BLCA	Bladder Urothelial Carcinoma
BRCA	Breast invasive carcinoma
bp	Base pair
bZIP	Basic leucine zipper
C	Cytosine
cDNA	Complementary deoxyribonucleic acid
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
ChIP	Chromatin immunoprecipitation
ChIP-Seq	Chromatin immunoprecipitation sequencing
CHOL	Cholangiocarcinoma
CNTL	Controls
COAD	Colon adenocarcinoma
CUL3	Cullin 3
DDR1/DDR2	Discoidin domain receptor family, member 1 and 2
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
DNA	Deoxyribonucleic acid

EGFR	Epidermal Growth Factor Receptor
EMA	European Medicines Agency
ESCA	Esophageal carcinoma
ESTs	Expressed sequence tags
FDA	US Food and Drug Administration
G	Guanine
GBM	Glioblastoma multiforme
GCLM, GCLC	Glutamate-cysteine ligase, modifier and catalytic subunits
GSEA	Gene set enrichment analysis
GSH	Reduced glutathione
GSSG	Oxidized glutathione
GTP	Guanosine triphosphate
HGP	The Human Genome Project
HMOX1	Heme oxygenase-1
HNSC	Head and Neck squamous cell carcinoma
HOTAIR	HOX transcript antisense RNA
HSF1	Heat shock factor 1
HSP	Heat shock protein
ICGC	International Cancer Genome Consortium
Ile	Isoleucine
KEAP1	Kelch-like ECH associated protein 1
KICH	Kidney Chromophobe
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
LAML	Acute Myeloid Leukemia
LCML	Chronic Myelogenous Leukemia
LC-MS/MS	Liquid chromatography-tandem mass spectrometry
Leu	Leucine
LGG	Brain Lower Grade Glioma
LIHC	Liver hepatocellular carcinoma
lincRNA-p21	Large intergenic non-coding RNA p21

lncRNA	Long non-coding RNA
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MALAT1	Metastasis Associated Lung Adenocarcinoma Transcript 1
MAPK	Mitogen-activated protein kinase
MDK	Midkine
MESO	Mesothelioma
miRNA	MicroRNA
mRNA	Messenger RNA
MSigDB	Molecular signatures database
NADPH	Nicotinamide adenine dinucleotide phosphate
NCI	National Cancer Institute
ncRNA	Non-coding RNA
NFE2L2	Nuclear factor erythroid-derived 2-like 2 gene
NF-YA	Nuclear transcription factor Y subunit alpha
NGS	Next Generation Sequencing
NHGRI	National Human Genome Research Institute
NPM1	Nucleophosmin gene
NRF2	Nuclear factor erythroid 2–related factor 2
NQO1	NAD(P)H quinone dehydrogenase 1
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PANDA	P21-associated ncRNA DNA damage-activated
PCA3	Prostate cancer antigen 3
PCAWG	Pan-Cancer Analysis of Whole Genomes Project
PCGEM1	Prostate-specific transcript 1
PCPG	Pheochromocytoma and Paraganglioma
PCR	Polymerase chain reaction
PDGF	Platelet-derived growth factor
Phe	Phenylalanine
PI3K	Phosphoinositide 3-kinase

PPI	Protein-protein interactions
PRAD	Prostate adenocarcinoma
PTN	Pleiotrophin
READ	Rectum adenocarcinoma
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
rRNA	Ribosomal RNA
RNA Poly II	RNA polymerase II
ROS	Reactive oxygen species
RPPA	Reverse phase protein array
RSV	Rous sarcoma virus
RT-PCR / q-PCR	Real time quantitative PCR
SAGE	Serial analysis of gene expression
SARC	Sarcoma
Ser	Serine
SKCM	Skin Cutaneous Melanoma
SMRT	Single molecule real time sequencing
SMT	Somatic mutation theory
SNP	Single nucleotide polymorphism
snRNA	Small nuclear RNA
snoRNA	Small nucleolar RNA
SODs	Superoxide dismutases
SOLiD	Sequencing by oligonucleotide ligation and detection
STAD	Stomach adenocarcinoma
T	Thymidine
TCGA	The Cancer Genome Atlas
TERC	Telomerase RNA component
TF	Transcription factor
TGCT	Testicular Germ Cell Tumors
THCA	Thyroid carcinoma

THYM	Thymoma
TKI	Tyrosine kinase inhibitor
tRNA	Transfer RNA
Trp	Tryptophan
TOFT	Tissue organization field theory
TXN	Thioredoxin
TXNRD1	Thioredoxin reductase 1
Tyr	Tyrosine
U	Uracil
UCEC	Uterine Corpus Endometrial Carcinoma
UCS	Uterine Carcinosarcoma
UV	Ultraviolet
UVM	Uveal Melanoma
WGS	Whole genome sequencing
XIST	X-inactive specific transcript
XRE	Xenobiotic response element
ZMW	Zero-mode waveguide

1 BACKGROUND

Cancer is a genetic disease caused by alterations in cellular DNA, leading to abnormal intracellular signaling and gene expression. The main mechanism of gene alteration is mutation, which in general terms is an error in the DNA that inactivates the gene or changes its normal function. Mutations can be inherited from one's parents if they have been present in the germ cells (egg or sperm); also called germline mutations. Another type of mutation, namely somatic mutations, are accumulated in the body after the formation of the zygote. These mutations can be a consequence of either cell-intrinsic mutagenic processes or external mutagen exposure. There are also other ways in which gene expression can be altered by changes in DNA; these alterations are called epigenetic modifications and occur through addition or removal of different chemical groups on DNA (Hanahan and Weinberg, 2011).

Cancer cells typically carry acquired somatic mutations in key cancer driver genes, which are genes that have the ability to drive tumorigenesis. Identification of cancer drivers has been the main goal of cancer research since the discovery of the DNA and genes during the last century (Martínez-Jiménez, et al., 2020). The downstream effect of alterations in driver genes can vary a lot, depending on the gene's function and the type of mutation occurring. Tumor progression is driven by Darwinian-like somatic evolution, meaning that although the mutations in the cell are often blind, if they are advantageous for the cell in the tumour microenvironment, positive selection will select for these cells, resulting in clonal expansion of these cells (Greaves and Maley, 2012). There are several mechanisms which can be beneficial for tumour growth and survival, for example converting a proto-oncogene, which are normal cellular genes that are usually involved in regulation of cell growth or proliferation, into an oncogene. Another mechanism that helps cancer cell growth is by stopping the function of tumour suppressor genes, which are genes that have a role in prohibiting abnormal cell proliferation (Lodish H, 2000). When one or several such changes is present, signaling pathways and gene expression patterns might be altered, such that the cell is driven toward an oncogenic phenotype.

Gaining insight into altered pathways in cancer cells that benefit the cells towards tumourigenesis is one of the biggest challenges in the field of cancer genomics, and it is fundamental for translating our knowledge of the cancer genome into precision cancer medicine (Martínez-Jiménez, et al., 2020).

1.1. THE HUMAN GENOME

The haploid human genome comprises approximately $3.2x10^9$ (3.2 billion) nucleotides of DNA divided into 22 autosomes and two sex chromosomes, X and Y. DNA was first discovered in 1869 by the Swiss physician Friedrich Miescher, and since he had isolated it from the cells' nuclei, he named it nuclein, a name preserved in today's designation deoxyribonucleic acid (DNA) (Dahm, 2005). Later in 1953, James Watson and Francis Crick characterized the structure of DNA, based on an important contribution from Rosalind Franklin. But the genetic code was not deciphered until long after by the 1968 Nobel Laureates in Medicine or Physiology: Robert W. Holley, Har Gobind Khorana and Marshall W. Nirenberg. They discovered that for DNA to be translated to proteins, a combination of three bases, also called ‘codon’, is needed to specify which amino acid is to be added at each step for generating the full protein. With the invention of Polymerase chain reaction (PCR) by the American biochemist Kary Mullis, studying small amounts of DNA in detail became feasible (Saiki, et al., 1985). All these discoveries led to the most recent breakthrough in the history of DNA research: the very nearly complete sequence of the human genome in 2001 (Lander, et al., 2001; Venter, et al., 2001). The findings from The Human Genome Project were a starting point for interesting research with the aim to complete our knowledge about the human genome.

1.1.1. FROM DNA TO RNA AND MORE

Since the discovery of the double helical structure of DNA in early 1950s, and the knowledge that only four different nucleotide subunits make up the DNA, we have come a long way in understanding how cells read this information and use it to grow, replicate and differentiate into different cell types. A gene is a unit of hereditary information within DNA, varying in length between a few hundred bases to about two million bases, on a fixed location (locus) on a chromosome, which accomplishes its effect by directing the synthesis of proteins or functional RNAs. Genes are typically present in two copies, due to the genome being diploid (two of each chromosome) in most cells (Daxinger and Whitelaw, 2012). A gene often consists of a promoter region and alternating regions of exons (coding sequences) and introns (non-coding sequences). Transcription and translation are the means by which the cells read the DNA into functional molecules.

Briefly and simplified, the initial step is for the cell to transcribe the DNA into RNA, which can be functional non-coding RNAs (ncRNAs), or be an intermediate template messenger RNA (mRNA), that is translated into proteins providing a function (Figure 1) (Eldra Solomon, 2019).

In Figure 1, the process of transcription and translation is presented in a simplified schematic graphic (Eldra Solomon, 2019). Briefly, the process of transcription contains the following general steps:

- RNA polymerase (in this case RNA polymerase II, responsible for transcription of mRNA and ncRNAs) binds to the promoter region of a given gene, together with one or several transcription factors (TFs).
- A transcription bubble (complex of various molecules) is generated that separates the DNA helix strands by breaking hydrogen bonds.
- RNA polymerase II adds RNA nucleotides (adenine (A), cytosine (C), guanine (G), or uracil (U)) forming a pre-mRNA (if the gene is a protein coding) or a pre-lncRNA (if the gene is a ncRNA).
- The mRNA or ncRNA will be further processed including capping, polyadenylation and splicing and thereafter either remain in the nucleus or be transported into the cytoplasm.

- Some ncRNAs can fold into secondary and tertiary structures, which is their functional and conserved forms (Johnsson, et al., 2014).

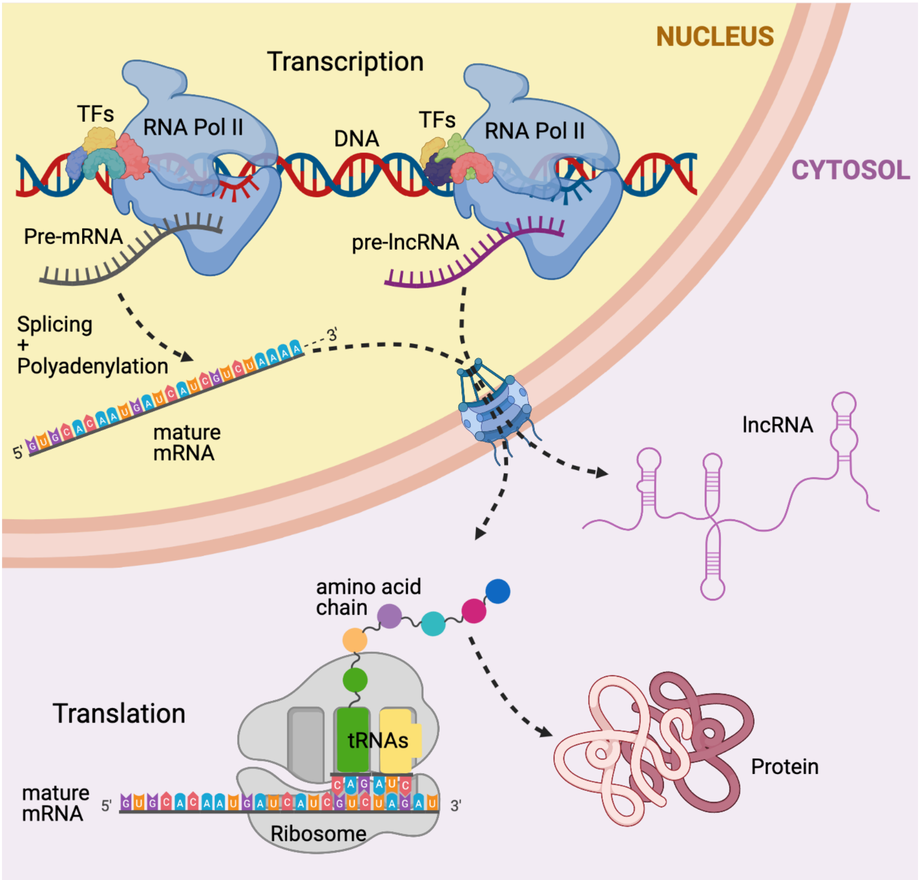

Figure 1. Transcription and Translation of protein coding and long non-coding RNA genes. Figure generated with BioRender.com.

The process of translation contains three main steps:

- Initiation: The first tRNA is attached at the start codon after the ribosome has been assembled around the mRNA.
- Elongation: Through accommodation, transpeptidation and translocation an amino acid chain is gradually built up by the ribosome and various tRNA molecules.

- Termination: Eventually when a stop codon is reached the polypeptide is released from the ribosome and thereafter posttranslational modifications can occur.

Over recent decades a rapid development of high-throughput sequencing technologies has occurred which led to a vast amount of research being performed to understand the nature and role of genes and ncRNAs. There are different types of genes that code specifically for RNAs which have a function without involvement in protein translation (Birney, et al., 2007). These ncRNAs can be small RNAs, such as transfer RNA (tRNA), microRNAs (miRNAs), small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), which are enzymatic RNA molecules, or larger ncRNA such as long non-coding RNAs (lncRNAs) and ribosomal RNAs (rRNAs) (Eddy, 2001). This shows that different types of genes coding for RNA molecules in various forms have a critical role in performing functions in cells by aiding in synthesizing, regulating and processing proteins in various ways.

While 75% of the human genomic DNA is believed to be transcribed, the exon sequences in mature mRNA represent only ~2% of the genome, and the rest are the introns and other non-coding RNAs (Li and Liu, 2019). Most of these mRNAs are polyadenylated, which is a process that produces mature mRNA by adding a poly(A) tail consist of adenosine bases to the end of the mRNA (Djebali, et al., 2012). In general, lower levels of gene expression has been measured in ncRNAs compared to coding genes, however, in certain tissues the amount of ncRNAs can be abundant (Liu, et al., 2016).

1.2. CARCINOGENESIS

Theories of carcinogenesis

The somatic mutation theory (SMT) hypothesis was originated in 1914 by Theodor Boveri, who proposed that a combination of chromosomal defects could result in cancer. Later on, in 1953, a Finnish architect and urban planner, Carl O. Nordling, summarized his findings from cancer mortality reports from several countries and suggested the SMT of cancer (Nordling, 1953). The essence of SMT are as follows:

- Cancer is originated form a single somatic cell that has accumulated multiple DNA mutations.
- These mutations are caused in genes related to proliferation and the cell cycle.
- The default state of cell proliferation is quiescence (Sonnenschein and Soto, 2020).

In summary, SMT states that cancer initiates if a mutation gives a growth advantage to a cell, and it is followed by clonal expansion of those cells. Over the last century, this theory has grown into a more complete picture of how cancer occurs in the cells, and the reported number of mutations associated with tumors has increased dramatically. Certain cancer cells contain tens of non-synonymous mutations, and others can harbor a few thousand. Perhaps, the most widespread explanation is that only a small number of these mutations are drivers, and the remaining passenger mutations do not play a role in carcinogenesis of these cells (Baker, 2015).

At the end of 20th century there were still many questions unanswered regarding tumourigenesis, hence researchers developed a new theory that was based more on the tissue organization than the cells itself. The tissue organization field theory (TOFT) was first introduced in 1999 (Sonnenschein, 1999). According to TOFT, carcinogenesis is a problem of tissue organization caused by carcinogens such as environmental factors that destroy the normal tissue architecture consequently disrupting normal cell functions, and the DNA mutations are the consequence of these effects, and not the cause. SMT considered cancer to be an irreversible disease, whereas TOFT introduce it as reversable and curable (Rosenfeld, 2013).

Even though our knowledge about the exact mechanism in which the cancer initiates is still incomplete, it is certain that neither the one single event in a cell (as proposed by SMT), nor the persistent damage to the tissue architecture (as suggested by TOFT) can explain the basis of carcinogenesis alone. SMT and TOFT do not contradict each other but come into confluence and complement each other in a single unified theory of carcinogenesis (Rosenfeld, 2013).

1.2.1. HALLMARKS OF CANCER

In 2000, Douglas Hanahan and Robert Weinberg suggested that the majority of cancer cells share a common set of acquired “hallmark” capabilities: 1) evading apoptosis, 2) self-sufficiency in growth signals, 3) insensitivity to antigrowth signals, 4) sustained angiogenesis, 5) limitless replicative potential, 6) tissue invasion and metastasis (Hanahan and Weinberg, 2000). These hallmarks of cancer have been extended further since then. A decade later, in 2011, same authors added four new hallmarks to the original six, two of which were called as emerging hallmarks as they were not fully validated yet at that time: 7) genome instability and mutation, 8) tumour-promoting inflammation. The emerging hallmarks: 9) deregulating cellular energetics, and 10) avoiding immune destruction (Hanahan and Weinberg, 2011). The hallmarks of cancer are presented in Figure 2. Here by, we will mainly focus on one of the hallmarks, genome instability and mutation, although we will also touch upon some other hallmarks throughout the way.

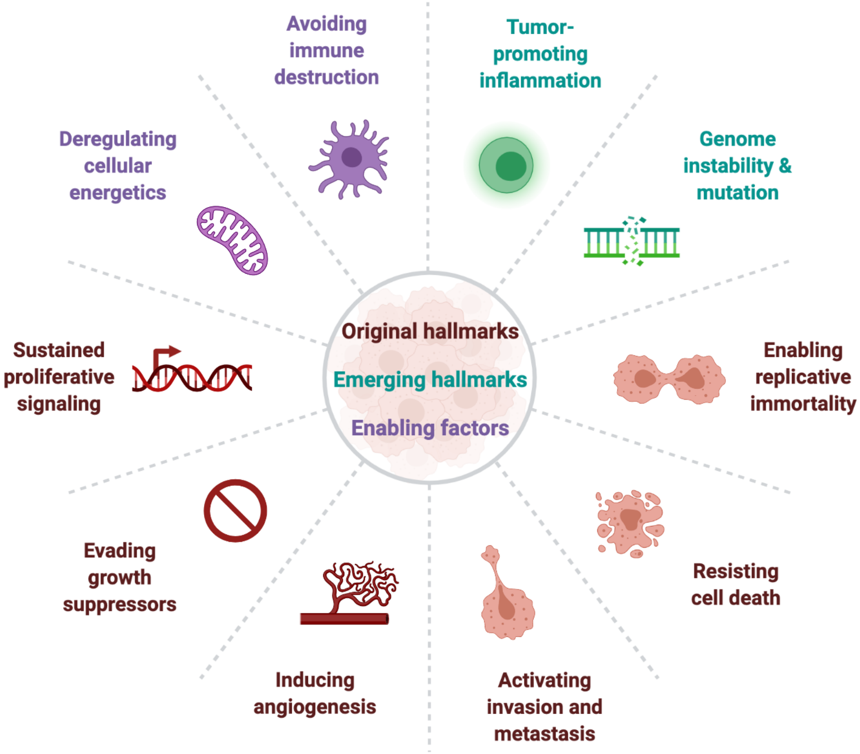
Avoiding immune destruction
Tumor-promoting inflammation
Deregulating cellular energetics
Genome instability & mutation
Sustained proliferative signaling
Original hallmarks
Emerging hallmarks
Enabling factors
Enabling replicative immortality
Evading growth suppressors
Resisting cell death
Inducing angiogenesis
Activating invasion and metastasis

Figure 2. The hallmarks of cancer. Figure adapted from "Hallmarks of Cancer" template by BioRender.com (2021). (Hanahan and Weinberg, 2011)

1.2.2. GENOME INSTABILITY AND MUTATION

There are around 4 to 5 million sites in which the genome differs between two individuals, or between an individual and the reference human genome, and of these, around 25'000 variants are distributed throughout the exome. Although more than 99.9% of these variants consist of non-synonymous single nucleotide polymorphisms (SNPs), there is an estimation of around 2500 large structural variants in the human genome (Auton, et al., 2015). These individual DNA polymorphisms, also called germline variants, determine each person's unique features and influence susceptibility to disease. The germline variants are present in all cells in an individual, are inherited from their parents, and will be passed on to their offspring.

If a mutation occurs in a non-germ cell, also called a somatic cell, it is not transmitted to the progeny. These somatic mutations are a result of natural cell growth and are usually repaired by different DNA repair machineries (Kandoth, et al., 2013). There are various types of molecular changes in the cell, but the word 'mutation' refers to changes that affect the nucleic acid or DNA. Although it should be noted that if the organism is a virus, the word nucleic acid could refer to either DNA or RNA depending on the virus.

Mutations can be divided into two main categories: synonymous and non-synonymous. Synonymous mutations are point mutations that only affect the DNA and the mRNA, but not the encoded protein. This is mainly due to the fact that several amino acids are degenerate and are coded by more than one codon (Chamary, et al., 2006). Synonymous mutations, also called silent mutations, were previously known to have no impact on the DNA, RNA, or the cell in general, hence were assumed to be under no selective pressure (Kimura, 1977). But in recent years, several studies have shown that they have a role in splicing, RNA stability, RNA folding, translation or co-translational protein folding, among others (Sauna and Kimchi-Sarfaty, 2011).

Non-synonymous mutations can be either a point mutation, which involves a change in a single base pair, or structural variations which engage much larger segments of DNA (Loewe, 2008). Point mutations are the most common types of mutations, they usually affect one gene and are divided into three subcategories: missense, nonsense, or frameshift mutations (Kandoth, et al., 2013) (Figure 3).

Missense mutations alter the amino acid sequence and can render the resulting protein non-functional. One example of this type of mutation is in sickle-cell anemia, where a single nucleotide changes from adenine

(A) to uracil (U) in the codons of glutamic acid within the hemoglobin protein, changing it to a valine (Mandal, et al., 2020). Missense mutations are the most common type of non-synonymous mutations in cancer, having a high impact in the formation of cancer driver genes (Stehr, et al., 2011) (Figure 3).

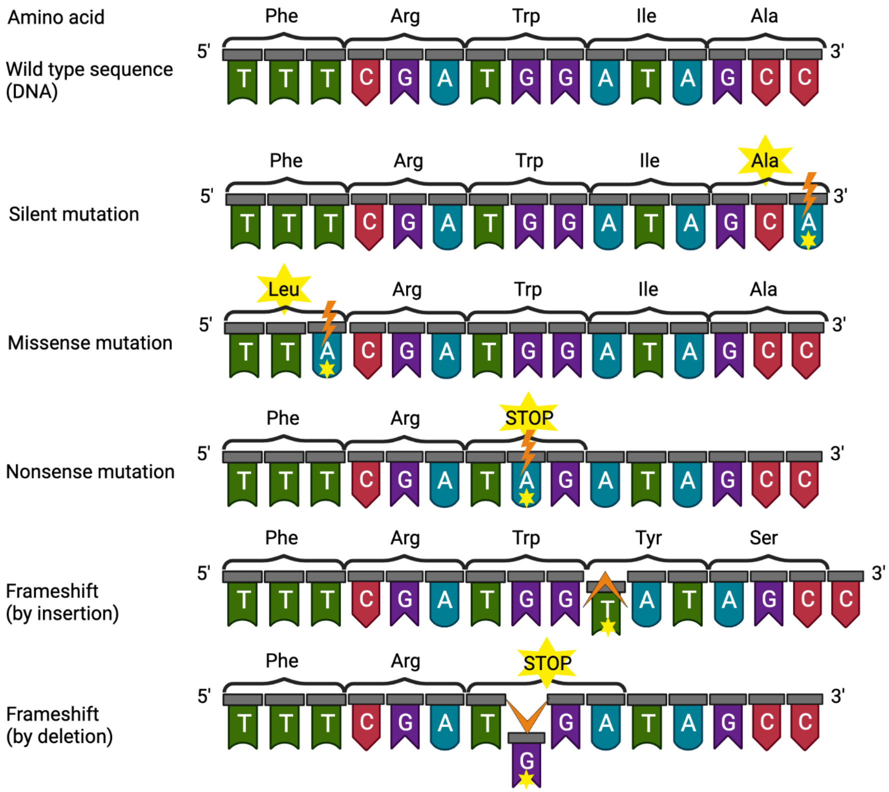

Figure 3. Different types of point mutations. Figure generated with BioRender.com.

Nonsense mutations lead to replacement of an amino acid encoding codon with a chain-terminating codon, thus cause premature termination of protein synthesis and a truncated protein is made as a result (Figure 3). This truncation can interfere with the normal function of the protein in various ways, and the non-functional proteins can also displace functional versions of the same protein from multi-protein complexes (Jopling, 2014). Nonsense mutations are less frequent in cancer-related genes compared to other genes and are often located closer to stop codons in those genes, which putatively minimize their deleterious effects (Chu and Wei, 2019).

Frameshift mutations arise when the DNA is disrupted by the insertion or deletion of one or more nucleotides which shifts the way the sequence is read; provided that the number of nucleotides added or removed is not a multiple of three, or in other words it is not an in-frame mutation (Figure 3) (Loewe, 2008). Frameshift mutations are common in several cancer types, more specifically in tumours with microsatellite instability, which are characterized by alterations in the genome-wide microsatellite repeats (Yamamoto, et al., 2020).

The second type of mutation are the ones involving large-scale changes in the chromosomes and alter the function of several genes. The structural variations or chromosomal abnormalities are a striking feature of cancer cells and are caused by genomic instabilities inherent to most cancers. These alterations either affect the whole chromosome resulting in an aneuploidy (an abnormal number of chromosomes), or create rearrangements through chromosome structure instability as a consequence of improper repair of DNA damage (Thompson and Compton, 2011).

1.2.3. ONCOGENES AND TUMOUR SUPPRESSORS

The majority of driver alterations found in cancer cells can be categorized into two main types of genetic changes: gain-of-function mutations in proto-oncogenes, and loss-of-function mutations in tumor suppressor genes.

Proto-oncogenes

Proto-oncogenes are normal genes that are likely to be involved in cell growth regulation or differentiation, and when mutated, the cells can become cancerous (Adamson, 1987). Oncogenes arise when a proto-oncogene is mutated, resulting in an increased expression level or activity. Most of the mutations in the proto-oncogenes act in a dominant manner, meaning that a single mutated allele is sufficient to contribute to oncogenesis, and these mutations are often well-defined hot spots of missense or in-frame insertion–deletion (indel) mutations.

In 1911, Peyton Rous showed that by injecting cell free filtrate from fowl sarcoma, new tumours could form in other fowl's breast tissue. This was the first report in which a cell free substance could transmit the tumourigenesis into new cells (Rous, 1911). This study met with considerable scepticism, because cancers were believed to be endogenous, rather than infectious. But later between the years 1941-1958, several reports confirmed that a tumour induced by Rous sarcoma virus (RSV) could release infectious viruses (Rubin, 1955). In 1970s, it had become clear that the human genome could carry retroviruses that can transmit cancer. The term proto-oncogene was first used to describe the cellular precursor of a retroviral transforming gene, and to distinguish these from the oncogenes (Martin, 2001). In 1976, a study showed that the *v-SRC* gene is in fact a transduced allele of a human cellular gene (*c-SRC*) picked up by recombination during the retroviral life cycle (Stehelin, et al., 1976). This breakthrough was one of the most influential discoveries in the field of cancer research, and it was the beginning of many more discoveries of other oncogenes such as *MYC* and *EGFR* in the following years (Bister, 2015). Today, we know at least a few hundred cancer driver genes that are mutated in several cancers, with several being proto-oncogenes containing viral counterparts; some examples are: *H-ras, N-ras, K-ras, EGFR, ALK, cMYC* and *nMYC* among others (Futreal, et al., 2004).

Oncogenes can be divided into four different categories based on their function:

1) **Growth factors**. The discovery of the *SIS* oncogene and its encoded protein, the beta chain of Platelet-derived growth factor (PDGF), established the principle that growth factors can act as an oncogene (Heldin and Westermark, 1999). By now, there are several oncogenes identified with mutations that result in a clonal growth advantage, with many of them being growth factor receptors.

2) **Growth factor receptors**. These receptors that are called receptor tyrosine kinases in general terms, are cell membrane receptors that can be bound by extracellular growth factors that will result in the activation of the intracellular tyrosine kinase catalytic domain. Mutations in these receptors can for instance result in constitutive activation of receptor tyrosine kinases without a binding factor, which can give a growth advantage to the tumour cells (Arteaga, 2002). Some examples of growth factor receptors are *ALK*, *EGFR*, *ROS*, and *TRK* among others. The ALK receptor is discussed further in study III.

3) **Signal transducers**. The signals received by the growth factor receptors is transferred into the nucleus with the help of a series of complex pathways also called signal transducers. Many proto-oncogenes are members of this group, including tyrosine kinases such as *SRC*, serine/threonine kinases such as *RAF-1*, and guanosine triphosphate (GTP)-binding proteins such as RAS family of protooncogenes (Kaziro, et al., 1991).

4) **Transcription factors**. Transcription factors are proteins that regulate the expression of several downstream targets. Often either tyrosine or serine kinases activate these factors which then enter the nucleus to regulate the transcription of gene of interest. Some examples of proto-oncogene transcription factors are: *ETS*, *FOS* and *JUN* (Darnell, 2002).

Tumour suppressors

Tumour suppressor genes are normal genes that are often involved in controlling inappropriate cell growth and division, stimulating planned cell death, and DNA repair machineries to protect the cells from the accumulation of dangerous mutations (Chial, 2008). These genes often show a broad range of inactivating mutations that are usually recessive, meaning that both alleles need to be mutated for the function to be lost.

RB1 was first identified as a tumour suppressor gene in 1971, in a study based on retinoblastoma patients. Retinoblastoma is a rare childhood eye tumor and while susceptibility to this cancer is transmitted as a dominant trait, it is not sufficient to form a tumour. In that study, Alfred Knudson suggested that both alleles of *RB1* need to be mutated in order for the cells to form a tumour (Knudson, 1971). This concept, also called the two-hit hypothesis, was advanced with many following studies, with a conclusion that the susceptibility to retinoblastoma can be inherited from one's parents, but a second mutation in the other allele is essential for the tumour formation, hence this cancer is rare in people with no family background since there is a need for two independent somatic mutations to inactivate both normal copies of *RB1* in the same cell (Cavenee, et al., 1986). Although *RB1* was first discovered in eye tumours, today we have evidence that this gene is also lost in several other cancers including bladder, breast and lung carcinomas. The identification of *RB1* was a starting point for characterization of many other tumour suppressors, either with the similar inherited manner, or non-inherited adult cancers such as colon carcinoma (Cooper, 2000). In 1979, only a few years after the discovery of *RB1*, the second tumour suppressor gene, tumour protein 53 (*TP53)* was discovered to be the most common target of genetic alterations in human malignancies, with frequent mutations in up to 50% of all cancers. P53 protein can also behave in a dominant-negative manner, meaning that the mutated P53 suppresses the activity of the wild type P53 (Hofseth, et al., 2004). Although most tumour suppressors are found to have frameshift and nonsense mutations, *TP53* mutations are often missense mutations resulting in P53 mutant protein. It has been shown that some tumours endow the mutant protein with new activities and use it to increase tumour progression and to resist anticancer treatments. These activities are referred to as *TP53* gain-of-function mutations (Oren and Rotter, 2010).

Tumour suppressors can be classified into three major categories: caretakers, gatekeepers and landscapers. However the distinction between these groups is not always clear. Caretakers are genes that play a role directly or indirectly in DNA repair, meaning that without these genes, the mutation rate will be much higher, and the cells can become cancerous. Some examples of DNA caretaker genes involved in DNA repair are *MSH2* and *MLH1* genes (Macleod, 2000). The term gatekeeper was first introduced to describe the function of the adenomatous polyposis coli (*APC)* tumour suppressor gene. Gatekeeper genes directly control the cell cycle by different mechanisms and mutations in these genes can result in permanent imbalance of cell division over cell death. Patients with APC germline mutations are at great risk for developing colorectal cancer, as this gene controls several cellular processes including migration and adhesion, transcriptional activation, and apoptosis (Kinzler and Vogelstein, 1996). An example of a gene that is categorized as both caretaker and gatekeeper is the previously mentioned *TP53*, which plays a critical role in a cell cycle check point prior to DNA replication. Depending on the amount of DNA damage, the cell will either pause the cell cycle for some DNA repair machinery to remove the damage (caretaker role), or will go through apoptosis or programmed cell death if the damages are unrepairable (gatekeeper role) (Deininger, 1999). The third and last group of tumour suppressors are landscapers, which control the microenvironment surrounding the cells, by regulating the extracellular proteins, cell surface markers, or secreted growth factors among others. Mutations in these genes can make the microenvironment suitable for tumour growth (Macleod, 2000).

Currently we know about 719 cancer driver genes, which of 554 are classified into either oncogenes or tumour suppressors, with a great overlap, as genes can have different roles depending on the cell/tissue type or under different environmental stress (Figure 4) (Sondka, et al., 2018). An interesting example would be NRF2, which has been traditionally considered to be a tumor suppressor because of its cytoprotective functions against oxidative stress, but recent studies have found evidence that NRF2 promotes the survival of cancer cells by protecting them from excessive oxidative stress. Therefore it is still not clear if NRF2 acts as a tumor suppressor or as an oncogene (Menegon, et al., 2016). In addition, it has been observed that some genes can gain either tumour suppressor or oncogenic function by fusing to either of these group of genes. These are called the fusion partners in the classification below (Figure 4). An extensive work is done by thousands of cancer researchers to determine the function of cancer driver genes, and the work is continually underway (Colaprico, et al., 2020; Dietlein, et al., 2020; Sondka, et al., 2018).

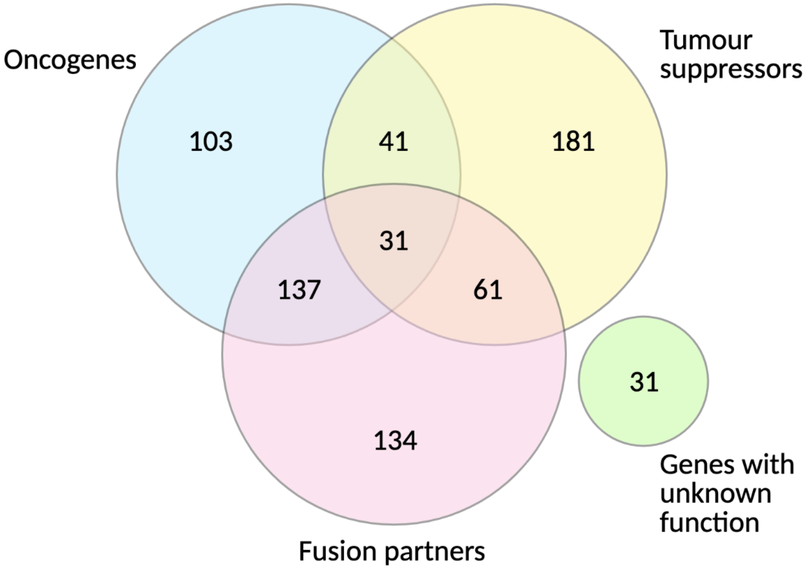

Figure 4. Quantification of cancer genes. Figure adapted from (Sondka, et al., 2018).

1.3. LNCRNAS IN CANCER

In the 1970s, researchers started to realize that the transcribed genome include more than just coding genes and the large RNAs such as rRNA and tRNA known at that time. For a long time the remaining non-coding regions were called ‘junk DNA’ (Ohno, 1972). By the discovery of introns in 1977, a small portion of the non-coding regions were clarified (Berget, et al., 1977; Chow, et al., 1977), and by the 1980s snRNAs and snoRNAs were recognized (Busch, et al., 1982). But it wasn’t until early 2000s, with the arrival of high throughput technologies that the term ‘pervasive transcription’ was introduced, which suggests that a large portion of the human genome is transcribed. It is now estimated that more than 75% of the human genome is transcribed (Li and Liu, 2019). However, the pervasive transcription concept has been questioned in many ways by researchers, mainly due to the low conservation rate and low abundancy for most non-coding transcripts (Wang, et al., 2004). The main question of whether such transcripts have any biological function is still to be answered for the majority of these RNAs.

LncRNAs are generally defined as transcripts with a length of more than 200 nucleotides (typically between 1 to 10 kb) that lack obvious protein-coding capacity. The recent estimate is that there are as many as 28’000 lncRNA transcripts driven by RNA polymerase II in the human genome (Figure 1) (Hon, et al., 2017; Huarte, 2015; Iyer, et al., 2015). Although the abundance of these transcripts was previously thought to be transcriptional noise, recent studies have confirmed the significant tissue and cell specific transcription of lncRNAs. It is also worth mentioning that the majority of lncRNAs are not conserved sequentially, but rather structurally (Diederichs, 2014).

LncRNAs are often classified into four main categories based on their genomic location in relation to other genes (Figure 5).

1) **Sense lncRNAs,** that overlap with a protein coding gene on the same strand (sense); they can overlap several exons and/or introns.

2) **Antisense lncRNAs,** that originate from the antisense RNA strand of a protein coding gene.

3) **Bidirectional lncRNAs,** that share a promoter with a protein-coding gene, yet they are located on the opposite strand.

4) **Intergenic lncRNAs,** that are located between two protein coding genes and are not overlapping with either of them (Figure 5) (Balas and Johnson, 2018).

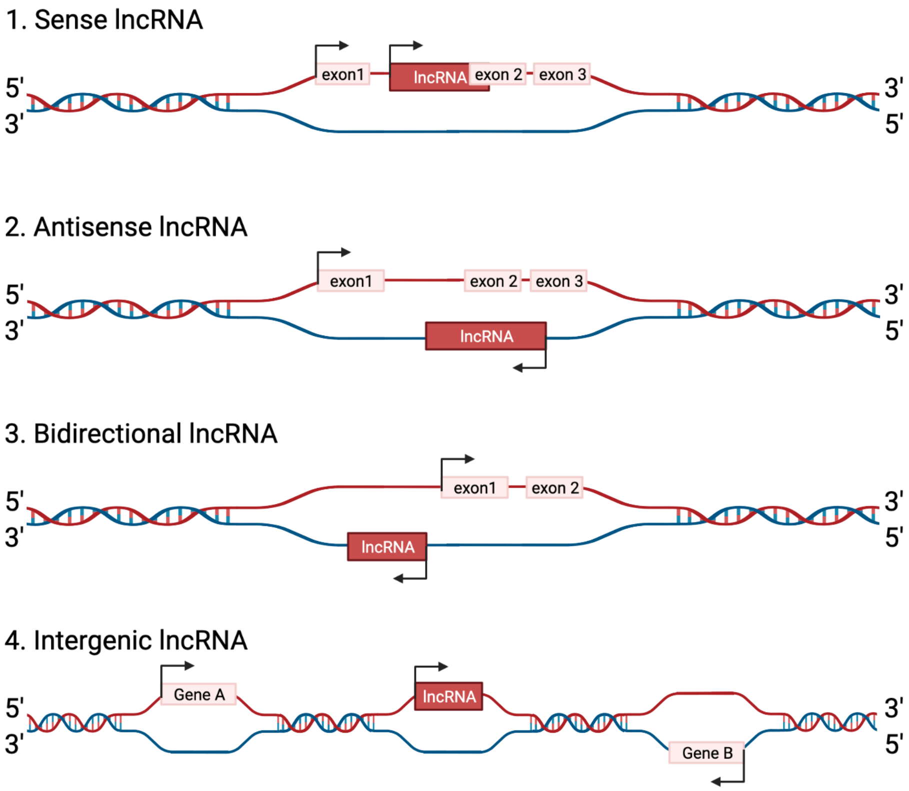

Figure 5. LncRNA classification based on genomic location. Adopted from (Balas and Johnson, 2018) and generated with BioRender.com.

In early 1990s, some of the first lncRNAs with possible biological functions were uncovered; two main examples are *H19* and *XIST* (Brannan, et al., 1990; Brown, et al., 1992). With several more functional lncRNAs discovered recently, these RNAs are also classified based on their function. There are three main categories of known functional lncRNAs:

1) **Guide lncRNAs**, which are RNAs that can bind to regulatory or enzymatically active proteins, such as transcription factors and chromatin modifiers, to guide them to their functional location (Wang and Chang, 2011). The main example of these type of lncRNAs is *HOTAIR* (Rinn, et al., 2007).

2) **Molecular scaffolds lncRNAs**, which provide a platform for the assembly of various regulatory proteins and co-factors. Some examples are: *TERC*, *XIST* and *MALAT1* (Brown, et al., 1991; Feng, et al., 1995; Ji, et al., 2003).

3) **Decoy lncRNAs,** that are negative regulators that often titrate the regulatory factors away from their target sites. For instance, the lncRNA PANDA binds to a transcription factor called NF-YA, limiting its role in inducing apoptosis, resulting in decreased expression of apoptotic genes (Hung, et al., 2011).

During the last years of the 20th century, the first two lncRNAs were associated with cancer because of their aberrant expression. These lncRNAs, prostate cancer antigen 3 (*PCA3*), which is used as a biomarker, and prostate-specific transcript 1 (*PCGEM1*), that is involved in *c-MYC* activation, have important roles in prostate cancer development (Bussemakers, et al., 1999; Srikantan, et al., 2000). Another lncRNA that was also identified early on is metastasis-associated lung adenocarcinoma transcript 1 (*MALAT1*), which was used as a prognostic parameter for lung cancer survival (Ji, et al., 2003). By now *MALAT1* is known to be extremely abundant in both normal cells and several malignancies including in the liver, breast and colon, although the exact mechanism of action of this lncRNA in cancer is yet to be discovered (Huarte, 2015; Sun and Ma, 2019).

In addition, several studies have described lncRNAs that act as downstream effectors in established cancer-relevant pathways. One of the first studies revealed a lncRNA called large intergenic non-coding RNA p21 (*lincRNA-p21*), as a bona fide p53 transcriptional target that mediates global gene repression and apoptosis in the p53 pathway (Huarte, et al., 2010). Similarly, other studies have linked various lncRNAs to known cancer driver genes such as *MYC* and *BRAF* (Flockhart, et al., 2012; Kim, et al., 2015). While intriguing, these studies are still limited in number, and it is likely that lncRNAs have roles also in many other oncogenic programs.

In study I, we make use of the increasing availability of mutational and transcriptomic data from tumours, to identify associations between coding oncogenic drivers and potential effector lncRNAs in a more systematic way (Ashouri, et al., 2016).

1.4. STRESS RESPONSE

Cells respond to extracellular or intracellular stress factors in various ways, and they either survive or move towards a programmed cell death. The way cells react to different stimuli depends on the nature and duration of the stress and also the cell/tissue type. Cancer cells depend on several stress response pathways to survive, including heat shock, DNA damage and oxidative stress responses.

Stress-Induced Cell Death

There is an equilibrium between the growth rate and the rate of cell death when cells are in homeostasis. If, after being exposed to stress, the cell's response is not successful, cell death programs will be activated to eliminate these damaged cells. Although the most used term for the programmed cell death is apoptosis, there are also two other types of cell death namely necrosis and autophagic cell death. Depending on the stimuli cells are exposed to, either one, two or all three types of cell death can be initiated simultaneously. Autophagy is the first to be activated prior to apoptosis, and necrosis will be the last to act (Chen, et al., 2018).

Cellular Stress Responses

Depending on the type of stress, cells will activate different protective responses; for instance there will be different pathways activated in response to heat shock, DNA damage, or oxidative stress (Fulda, et al., 2010).

The heat shock response was first discovered in cells that were exposed to heat, resulting in an increased temperature (3-5°C above normal). By now, many studies have shown that other types of stress such as oxidative stress and heavy metals can also induce a similar response in cells. The main mechanism of response to these stimuli is protein damage leading to the aggregation of unfolded proteins. To help removing these aggregations, cells increase the expression of molecular chaperones, also called heat shock proteins (HSP). It is worth mentioning that the heat shock response is known to be one of the most evolutionarily conserved cytoprotective mechanisms found in nature (Kennedy, et al., 2014). In cancer cells, the heat shock response is activated through induction of heat shock factor 1 (HSF1), which is the master regulator of the heat shock response. This protein has been found to have several functions in tumourigenesis and metastasis, and many studies and clinical trials have considered this gene both as a biomarker and as a therapeutic target (Carpenter and Gökmen-Polar, 2019).

DNA damage is one of the first common initial events in the cells exposed to extracellular stress, such as ultraviolet (UV) light, chemotherapeutic agents, or other environmental toxins. Several DNA repair mechanisms exist in the cells, each related to a different kind of damage. There are many different kinds of DNA lesions, but they can be divided into two main categories: affecting either a single DNA strand or both DNA strands (Jackson and Bartek, 2009). Some examples of DNA damage repair mechanisms are: DNA mismatch repair, which repair mismatches that were introduced during DNA replication, base excision repair, which corrects the abnormal DNA bases in a single DNA strand and nucleotide excision repair, which operates as the main pathway responsible for the removal of bulky DNA lesions induced by UV irradiation, among many others (Jackson and Bartek, 2009). DNA damage response defects are known in several cancers. For example, in about 15% of colorectal cancers there is evidence for DNA mismatch repair deficiency whereby errors in DNA replication are not corrected, which results in microsatellite instability in these tumours (Lord and Ashworth, 2012).

Oxidative stress is another common form of stress. A balanced level of oxygen and reactive oxygen species (ROS) is necessary for maintaining cellular homeostasis. ROS are derived from oxygen, an obligate component of eukaryotic organisms, and they are often in forms of either superoxide (O_2^-), hydroxyl radical (HO^-) or hydrogen peroxide (H_2O_2) (Trachootham, et al., 2008). Both intracellular and extracellular factors can result in an increase in ROS levels. Some examples of extracellular factors are irradiation such as UV or x-ray exposure, and chemicals such as metabolites. Cells need a balanced level of ROS and active antioxidant defense mechanisms to survive. Such defense usually includes ROS-metabolizing enzymes and proteins including glutathione peroxidase and glutathione. When there is a disturbance in this equilibrium, cells begin the oxidative stress response (Fulda, et al., 2010). It has also been shown that occasionally there is a cross talk between different stress responses, for instance high levels of ROS can also activate some heat shock proteins. But this is not surprising as ROS can damage all kinds of molecules in the cells, including DNA, RNA, proteins, carbohydrates, and lipids (Trachootham, et al., 2009).

It is long known that ROS levels are generally increased in cancer cells, although it should be noted that the reduction–oxidation (redox) alteration in these cells is quite complex as there are several factors involved. Cancer cells can tolerate high amounts of ROS by modulating the activity of proteins and transcription factors involved in the stress response, discussed further below.

Figure 6 illustrates a summary of all three types of stress responses in the cancer cell. Briefly, cancer cells are exposed to different kinds of stress, such as oxidative damage, heat shock and DNA damage as presented here. Increased amount of ROS result in DNA damage and activates both oxidative stress and DNA damage response. On the other hand, aneuploidy, which is a common phenomenon in cancer cells where the number of chromosomes is abnormal, results in increased protein dosage, which can activate the heat shock response. Through activation of HSF1 and other chaperones, cancer cells try to alleviate the negative effect of the abnormal chromosome count (Solimini, et al., 2007).

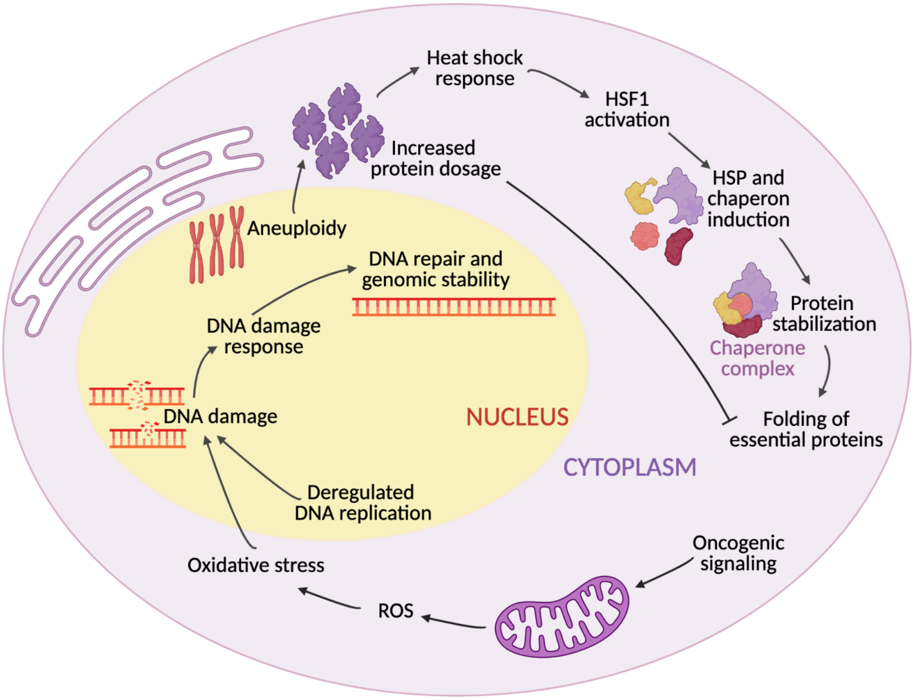

Figure 6. Stress response in cancer cells. (Figure adapted from (Solimini, et al., 2007) and generated with BioRender.com.

1.4.1. ALTERED REDOX BALANCE IN CANCER

In cells, ROS are produced as a by-product of cellular metabolic pathways mainly in mitochondria. Elevated levels of oxidative stress and oxidative damage products have been observed in many cancer types such as leukemia and in various types of solid tumours. Both extrinsic and intrinsic mechanisms are found to be causing the increased ROS levels, although the precise underlying cause is still unclear (Trachootham, et al., 2009).

Intrinsic mechanisms of redox imbalance

Some examples of intrinsic mechanisms are through the dysfunction of mitochondria, various oncogenic activities, and the loss of function of P53. Mitochondria are responsible for production of the major part of cellular energy, and also regulate programmed cell death or apoptosis. Cancer cells need high amounts of adenosine triphosphate (ATP) as they have a high proliferation rate, and mitochondria generate ATP by oxidizing lipids, glucose and amino acids, with ROS produced as a side product (Sabharwal and Schumacker, 2014). Numerous germline variants and somatic mutations have been detected in the mitochondria of cancer cells (Jiménez-Morales, et al., 2018). The somatic mutations in mitochondria can be divided into two groups, tumourigenic (or pathogenic) and adaptive (or beneficial). The tumourigenic mutations result in increased ROS levels that can often initiate tumourigenesis by mutagenizing proto-oncogenes into oncogenes. The adaptive mutations on the other hand can aid tumour cells to survive through environmental changes such as increased ROS toxicity and reduced oxygen tension (Chinnery, et al., 2002). Interestingly, one study showed that after replacing the mitochondrial DNA (mtDNA) in non-metastatic cancer cells by mtDNA from highly metastatic cells, the metastatic potential was acquired by the recipient cells. The metastatic mtDNA contained two mutations that were associated with overproduction of reactive oxygen species (ROS) (Ishikawa, et al., 2008). Some of the oncogenes that are found to induce ROS activity are *RAS* and *c-MYC*. High expression levels of *c-MYC* produce a sufficient amount of ROS to induce DNA damage and this activates *TP53* as a result (Vafa, et al., 2002).

As mentioned before, cancer cells often contain high levels of ROS, resulting in an increased mutation rate, especially in the mitochondrial genome, which leads to even more ROS generation. P53 has an important role in removing the DNA damage from both nuclear and mitochondrial genome. Therefore, in tumour cells with loss of function mutations of *TP53,* imbalanced redox levels are observed which are followed by high mutation rate and aggressive tumour growth (Trachootham, et al., 2009).

Extrinsic factors

In addition to the cell itself, various external factors can increase the ROS levels in the cells, including hyperthermia, hyperoxia, chemotherapeutic agents, UV radiation, among others. For example, in cigarette smokers, high levels of ROS are generated in the cells after exposure to tobacco, which result in inflammation and DNA damage, a possible cause of lung cancer in smokers (Prasad, et al., 2017). Sunlight, specifically UVB photons can be directly absorbed by DNA, which results in lesions. In addition, the non-DNA chromophores that are present in skin cells can also absorb UV protons, which lead to the formation of ROS and other toxic photoproducts that may cause DNA damage (Schuch, et al., 2017).

Cellular defense against reactive oxygen species

Human cells are equipped with several mechanisms that balance ROS levels in the cells, which are called antioxidants in general terms. Antioxidants can be divided into two main categories: enzymatic and nonenzymatic; some of which are synthesized in the cells, and others are absorbed through the diet (Tebay, et al., 2015). There are three main types of enzymatic antioxidants: superoxide dismutases (SODs), catalases, and glutathione (GSH) peroxidases (Birben, et al., 2012). SOD is one of the major antioxidants in the cell that is present in three different forms and localized in different parts of the cell, such as in the extracellular matrix and mitochondria. As clear from the name, these enzymes have a role in superoxide dismutation, which is catalyzing the dismutation of superoxide (O_2^-) into oxygen (O_2) and peroxide (H_2O_2). Superoxide is one of the primary forms of ROS produced by various sources in the cell (Zelko, et al., 2002).

H_2O_2 is a product of SODs and other oxidates action, and is often reduced to water by either catalases or GSH peroxidases. Catalases also bind to nicotinamide adenine dinucleotide phosphate (NADP(H)) as a reducing equivalent to protect itself from oxidation by hydrogen peroxides (H_2O_2) as it is reduced to water (Kirkman, et al., 1999).

GSH peroxidases are a family of enzymes that work mainly with the aim to detoxify H_2O_2 in the cells. GSH itself is the key substrate of these enzymes, discussed further below. Common between all these enzymes is that they all require NADPH as a reducing equivalent (Birben, et al., 2012).

Nonenzymatic antioxidants include vitamins (Vitamin C and E), uric acid, β-carotene, and GSH. GSH is one of the main regulators of intracellular redox balance, and it is highly abundant in all cell types and cell organelles at millimolar concentration. The ratio of reduced glutathione (GSH) to oxidized glutathione (GSSG) is a determinant of cellular oxidative stress (Masella, et al., 2005).

The antioxidant responsive element

The presence of antioxidant response elements (AREs) was first discovered in 1991 in the promoter region of a subunit of glutathione S-transferase as well as the NADPH reductase gene (Rushmore, et al., 1991). By now we have evidence that the transcriptional activation of most of antioxidant enzymes and genes are regulated by antioxidant response elements (AREs). The main transcription factor binding to AREs is nuclear factor erythroid 2–related factor 2 (NRF2). The role of NRF2 in cancer redox balance and its mechanism of action will be discussed further here after.

1.4.2. THE NRF2 PATHWAY IN CANCER

NRF2 is a leucine zipper (bZIP) transcription factor known as the master regulator of oxidative stress, and it regulates a wide variety of biological processes. NRF2 is regulating several drug-metabolizing, antioxidant and anti-inflammatory genes and also regulating mitochondrial bioenergetics, mainly by binding to ARE sites in the regulatory regions of the downstream targets (Holmström, et al., 2013). Therefore, it is important to first understand how NRF2 activity is regulated in the cell, both transcriptional and at protein levels.

Molecular regulation of NRF2

Transcription of *NRF2*, is regulated in many different ways. There are several xenobiotic response element (XRE) sequences identified in the promoter regions of *NRF2* that are bound by Aryl hydrocarbon receptor (AHR) transcription factor, which is another xenobiotic-sensing transcription factor (Ma, et al., 2004). NRF2 also appears to autoregulate its own expression through an ARE element located in the proximal region of its promoter (Kwak, et al., 2002).

NRF2, like most other stress-responsive transcription factors, is regulated at the protein level. NRF2 protein is composed of seven functional domains known as Neh1–Neh7, where the Neh2 domain is the major regulatory domain. The Neh2 domain contains seven lysine residues that are responsible for ubiquitin conjugation, as well as two binding sites namely ETGE and DLG motifs that interact with Kelch-like ECH associated protein 1 (KEAP1) (Figure 7) (McMahon, et al., 2006). KEAP1 is a substrate adaptor protein for the Cullin 3 (CUL3)-dependent E3 ubiquitin ligase complex. In the absence of stress stimuli, the main mechanism which controls the NRF2 stability is through KEAP1 (Figure 7).

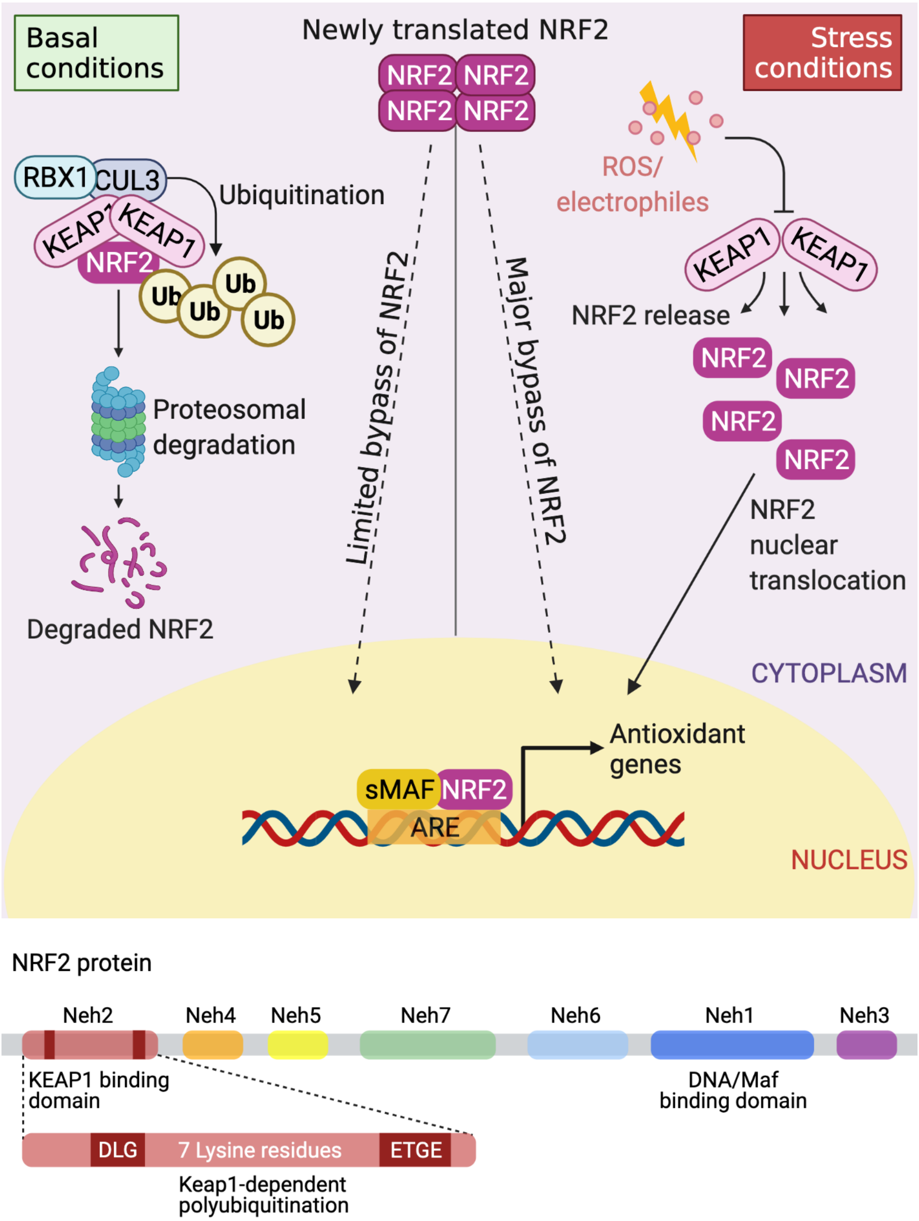

Figure 7. NRF2 pathway and NRF2 protein domains. Figure was adapted from "KEAP1-NRF2 pathway" template by BioRender.com (2021) and (Lee and Hu, 2020).

KEAP1 binds to NRF2 Neh2 domain, on the ETGE and DLG motifs as a dimer and promote NRF2 polyubiquitination and its subsequent proteasomal degradation by the 26S proteasome (Zhang, et al., 2004). Therefore, the NRF2 protein concentrations are tightly regulated by this complex. Multiple stresses as well as small molecules of endogenous and exogenous origin can activate NRF2, mainly through KEAP1 inactivation or disruption of its binding to NRF2. Electrophiles and ROS above a critical threshold can alter the chemical modification of critical cysteine residues of KEAP1 and disrupt KEAP1-mediated NRF2 ubiquitination, resulting in accumulation of NRF2 in the nucleus. NRF2 then binds to ARE sites of the target genes, and activate the stress response pathway (Figure 7) (Dinkova-Kostova, et al., 2005). However, there are conflicting views on how chemical agents can activate NRF2, and it is still unclear if the translocation of NRF2 into the nucleus is regulated (Li, et al., 2012).

The dual role of NRF2 in carcinogenesis

NRF2 plays an important but somewhat ambiguous role in the initiation, promotion, and progression of cancer. For the last 50 years, natural antioxidants have been known for their chemopreventive function by activating a NRF2 related response. Therefore the ‘good side’ of NRF2 that eliminates the chemical toxins and carcinogens from the cells has been well established (McMahon, et al., 2001). But in early 2000, researchers discovered a new side of NRF2, namely ‘the dark side of NRF2’, that is upregulation of NRF2 in cancer cells provides them with a growth advantage under harmful environments (Wang, et al., 2008). Thereafter, more evidence has indicated the role of *NRF2* in stimulating cancer tumorigenesis and chemoresistance. Activating mutations in *NRF2* and *KEAP1* have been identified in several cancers, and constitutive activation of *NRF2* is known to favor the survival of malignant cells, protecting them against apoptosis and senescence, oxidative stress, chemotherapeutic agents, and radiotherapy (Menegon, et al., 2016). Therefore, *NRF2* promotes the survival of not only normal cells but also cancer cells. High levels of ROS are generated in cells by ongoing mitochondrial oxidative phosphorylation, however cancer cells might switch to aerobic glycolysis for generating ATP instead (also known as the Warburg effect) (Vander Heiden, et al., 2009). Cells can also encounter high ROS level upon extracellular exposure to xenobiotics (drugs, radiotherapy, UV light). Because of this, the constitutive activation of *NRF2* presents a selective pressure, which favors cancerous cells with activated *NRF2*. Many NRF2 target genes contribute to NRF2 dependent chemoresistance and cancer promotion. Some examples of the known direct targets of NRF2 are: Heme oxygenase-1 (*HMOX1*), NAD(P)H quinone dehydrogenase 1 (*NQO1*), Glutamate-cysteine ligase, modifier and catalytic subunits (*GCLM* and *GCLC*), Thioredoxin and thioredoxin reductase 1 (*TXN* and *TXNRD1*) among many others (Lacher and Slattery, 2016). Despite key roles in human malignancies as well as in normal cell physiology, the NRF2 targetome remains incompletely characterized. In study II, we provide a comprehensive map of genome-wide gene regulation downstream of NRF2.

1.5. ALK IN CANCER

The anaplastic lymphoma kinase (*ALK*) gene was originally identified in an anaplastic large-cell lymphoma (ALCL) cell line, as a fusion gene to the nucleophosmin (*NPM1*) gene (Morris, et al., 1994). ALK is a transmembrane protein with a strong homology to the insulin receptor subfamily of transmembrane tyrosine kinases, and it contains an extracellular domain, a transmembrane part, and an intracellular domain containing a tyrosine kinase region (Morris, et al., 1997). Two secreted growth factors pleiotrophin (PTN) and midkine (MDK) are known to bind to the ALK receptor and activate either of the mitogen-activated protein kinase (MAPK) or the phosphoinositide 3-kinase (PI3K) pathways; although activation of ALK independently of direct ligand interactions has also been proposed (Palmer, et al., 2009). The full length ALK protein has been shown to be expressed in several cancers, including neuroblastomas, melanoma, neuroectodermal tumors and glioblastomas, in addition to increased copy number and activating mutations (Holla, et al., 2017). The most abundant form of oncogenic ALK in human cancers are the ALK fusion proteins that is often the 3' half of ALK, containing its kinase catalytic domain, fused with the 5′ portion of a different gene, which often help ALK bypass the requirement of binding of ligands to its extracellular domain for its activation, resulting in increased oncogenic potential of ALK (Webb, et al., 2009).

The high frequency of ALK mutations in primary neuroblastoma tumours (10-14%) and even higher percentage (up to 26%) in the relapsed tumours has made it an interesting target for treatment of these tumors (Martinsson, et al., 2011). In the last decade, huge efforts from many scientists in both academia and the pharmaceutical industry have led to the development of numerous ALK tyrosine kinase inhibitors (TKIs), a few of which are approved by regulatory authorities, such as the US Food and Drug Administration (FDA) and European Medicines Agency (EMA). About 5% of non-small-cell lung cancer patients, and often the adenocarcinoma subtype, have a rearrangement in the *ALK* gene. The approved drugs, such as crizotinib, ceritinib, alectinib, and lorlatinib are used for the treatment of advanced non-small-cell lung cancer patients with *ALK* rearrangements (Shaw, et al., 2016; Shaw, et al., 2014; Shaw, et al., 2013). But unfortunately, similar to most targeted therapies, the tumour cells evolve resistance to *ALK* TKIs, often by developing secondary resistant mutations in *ALK*. New generations of *ALK* inhibitors have been developed that can overcome resistance to the first-generation (crizotinib) and second-generation (e.g., ceritinib, alectinib, brigatinib). The third-generation *ALK* TKI, lorlatinib, can inhibit the growth of cell lines harboring ALK resistance mutations, and it has now been approved (in March 2021) for patients with metastatic non-small cell lung cancer whose tumours are ALK positive (Johnson, et al., 2014; Zou, et al., 2015).

Neuroblastoma, the most common solid tumour of childhood, often happens in children with no family history of the disease (also called sporadic neuroblastoma) with a small percentage of the patients developing this cancer inheriting the disease (familiar neuroblastoma). Gain-of-function mutations of *ALK* have been reported in both primary and relapsed neuroblastoma tumours, where the full-length ALK is activated predominantly in the kinase domain (Chen, et al., 2008). Therefore, *ALK* is an attractive therapeutic target in neuroblastoma, although, the picture regarding the role of *ALK* as an oncogenic driver in these tumours is less clear.

Most of the FDA approved *ALK* TKIs have also been tested in different preclinical neuroblastoma models and on neuroblastoma patients in clinical trials. Initial results from clinical trials have reported positive individual patient case data, for example in one study a patient with ALK-positive neuroblastoma reached a complete response after treatment with ceritinib (Guan, et al., 2018). In addition, one patient with a metastatic ALK neuroblastoma case showed a partial response after receiving alectinib (Heath, et al., 2018). However, currently there is no approved *ALK* TKI available for neuroblastoma patients.

In study III, based on integrative proteomics and gene expression analyses of neuroblastoma cells exposed to first- and third-generation ALK TKIs (crizotinib and lorlatinib), we identified several relevant biomarkers, signaling networks, and new potential therapeutic targets of neuroblastoma (Van den Eynden, et al., 2018).

1.6. HIGH-THROUGHPUT TECHNOLOGIES

We have come a long way since the discovery of DNA, starting in 1869 by the Swiss physician Friedrich Miescher, continued by revealing the DNA helix structure based on an X-ray diffraction picture by Rosalind Franklin of a DNA molecule. Later on, Watson and Crick revealed the DNA double helix structure, and finally in 1968 when the genetic code was identified by Holley, Khorana and Nirenberg. Many researches have contributed to the understanding and developing of new methods for DNA and RNA sequencing over the years. More than 50 years after the discovery of the codons, we can now read the nearly complete 3 billion base pair sequence of the human genome in the matter of hours (Heather and Chain, 2016).

1.6.1. THE HISTORY OF DNA SEQUENCING

First generation DNA sequencing

The first ever complete nucleic acid sequence, from a tRNA of Saccharomyces cerevisiae (77 bps long), was sequenced in 1965 (Holley, et al., 1965). Together with several other parallel studies, a method called two-dimensional fractionation was then developed (Sanger, et al., 1965). Using this method, the first complete protein coding gene of the coat protein of a bacteriophage (containing 129 amino acids, and 387 bps) was sequenced in 1972, followed by the sequence of the complete genome of the same organism in 1976 (Fiers, et al., 1976; Min Jou, et al., 1972). Another method was developed during the 1970s namely the plus and minus technique, which was a joint effort from two different group of researchers (Maxam and Gilbert, 1977; Sanger and Coulson, 1975). Between these two methods, the one developed by Maxam and Gilbert was adopted world-wide, and is considered to be the real birth of 'first-generation' DNA sequencing (Heather and Chain, 2016). However, in the same year, 1977, Sanger and his colleagues published a much more accurate and robust method, called the dideoxy chain-termination method or as we know it today, Sanger sequencing (Sanger, et al., 1977). This method was a major breakthrough in DNA sequencing, and it became the most commonly used technique for many more years to come. As Sanger sequencing was only able to sequence DNA fragments of less than one kb, a new strategy was developed for further sequencing of longer fragments, namely shotgun sequencing, where the smaller fragments from the Sanger method were then assembled in silico (Anderson, 1981). These methods were then used in the Human Genome Project, between the years 1990 to 2001, resulting in the first complete sequence of the human genome (Lander, et al., 2001; Venter, et al., 2001).

Second generation DNA sequencing

Around the same time as Sanger and shotgun sequencing methods, another technique was developed that had a similar base principle, but several beneficial aspects compared to the previous sequencing approaches. The new method, called pyrosequencing (Nyrén and Lundin, 1985), had two main advantages; first that it could be observed in real time, and second that it could be performed using natural nucleotides (Heather and Chain, 2016). Pyrosequencing was later licensed to 454 Life Sciences and was the first commercially available next generation sequencing (NGS) technology. The 454 sequencing machines, first launched in 2005, were also the first to perform massively parallel sequencing, with about 200'000 reads of 110 bp long DNA fragments (Margulies, et al., 2005). 454 was then acquired by Roche in 2007. In parallel with this, Solexa had started developing a new sequencing method, that was able to sequence the whole genome of a bacteria in 2005, with a high coverage of about three million base pairs in one run. Solexa's first commercial machine, the Genome Analyzer, was launched in 2006, and it could sequence up to one giga base pairs of data in one run. Solexa was then acquired by Illumina in 2007. Although several researchers were developing new methods for sequencing of DNA throughout the years, the third most popular method used during those years, after 454 and Illumina was the sequencing by oligonucleotide ligation and detection (SOLiD) which was launched in 2009 (McKernan, et al., 2009). Illumina and SOLiD produced many more reads compared to 454 (almost 100 million reads for Illumina and SOLiD, compared to 200’000 for the 454 sequencer), but the read lengths were quite short at around 35 bps. It should also be noted that in 2013, Roche announced that it will be closing 454 Life Sciences only six years after acquisition; and by mid 2016, the 454 technology was completely shut down. In 2010, the founder of 454 introduced a new method called Ion Torrent that was mainly based on the 454 technology, but could sequence up to 270 mega base pairs of DNA with 100 base pair read lengths (van Dijk, et al., 2014).

Out of all these methods, Illumina has been the most successful, and has continued to develop the technology and is now ensconced as a near monopoly (Greenleaf and Sidow, 2014). The latest high-capacity platform from Illumina, Novaseq 6000, can generate 3000 GB of data, comprising 20 billion paired end reads with read pairs up to 2 x 250 bp in length. For Illumina’s main competitor Ion Torrent, Ion GeneStudio S5 can generate 50 GB of data comprising 130 M reads of 200 bp length. Therefore it is safe to say that Illumina has been the most influential method among all second generation sequencing technologies (Heather and Chain, 2016).

Third generation DNA sequencing

Third generation sequencing is generally characterized by methods with longer read lengths and shorter processing times compared to the second generation technologies (Deamer, 2010; Eid, et al., 2009; Schadt, et al., 2010). The main technologies used today which can be classified as “Third Generation” are from Pacific Biosciences (PacBio) and Oxford Nanopore. PacBio sequencing was founded in 2004, and is based on single molecule real time sequencing (SMRT) (Eid, et al., 2009), which uses a DNA polymerase as an engine and measures the incorporation of bases using the zero-mode waveguide (ZMW). This allows sequencing of DNA fragments up to 10 kb in length. The latest release from PacBio, the Sequel IIe system, sequences up to 4 million reads with a maximum length of 16 kb. Oxford Nanopore, first released in 2014, also uses a DNA polymerase pore motor which controls the movement of the DNA molecule through the pore and records the electrical changes with each base (Deamer, 2010). Nanopore sequencing can be used to sequence molecules up to 2 Mb in length. Nanopore sequencing is also very portable, with the MinION chip able to be attached to a USB port for use in the field. Another earlier technology from Helicos BioSciences (Pushkarev, et al., 2009) was based on single molecule fluorescent sequencing first suggested in 2003, where researchers showed that the activity of DNA polymerase can be studied at the single molecule level with single base resolution, which allowed degraded samples to be sequenced and avoided associated PCR-bias, (Braslavsky, et al., 2003). However, this technology was quickly overtaken, and Helicos entered bankruptcy in 2015. The longer read lengths from third generation sequencing allows for much greater knowledge of genome assembly and structural variations in the genome, however these technologies have much higher error rates than the shorter fragment based methods.

During the last few decades, many sequencing technologies have evolved, making it possible to study the genetic code in a level never before possible. Both the first generation, and the ‘next’ generation (second and third combined) sequencing (NGS) methods have followed the same trend in the case of sequencing dept and costs. It is fascinating to acknowledge the fact that the estimated cost for The Human Genome Project, the first human genome ever sequenced in 2001, was about three billion dollars over 13 years of work. Today, we can sequence the complete human genome in a matter of hours with the price of less than 1000 dollars (van Dijk, et al., 2014).

A summary of DNA discovery and sequencing history over the past 150 years can be seen in Figure 8.

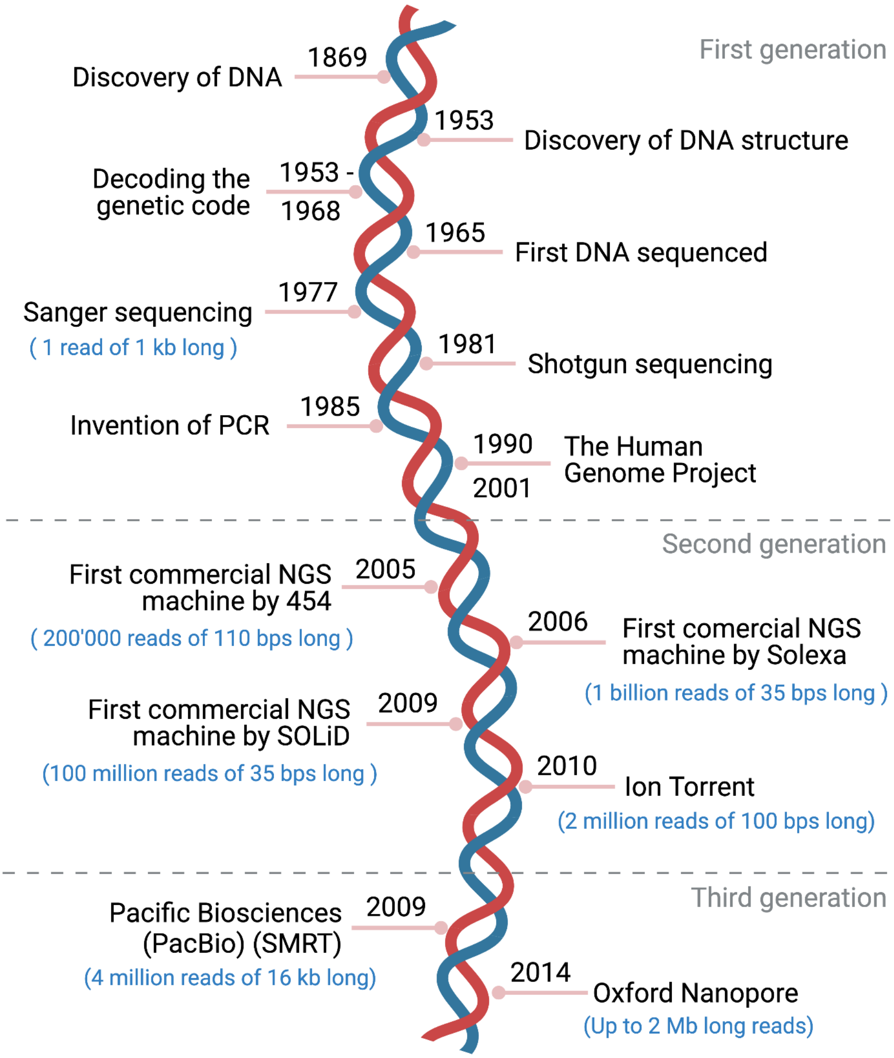
First generation
1869
Discovery of DNA
1953
Discovery of DNA structure
Decoding the genetic code
1953 - 1968
1965
First DNA sequenced
1977
Sanger sequencing
(1 read of 1 kb long)
1981
Shotgun sequencing
1985
Invention of PCR
1990
2001
The Human Genome Project
Second generation
First commercial NGS machine by 454
2005
(200'000 reads of 110 bps long)
2006
First comercial NGS machine by Solexa
(1 billion reads of 35 bps long)
First commercial NGS machine by SOLiD
2009
(100 million reads of 35 bps long)
2010
Ion Torrent
(2 million reads of 100 bps long)
Third generation
Pacific Biosciences (PacBio) (SMRT)
2009
(4 million reads of 16 kb long)
2014
Oxford Nanopore
(Up to 2 Mb long reads)

Figure 8. DNA discovery and sequencing timeline. Figure generated by BioRender.com.

1.6.2. RNA SEQUENCING

In multicellular organisms, almost all the cells have the same genomic code, however the transcriptome, including the level and the composition of expressed RNA molecules, varies significantly between cells. The term transcriptome was first used in the 1990s, even before the completion of The Human Genome Project, and was mainly used in studying human disease. Understanding how gene expression changes upon exposure to different environmental stimuli, or upon a mutation in a key gene, has always been an interesting research topic (Lockhart and Winzeler, 2000). The first technique that aimed to generate transcriptomic data was expressed sequence tags (ESTs) in 1991 (Adams, et al., 1991). Using PCR techniques, the complementary DNA (cDNA) clones that were reverse transcribed products of mRNAs were quantified. Two of these techniques, namely SAGE (serial analysis of gene expression) and microarray were developed in 1995, both aimed to quantify global gene expression using complementary probe hybridization (Schena, et al., 1995; Velculescu, et al., 1995). Although having their flaws in the beginning, microarray technology became the most valuable and used method for global transcriptomic studies for many more years to come. Shortly after the invention of these methods, in 1996 real time quantitative PCR (RT-PCR / q-PCR) was developed (Heid, et al., 1996), to become the gold standard method used for measuring transcript levels.

In 2008, the RNA sequencing (RNA-Seq) method was first described, which used the Illumina technology to sequence the cDNA (Nagalakshmi, et al., 2008). This was the first time researchers could define the exact exon and intron boundaries as well as comprehensive gene structure and transcriptional landscape of the yeast genome. Although it also has its downfalls, with the main disadvantage being the short reads which don’t allow the discovery of long distanced alternative exons (Casamassimi, et al., 2017). Adapting the third generation sequencing methods such as Oxford Nanopore and PacBio sequencing allows the detection of the entire transcriptome with full length RNA transcripts. This allows for the identification of a multitude of splice variants. The invention of RNA-Seq has revolutionized the transcriptomics era, and it provides useful information about the post-transcriptional RNA-editing, as well as splicing variants.

1.6.3. CHIP SEQUENCING

Protein-DNA interactions are essential components of all biological systems. Almost all aspects of cellular function depend on this interaction, some examples are: transcriptional regulation with RNA polymerases and TFs binding to DNA, chromosomal maintenance with histones binding to DNA, replication with DNA polymerases binding to DNA, among many others (Dey, et al., 2012). Because of the importance of this phenomenon, there have been many techniques developed, both *in vitro* and *in vivo*. A ground-breaking study that detected the protein-DNA interaction in live cells for the first time was by John T. Lis and David Gilmour in 1984 (Gilmour and Lis, 1984). In this study, the crosslinking was performed using UV radiation on bacterial cells, and immunoprecipitation of both the RNA polymerase molecules, and the DNA that was bound to these molecules, were then studied further to give a genome wide readout of regions bound by RNA polymerase. This method has matured to a method currently known as chromatin immunoprecipitation (ChIP). In ChIP, chromatin is sheared, then antibodies are used to isolate specific DNA binding proteins and the corresponding bound DNA fragments. The purified DNA from ChIP experiments can be detected in different ways, but the most common techniques used are PCR using primers to known bound regions (ChIP-PCR) and high-throughput sequencing (ChIP-Seq) which sequences all bound DNA fragments. ChIP-seq was developed in 2007 in a study that identified binding sites of the transcription factor STAT1 *in vivo*. (Dey, et al., 2012; Robertson, et al., 2007). Since then, ChIP-Seq has played a major role in discovering transcription factor binding sites, in addition to many other DNA binding proteins.

1.6.4. CANCER GENOMICS

Since the somatic mutation theory (SMT) was first described in 1914, we have come a long way in understanding carcinogenesis. Using high-throughput sequencing technologies, tumours are not only categorized by their location in the body, but rather with their genetic information. In the beginning of NGS era, many research groups started to perform these techniques on tumour cells. To make the most of the increase in DNA sequencing data available, in 2005, a joint project of the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI), called The Cancer Genome Atlas (TCGA) was launched, to become the largest and most comprehensive tumour dataset available for many more years to come (www.cancer.gov/TCGA). The first study published from this project was in 2008, presenting a comprehensive genomic characterization of glioblastoma (The Cancer Genome Atlas Research Network (2008)). The data collection for TCGA was completed in 2016 comprising matched tumour and normal tissues from more than 11’000 patients across 34 tumor types (Figure 9) and the last marker papers were published in 2018 (Hutter and Zenklusen, 2018). There are seven different data types available for almost all of these patients, and as the general pipelines have been similar, the techniques and tools used have undergone a development path throughout the years.

1) **Mutations**, using exome sequencing, which restricts the sequencing reads to only the coding regions of the genome with probe hybridization, and a mutation calling tool called MutSig (Lawrence, et al., 2013), mutations in genes or intergenic regions with non-synonymous mutations are reported.

2) **Copy number alteration**, using Affymetrix SNP 6.0 arrays, the copy number of each DNA region is calculated using an established pipeline (McCarroll, et al., 2008). The focal copy number alterations are then called with the GISTIC tool (Beroukhim, et al., 2007).

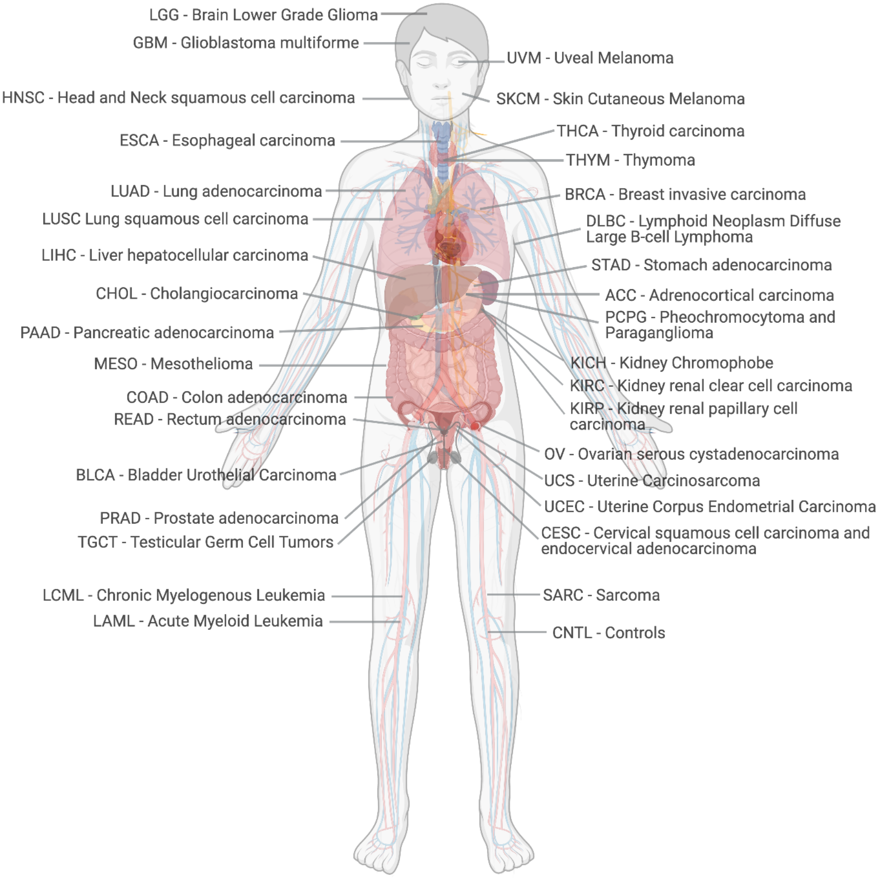
LGG - Brain Lower Grade Glioma
GBM - Glioblastoma multiforme
HNSC - Head and Neck squamous cell carcinoma
ESCA - Esophageal carcinoma
LUAD - Lung adenocarcinoma
LUSC Lung squamous cell carcinoma
LIHC - Liver hepatocellular carcinoma
CHOL - Cholangiocarcinoma
PAAD - Pancreatic adenocarcinoma
MESO - Mesothelioma
COAD - Colon adenocarcinoma
READ - Rectum adenocarcinoma
BLCA - Bladder Urothelial Carcinoma
PRAD - Prostate adenocarcinoma
TGCT - Testicular Germ Cell Tumors
LCML - Chronic Myelogenous Leukemia
LAML - Acute Myeloid Leukemia
UVM - Uveal Melanoma
SKCM - Skin Cutaneous Melanoma
THCA - Thyroid carcinoma
THYM - Thymoma
BRCA - Breast invasive carcinoma
DLBC - Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
STAD - Stomach adenocarcinoma
ACC - Adrenocortical carcinoma
PCPG - Pheochromocytoma and Paraganglioma
KICH - Kidney Chromophobe
KIRC - Kidney renal clear cell carcinoma
KIRP - Kidney renal papillary cell carcinoma
OV - Ovarian serous cystadenocarcinoma
UCS - Uterine Carcinosarcoma
UCEC - Uterine Corpus Endometrial Carcinoma
CESC - Cervical squamous cell carcinoma and endocervical adenocarcinoma
SARC - Sarcoma
CNTL - Controls

Figure 9. TCGA available tumours. Figure generated by BioRender.com.

3) **Gene expression**, in the primary studies, the mRNA profiling was performed using microarrays, while in the more recent studies RNA-Seq was performed using first Solexa Genome Analyzer machines and later with Illumina machines.

4) **DNA methylation**, methylation at CpG dinucleotides was measured using the Illumina machines GoldenGate assays primarily, and Infinium assays in the more recent studies.

5) **MiRNAs**, expression of miRNAs was measured using Affymetrix microarray platforms in the beginning, and in the more recent studies, miRNA-Seq technique was used (Chu, et al., 2016).

6) **Reverse phase protein array (RPPA)**, this data was not available in the initial studies, but in the last few studies, the protein expression profiling with the RPPA technique (Tibes, et al., 2006) was also presented in the marker papers.

7) **Clinical data**, the clinical and pathological data is available for all the patients included in TCGA, including information such as age, race, gender, ethnicity, stage and histological type of cancer, among others. In addition, there are follow-up data such as treatments received, relapsed tumours and the survival rate.

In addition to these data, whole genome sequencing (WGS) data has also been available from a few hundreds of patient samples. The International Cancer Genome Consortium (ICGC) was launched in 2008 (www.icgc.org) and was later joined with TCGA on a project called Pan-Cancer Analysis of Whole Genomes Project (PCAWG). By now PCAWG includes more than 2800 WGS data from more than 35 cancer types. The development of these consortia to compile and make available the abundant DNA sequencing data has allowed thousands of cancer genomics researchers the opportunity to investigate their hypotheses using real patient data.

2 AIM

The main aim of this thesis was to identify transcriptional responses to oncogenic driver mutations in various cancer types. To achieve this, we have made use of large datasets available at TCGA, used marketed cancer drugs, and generated several types of data in the lab.

More specifically the aim and objectives are as follows:

- ✓ To establish a methodology that make it achievable to identify the associations between key driver mutations and expression changes in various tumours (**Study I**).
- ✓ To generate a comprehensive list of recurrent associations between cancer driver mutations and putative downstream effectors including both protein coding and lncRNAs (**Study I**).
- ✓ To explore the associations and validate novel downstream lncRNAs and discover their possible functions (**Study I**).
- ✓ To provide a comprehensive map of NRF2 direct targets (**Study II**).
- ✓ To validate the known NRF2 targets and identify novel genes that can help understanding the NRF2 effect in the tumours (**Study II**).
- ✓ To investigate the proteomic and transcriptomic alteration in tumours upon inhibiting ALK using ALK TKIs (**Study III**).
- ✓ To identify the similarities and differences between the two marketed ALK TKI drugs and their downstream effects (**Study III**).

3 RESULTS AND DISCUSSION

3.1 ASSOCIATIONS BETWEEN CANCER DRIVER MUTATIONS AND LNCRNAS (**STUDY I**)

It is well known that mammalian genomes encode an abundance of mRNA-like transcripts with a length often between 1 to 10 kb and no protein coding capacity (Djebali, et al., 2012). Several of these transcripts, called long non-coding RNAs (lncRNAs), have been shown to have important roles in different biological processes as well as contributing to human diseases such as cancer. In addition, a few studies have identified lncRNAs that are downstream effectors in known cancer related pathways. One example is a lncRNA called *lincRNA-p21*, which is a target of p53, and it has been shown to mediate apoptosis in the p53 pathway (Huarte, et al., 2010). Although there are a handful of these functional lncRNAs identified in cancer, the list is still short and uncomplete.

In this study, we systematically looked for lncRNAs expression alteration in tumours with or without mutations in cancer driver genes. We made use of the massive mutation, expression and copy number data available at TCGA, in 19 cancer types and more than 7000 tumours (Figure 10).

We first obtained the data from TCGA. For the expression profiles, we realigned the RNA-Seq data from 7295 tumours to the human hg19 genome assembly, followed by annotation of all lncRNA and protein coding genes. All the genes were also assigned copy number values using Affymetrix SNP6 data available from TCGA. For the mutation data, based on known cancer driver genes from The Cancer Gene Census (Futreal, et al., 2004; Sondka, et al., 2018), we defined a set of 68 genes that were recurrently mutated in at least two tumour types (Figure 10). Using two-sided Wilcoxon rank sum test, we tested the associations between mutations in these genes and changes in mRNA levels, for each gene individually in each cancer type.

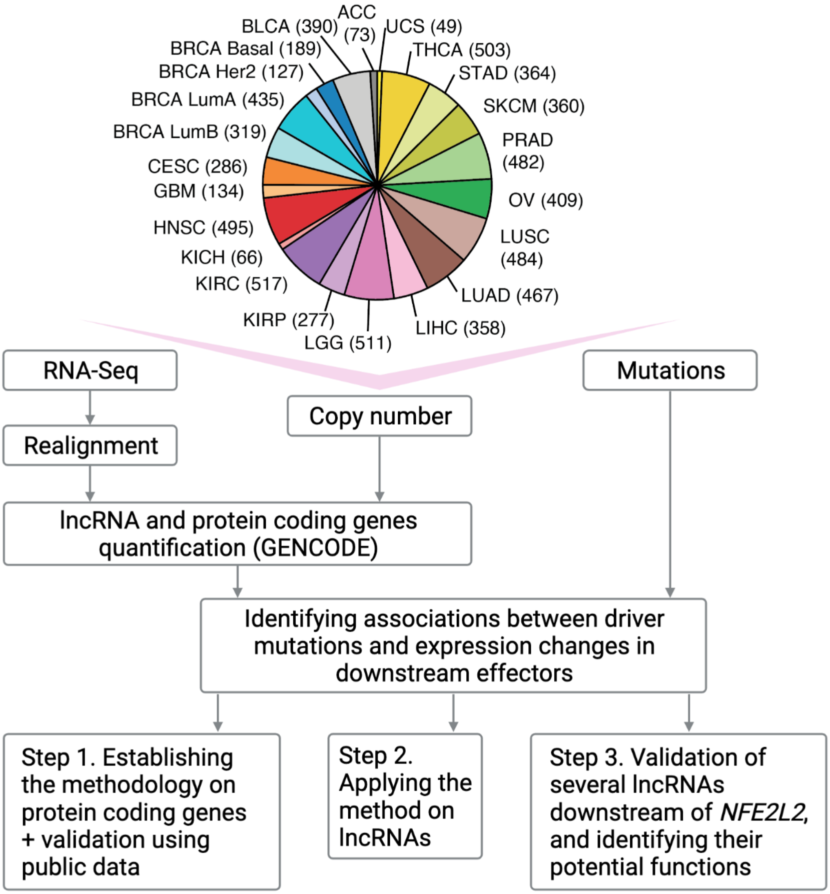
ACC (73)
BLCA (390)
UCS (49)
BRCA Basal (189)
THCA (503)
BRCA Her2 (127)
STAD (364)
BRCA LumA (435)
SKCM (360)
BRCA LumB (319)
PRAD (482)
CESC (286)
GBM (134)
OV (409)
HNSC (495)
LUSC (484)
KICH (66)
KIRC (517)
LUAD (467)
KIRP (277)
LIHC (358)
LGG (511)
RNA-Seq
Mutations
Copy number
Realignment
lncRNA and protein coding genes quantification (GENCODE)
Identifying associations between driver mutations and expression changes in downstream effectors
Step 1. Establishing the methodology on protein coding genes + validation using public data
Step 2. Applying the method on lncRNAs
Step 3. Validation of several lncRNAs downstream of *NFE2L2*, and identifying their potential functions

Figure 10. Study I overview.

We first established our methodology on protein coding genes. This approach resulted in more than 20’000 associations, with many representing indirect effects, for example due to transcriptional subtypes. To overcome this issue, and enrich for significant signals, we hypothesized that the true downstream effector of known cancer driver genes, would show consistent alteration in more than one cancer type. Based on this, we filtered the associations to only include the ones that are repeated in at least two tumour types. As a result, we identified 1121 such associations, involving 978 unique genes. Among these genes, we observed several canonical targets of known drivers, such as *TP53* and *NFE2L2*. We confirmed several downstream factors using publicly available data. In addition, gene set enrichment analysis using molecular signatures database (MSigDB) (Subramanian, et al., 2005) resulted in related pathways downstream of each cancer driver gene.

Having our methodology established on the protein coding genes, we next used the same approach on lncRNAs. In total, we identified 189 associations and 169 unique lncRNAs. Based on data from a recent study identifying lncRNAs that were targeted by p53 in response to DNA damage (Leveille, et al., 2015), we were able to validate several lncRNAs found as associated ones with p53 in our results. This again confirmed that our results are enriched for true targets.

We next focused on *NFE2L2*, the master regulator of oxidative stress, with often activating mutations in various tumour types. We found 15 lncRNAs that were differentially expressed downstream of *NFE2L2* gain-of-function mutations. To validate the responsiveness of these lncRNAs to this factor, we silenced *NFE2L2* in a lung cancer cell line, A549, followed by high coverage RNA-Seq. We were able to confirm the downregulation of 8 out of 15 (11 were expressed in this cells) lncRNAs predicted computationally. We next focused on three of these lncRNAs and performed q-PCR assays to once again confirm their expression changes upon silencing *NFE2L2*.

To check the ability of NRF2 (the protein coded by *NFE2L2*) to control the expression of these lncRNAs, we looked specifically at the promoter region of one of these lncRNAs, called *LINC00942*. Interestingly, we observed an ARE-like element only 120 bps upstream of TSS of this gene. We next used luciferase reporter assays for further analyzing the effect of NRF2 binding. Using vectors containing either the wild-type promoter of *LINC00942,* or with a point mutation in the ARE site (NRF2 binding consensus), we were able to show that binding of NRF2 is necessary for the expression of this lncRNA.

To further assess the function of *LINC00942*, we inhibited the expression of this lncRNA using antisense oligonucleotides (ASOs) in A549 cells. We were able to confirm that both expression and protein levels of GCLC, a well-known NRF2 target that is crucial for synthesis of the antioxidant glutathione, went down upon treatment with *LINC00942* ASOs. Collectively, our results suggest a role for *LINC00942* in the antioxidant response downstream of *NFE2L2*.

In conclusion, in this study we systematically investigated alterations in lncRNA expression in relation to key mutational driver events in human cancers. We provided a comprehensive catalogue of candidate lncRNAs that may play a functional role as part of oncogenic programs and may serve as a reference and starting point for future experimental studies.

3.2 NRF2 TARGETOME (**STUDY II**)

NRF2 is a transcription factor that regulates reactive oxygen species (ROS) levels in cells, by inducing the transcription of a wide range of genes involved in cellular antioxidant response. Therefore, NRF2 has traditionally been recognized as a cancer preventive agent (Jaramillo and Zhang, 2013). Recently, however, several studies have demonstrated that constitutive activation of NRF2 in cancer cells can work in favor of tumours by protecting cells from apoptosis, chemotherapeutic agents, and radiotherapy (Menegon, et al., 2016). This phenomenon that has been described as the 'dark side' of NRF2, has made this factor an interesting gene to study in tumours. Somatic mutations in *NRF2* and *KEAP1*, the protein that regulates the amount of active NRF2 in the cells, have been reported in variety of cancers such as lung squamous cell carcinoma and lung adenocarcinoma (Kerins and Ooi, 2018). Since the recognition of the consensus binding sites for NRF2, also known as antioxidant response elements (AREs), several studies have aimed to identify the downstream targets of this factor, with only a few using a genome-wide approach. In this study, we provide the most comprehensive genome-wide characterization of NRF2 direct targets to date.

We used two lung cancer cell lines, A549 and H838, both harboring activated *NRF2*. For determining the binding sites of NRF2, we performed ChIP experiments using two different antibodies (abcam and Diagenode) targeting NRF2, followed by high coverage DNA sequencing. In parallel, to monitor the expression changes of downstream targets of NRF2, we transfected both cell lines with siRNAs targeting *NRF2* (*n*=3) and control siRNAs (*n*=4) and sequenced the RNA from these cells. In addition, to identify early and late targets of NRF2, we designed a time course experiment in which using the same siRNAs, we extracted the RNA after 3, 6, 12, 24 and 48 hours (Figure 11). By integrating these data, we defined the NRF2 targetome.

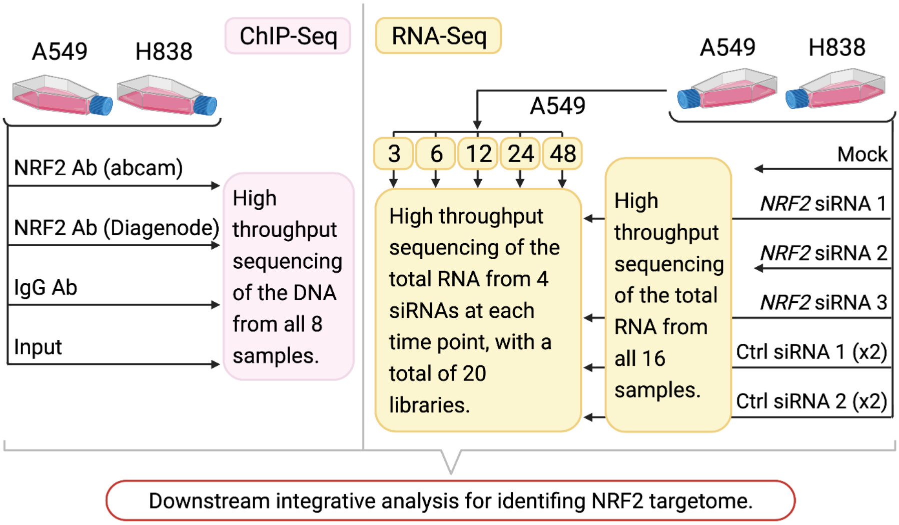
A549
H838
ChIP-Seq
NRF2 Ab (abcam)
NRF2 Ab (Diagenode)
IgG Ab
Input
High throughput sequencing of the DNA from all 8 samples.
RNA-Seq
A549
H838
A549
3
6
12
24
48
High throughput sequencing of the total RNA from 4 siRNAs at each time point, with a total of 20 libraries.
High throughput sequencing of the total RNA from all 16 samples.
Mock
NRF2 siRNA 1
NRF2 siRNA 2
NRF2 siRNA 3
Ctrl siRNA 1 (x2)
Ctrl siRNA 2 (x2)
Downstream integrative analysis for identifing NRF2 targetome.

Figure 11. Study II overview. Figure generated by BioRender.com.

First, we identified NRF2 binding sites in both cell lines, based on the ChIP-Seq data and the MACS tool for peak-calling (Zhang, et al., 2008). Using different filters on the number of mapped reads in each region, and enrichment values for each peak, we generated a map of NRF2 sites, although number of peaks varied a lot between different samples. Peaks were spread throughout the genome, and only about 8% of all peaks resided in gene promoter regions. To identify the true binding sites of NRF2 and not the false positives, we made use of the expression data. By combining ChIP-Seq and RNA-Seq data, we could confirm which binding sites were true positives, by selecting for peaks with the closest gene being differentially expressed after *NRF2* silencing. Using these stringent criteria, we were able to identify 87 genes that were directly bound by NRF2. Among these, most of the known direct targets of NRF2 were identified, in addition to several novel targets. Interestingly, the NRF2 binding sites were not always in proximal distance to the genes, nor in the promoter regions. We identified several genes that were regulated by NRF2 binding more than 10 kb away from the transcription start site (TSS).

We next explored the time course data, with the aim to identify early and late targets of NRF2. We could see a clear pattern where the expression changes increased over time, starting already six hours after silencing *NRF2*. We defined several early and late responders including both known and novel targets of NRF2.

In conclusion, in this study we have provided a comprehensive map of the NRF2 targetome that includes several novel targets. We also showed that by combining ChIP-Seq and RNA-Seq data, it is more probable to identify the true positive binding sites of NRF2. Using a genome-wide approach, we were able to detect binding of NRF2 in distal locations. These discoveries would not have been achievable without the high throughput sequencing approach following the ChIP experiment.

3.3 CELLULAR RESPONSE TO ALK INHIBITORS (**STUDY III**)

The high frequency of *ALK* mutations in primary and relapsed neuroblastoma tumours has made it an interesting target in these tumours (Martinsson, et al., 2011). Several ALK tyrosine kinase inhibitors (TKIs) have been developed to date, with a few of them being approved by regulatory authorities, including the U.S FDA and European EMA, for the treatment of non-small-cell lung cancer patients, but none are yet approved for use in neuroblastoma treatment.

In this study, we investigated both the signaling and the transcriptional changes in neuroblastoma cells treated with marketed ALK tyrosine kinase inhibitors (TKIs). The three cell lines used in this study, CLB-GE, CLB-BAR and SK-N-AS are all of neuroblastoma origin. CLB-BAR (gain of function, truncated *ALK*) and CLB-GE (gain of function *ALK* mutation) are *ALK*-addicted cell lines. As a negative control we used the SK-N-AS cells that do not express *ALK* and therefore do not depend on *ALK* for survival, instead they harbor activating mutations in *NRAS*.

For the phosphoproteomic experiments, all cells were treated with either 250 nM crizotinib or 30 nM lorlatinib, which are first and third generation ALK TKIs (Johnson, et al., 2014; Shaw, et al., 2013), for one hour (Figure 12). To identify the correlation between phosphorylated proteins and ALK activity, we used an LC-MS/MS (liquid chromatography-tandem mass spectrometry) approach to detect phosphorylation changes in both ALK itself and ALK target proteins. Upon treatment with each drug, thousands of proteins were phosphorylated or dephosphorylated at the phosphorylatable amino acids tyrosine, serine or threonine. The phosphorylation signal intensities within the ALK protein were decreased after one hour of treatment with both drugs in both CLB-GE and CLB-BAR cells, with crizotinib treatment being more specific for ALK. Lorlatinib treatment also resulted in reduced phosphorylation of two other receptor tyrosine kinases, DDR1 and DDR2 only in the ALK-addicted cells, with no altered phosphorylation detectable in the control cell line, suggesting that the effect of lorlatinib is through ALK.

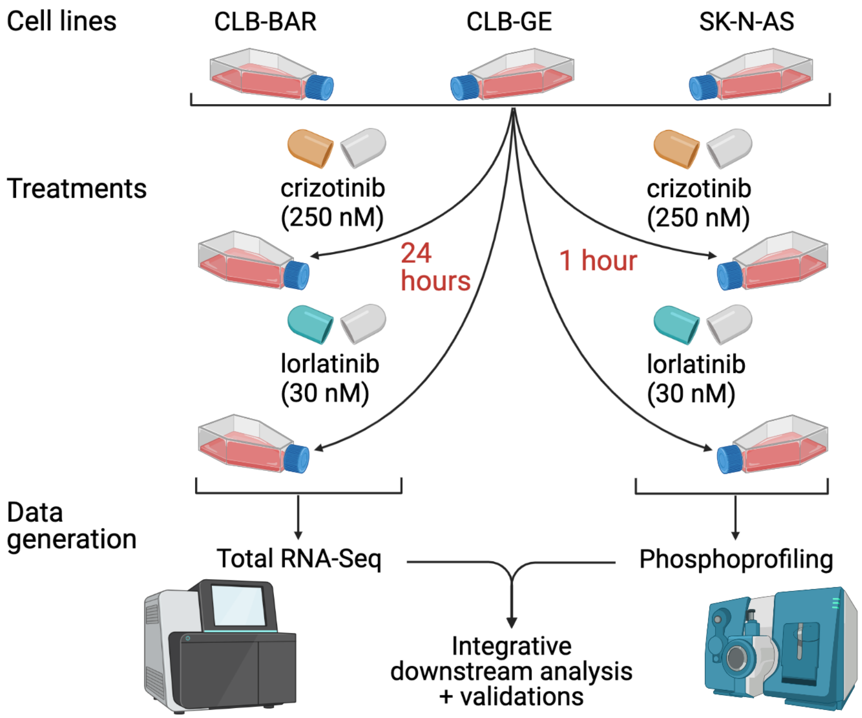

Figure 12. Study III overview. Figure generated by BioRender.com.

In total, in CLB-BAR cells we detected 74 proteins with decreased phosphorylation upon treatment with either or both drugs. Gene set enrichment analysis (GSEA) of these proteins showed an enrichment of several signaling pathways including the FGFR and INSR pathways, while the same analysis with proteins that were hyperphosphorylated after drug treatments showed no enrichment, suggesting non-specific effects on increased phosphorylation. In CLB-GE, only 23 of the proteins identified in the CLB-BAR cells showed decreased phosphorylation, and the GSEA on these genes showed a weak enrichment of the pathways identified for CLB-BAR cells. Lastly in the SK-N-AS cells only one protein (AKT3) was observed with lower phosphorylation after both treatments.

In parallel, we also performed RNA-Seq to identify downstream genes that are regulated by ALK. For this aim, all three cell lines were treated with each drug for 24 hours, and the total RNA was extracted and sent for deep sequencing (Figure 12). The response to both drugs was similar, with 302 down- and 462 up- regulated genes upon treatment with either of the drugs in CLB-BAR cells. As expression of the *ALK* gene itself was also increased in CLB-BAR cells, and not in the other two cell lines, the downstream expression changes were also much more significant in CLB-BAR. Interestingly, we identified 53 transcription factors among differentially expressed genes, which were then analyzed further.

To identify the main ALK-dependent response genes and proteins we performed an integrative biological network analysis based on known protein-protein interactions (PPIs) using the 74 proteins with decreased phosphorylation after drug treatment and the 53 differentially expressed transcription factors identified from RNA-Seq after drug treatment. This integrative analysis resulted in the identification of several known and novel factors and pathways, for example ETS family of transcription factors, as well as MAPK phosphatase *DUSP4* (also known as *MKP2*). Many of these findings were validated further using several experimental methods.

In conclusion, in this study we present a comprehensive map of downstream molecular changes in neuroblastoma cell lines, upon treatment with first- and third- generation *ALK* TKIs. We also show that although we use two *ALK*-addicted cell lines, they respond differently to these drugs, possibly due to the fact that they exhibit different chromosomal rearrangement profiles. This again emphasizes the fact that neuroblastoma is a complex heterogeneous disease and there is a need for further investigation of ALK signaling to find suitable drugs for patients with this type of tumour.

4 CONCLUSIONS

Cancer is a disease of the genome. Transformation of normal cells to tumour cells, or carcinogenesis, involves the accumulation of driver mutations that cause physiological changes, followed by gained selective advantages that are preferentially selected by the tumour environment. In this thesis, we aim to understand the mechanism by which these mutations drive the cells towards a cancerous phenotype. We use different approaches to understand this phenomenon in the three studies included here.

In Study I, we systematically investigated alterations in expression of lncRNAs in human tumours with key driver mutations. We provided a comprehensive map of lncRNAs as possible downstream effectors in key oncogenic programs. We also confirmed that some of these lncRNAs have a role in the oxidative stress response pathways in lung cancer cells. This broad list of lncRNAs may serve as a starting point for future studies.

In study II, we provided a comprehensive map of the NRF2 targetome. NRF2 is a transcription factor that is known as the master regulator of oxidative stress and is activated in several cancer types. Here we found several novel targets of NRF2 with both proximal and distal NRF2 binding sites. The dataset generated in this study constitute an important resource that will facilitate our understanding of NRF2 and its complex roles in cancer.

In study III, we went further than studying tumours genomic changes, by using marketed oncology drugs that inhibit the cancer driver gene *ALK*, followed by studying the transcriptomic and proteomic alterations upon treatment with the ALK TKIs. Here we identified relevant biomarkers and signaling pathways such as ETS family transcription factors and the MAPK phosphatase DUSP4 as targets of ALK signaling. This study reveals new targets that could be exploited to treat ALK-positive neuroblastoma.

5 ACKNOWLEDGEMENTS

My journey started in the summer of 2010 when I first visited Skövde (a small city in Sweden). That city has brought so much into my life! That summer I visited Sweden to meet my brother, we ended up in a trip around Europe, and that was probably the time I decided to stay! Earlier that year I had applied to a master program in Skövde, and that summer I travelled there for the first time. After providing some more documents and an IELTS exam a few months later, I finally got accepted to the Molecular Biology program. I moved to Sweden the next semester, in January 2011, and that is where my story starts.

Immigration is always hard, but even harder when done alone. Therefore, this acknowledgement is not only to the people that helped me during my PhD years, but to all of those who made my life better and made Sweden my home since 2011.

Erik. First of all, I would like to thank you for believing in me. You gave me the opportunity to learn, when I needed it the most. When I joined the group, more than seven years ago, I had a slight knowledge about bioinformatics and programming, but you agreed to let me learn, first as a project student for a year, and then finally, after I insisted for a long time, as a PhD student! Thank you for being patient all these years, specially when I went on parental leave, two times! and when my project took a really long time to finish (I hope the paper is submitted by the time you read this!). I have learnt a lot from you, from making figures, to writing papers and most importantly to do good and honest research. Seven years is a long time, you are like family now, and I hope we keep in touch even after I leave the group.

Babak. There is no doubt that without you, I would have been lost in the world of programming! You thought me so much the first few years when you were still in the group. I remember one day I asked you ‘what is an index?!’. You patiently explained all the basic stuff about MATLAB and Shell scripting to me, and I am truly thankful for that.

We also became good friends and you introduced me to many other Persian people, who have also become my friends during the years, and it helped me feel much more at home.

Johan. Thank you for all those lunch breaks, when all other people in the lab would go to Lyktan to eat and only we had our food with us! I learned a lot from you, specially about how to have a work/family balance, and how it is to have kids and what they are like!

Joanna. It is maybe time to tell the story of how I was introduced to Erik! Once upon a time, in a PhD dissertation party of someone called 'James', I met Joanna, and a few weeks after we went out for a coffee with our families. When I mentioned I was trying to learn bioinformatics by taking some courses and would like to find a job soon, Joanna told me that there is a PI at Sahlgrenska that has a small group and he only does bioinformatics. She was a part time post doc at Erik's group at that time. A few weeks later, there I was, at Erik's office! And he generously offered me a project with scholarship! And I joined Erik's group on the 3rd of April 2014. Here by, I thank you so much Joanna for introducing me to Erik! We have also become friends, and finding good friends is the most important of all!

Joakim. It was really nice to work with you during your PhD years. I always admired your hard work and wished I could work that good. Those few times we went out for afterwork with the group were also really fun! You have a good humor!

Jimmy. You are one of the most hard-working people I have ever met. During all those years you were in the group, I learned a lot from you, from basic biology concepts, to statistics, and more. Thanks for all your support during the years when I needed it (specially at occasions similar to the EMBL conference in 2015!). Also, thanks for the two great papers we have published together! I hope to be able to work with you again one day.

Kerryn. When you first joined the group, I thought to myself 'finally there is another girl joining the group!'. But it became much more than that, we became friends and spent so much time together both in the lab, and outside with our families. Thank you for all your help during the last five years. You were always there for me, no matter what problem or concern I had. I could talk to you about anything, and you would just listen. Specially during the last few months when I had lots of stress, you always listened to my complaints, and I felt so much better after talking about them. Thank you for all the proof-readings of my manuscript and thesis. I couldn't have done it without your help, and I hope to be able to return the favor sometime soon.

Swaraj. You are definitely one of the most fun people I have worked with. When it comes to work, you always knew the answer to my questions, and always knew of a tool or software that could complete the task. And when it came to out of office fun, you were always in for any activity, and your famous jokes and stories where the funniest moments of all! Thank you for all the good memories you made for me. It is sad that you couldn't be here for my dissertation, but I hope you can visit Gothenburg sometime soon.

Susanne. I always enjoyed your compony, even though you were not in the group for so long. During the last few months of your work when you sat in our lab, we talked a lot, and about everything! And it always feels good to have someone that understands! Thank you for all those days. Although I was sad that you left, I am really happy you found a job you like more!

Markus. Thank you for always being helpful! After Babak left the lab, I was worried that I would be left alone with all my bioinformatics / server related questions! But thanks to you that wasn't the case. You always took the time to answer all my questions and helped me a lot. I hope to be able to help you somehow in the next few months when you will be finishing up your PhD!

Martin. It has been really fun working with you during the last four years. Thanks for always answering my random R questions and taking the time to solve my MacFuse issues, and later on sshfs problems over and over again! You also have many other talents that we discovered later on! So, I have to ask you, could you please draw a portrait of me during my dissertation?

Arman. It was really nice to get to know you and have someone to talk Farsi to. We could talk about everything (and everyone), and that is rare! It was sad that you left the group, but I wish you all the best in your new job!

Vinod. Although you have been in the lab for more than a year, we haven't met so many times. But even those few times I really enjoyed your compony. And recently we have had a lot of discussions about our common interest, babies! As you and your wife had one on the way, and I have two at home! By the time you read this, you are already a father, so congratulations, and best of luck!

Isabella. I have basically known you for many years, as your previous office was so close to the coffee machine upstairs, and you talked to most people there! But I got to know you more since you joined our group recently, and it has been truly fun! We have been having long talks about everything in life, and that is always nice!

Emma. It has been really nice to get to know you. Thank you for listening to all my complaints during the last few months when I was really stressed while writing my thesis. Sometimes all you need is a good labmate that listens!

Alireza. We have known each other for a few years now, and I can certainly say that I only remember good things from you. It has always been nice to talk to you, as we have many things in common, specially our home country! It has been even nicer to have you in our group, sitting close by, so we can chat even more!

Thanks to all the other people that were in the LarssonLab for a short time during the last seven years, including **Niklas**, **Karl**, **Alejandro**, **Harsha**, and **Rada**, it has been nice knowing you all.

Katrin, thank you for all these years, and specially the last few months when we were both writing our theses! It was really nice to know that I am not alone in this, and someone else is going through the same pain with me! But it is all done now! We did it!

Josephine, Mahmood, Andranik, Martin, Liam, Anders and others next door in the Clausen lab, thank you all for the many hundreds lunch breaks during the last few years. It will be really hard to beat those days, when we were almost 10 people sitting around one little table eating lunch! I miss those days a lot!

Thanks to all the people in the **KanduriLab** for all their help during these years, specially **Chandra** as my co-supervisor. A special thanks to **Gendy** for all the help with setting up the lab, and for teaching me how to do human cell culture and qPCR among many other things. Your help was truly appreciated those days. **Sanhita** and **Tanmoy**, you helped me a lot those days when I was new here. Thanks for always answering my questions about different machines and reagents to use for my experiments. **Santhilal**, thanks for all the times I had bioinformatics questions and you helped me, specially during those months you sat in our office!

Bengt, **Ganesh**, and others on the first floor, thank you all for the collaboration on the ALK paper we have together. It was an amazing experience.

Laleh. It has been really nice to know you since the collaboration we had many years ago. Now you work even closer and it is really nice to chat whenever we meet! It is somehow easier to become friends when having the same mother tongue!

A special thanks to all the other people at the department that made my everyday life at work pleasant since 2014.

Markus Tamás. You are certainly one of the kindest people I have ever met. I joined your lab in November 2012, right after finishing my master's thesis with not a good experience. I learned for the first time in your group how to do true and honest research, and how my small contributions in projects are valued! Even though I couldn't start a PhD in your group at that time, I am really glad that we met, and we have two amazing papers together. Thanks for all you have thought me during those few but fruitful months.

Arefeh. Working with you has been the nicest collaboration I have ever had. You are extremely good at what you do, and know the details of all the work that has been performed in your group. Thanks for believing in me and my work, and for the two great papers I was lucky enough to be part of. Hope to meet you sometime soon, on the other side of the world!

Yogi moms. I would like to thank you all for being there for me when I needed good friends! It is really hard to immigrate and live far from family, and have kids! But you all have helped me a lot during the last four years, and made it all so much easier to handle and made Gothenburg a better place to live in! I am really thankful for that and I love you all!

Araz. You have always been more than just a brother to me. You are my role model and I always looked up to you since I was a kid. Since early school years, you always helped me, and it was truly amazing to have you around all the time. It was really hard when you left home to come to Sweden, so I had to follow you here as well! You pushed me so hard to apply, to study and to stay here. Although, sadly, you didn’t stay here when I came, all your supports during my first few years in Sweden was the only reason I could survive. Looking back, I think it is certain that without you, I wouldn’t have been where I am today. Thank you so much for all those years, and I am looking forward to live close to you again real soon. Dooset daram dadashe azizam!

Maman and **baba.** Thank you for helping me become the person I am today. Without your help and support I wouldn’t have been here. Thank you for letting me choose my own path at all steps in life (and sorry for never listening to you!). But I think you also agree now that I made quite alright choices throughout the years. Although the last one year has been really tough and I am truly sad that you are not here now, I am just happy to be able to have your support anytime I need it.

Maman. You are definitely the strongest and most successful woman I know. Growing up, I always saw you working hard, and I realize now how much I have learnt from you just by seeing that. Thank you for teaching me to be a strong person and woman, and to always encourage me to have a family in parallel to a good career. That is definitely the most important lesson in life! Dooset daram mamane azizam!

Baba. I definitely took most things after you. I have a strong mind and my own judgement on things without being affected by others' opinion, and that has helped me a lot during these years. I also learnt the most from you when it comes to literature and poetry, and that has become my favorite hobby recently. Thank you for always encouraging me to read by getting me lots of books. It has all paid off now, I just wrote my first little children book last year, and that is all because of your influence throughout the years. Dooset daram babaye azizam!

Dayi Kevin. You are definitely the most important person when it comes to feeling like at home in Gothenburg! Your home doors have always been open to me, and I am truly thankful for that. Whenever I had nowhere else to go, I could always come to you. Thank you for all your help and support since I moved to Sweden. You still help us a lot, and we don't know what we would have done without you!

Dayi Martin. Although we haven't met as much during these years, you would always be there if I needed any help. Specially during the year I lived in Linköping. Thank you for all the help and support, even though I know you are so modest, you wouldn't even agree!

Rest of my family in Iran, Canada and Sweden. Honestly it would be really hard to name the rest of my family here, as I have many more! In case you are wondering, I have 9 aunts, 7 uncles and more than 30 cousins (I should count them one day!), not even including in-laws in the numbers! All are spread around the world but mainly in Iran, Canada and Sweden. So here by, I would like to thank you all for being there when I needed you, and for showing me that 'family' is everything! I love you all, miss you so much and hope to see you soon!

خانواده‌ی عزیزم در ایران و جهان !
از همه‌ی شما برای تمامِی این سال‌ها سپاس‌گزارم. ممنون که همیشه هستید و همیشه به
من ثابت می‌کنید که خانواده به‌ترین حامی در دنیاست !
همه‌ی شما را دوست دارم و دلتنگ‌تان هستم.

All the others! If you are reading this book, you mean a lot to me. I cannot possibly name all my friends here, but I thank you all for making my life better in one way or another during these 10 years!

Liam and Lavin. My beautiful little kids, and the real achievements of my PhD years! You are still too young to read this fully, but Liam can read some words by now, and I will read the rest for you! I dedicate this book to you, and want you to come back to it when you are older, to understand what you meant to me during the hardest years of my life. You are the light of my life, and the reason I live.

Liam, having you has been the most amazing part of my life. You thought me to smile during my hardest days, and you have made me a better person in many different ways. You always want to make jokes and you make me laugh a lot! The best part of my day is when we just sit and talk about everything and imagine the impossible together! I am sorry for not being around so much recently, but I will soon be a full-time mom again! I am so proud of you and I am thankful every day for being your mom. Asheghetam pesare ghashangam!

Lavin, my life got complete the day you were born. Even though the past year was really tough in many ways, having you and being home with you was so amazing and definitely the best part of it all! You are really active and also strong minded just like me, and I love that about you. I am sorry for not spending so much time with you the last few months, but it will get much better soon, and I am so looking forward to teaching you everything and to see you grow into an interesting person! I know that one day you will be a great woman, but you will always be my little girl! Asheghtam dokhtare nazam!

Jonathan. Where to start! I have known you for more than 10 years now, as we met just one month after I moved to Sweden! And it will soon be 5 years since we got married. So that should say it all! It is not easy to thank you for all these years in a few sentences, but I will try! Thank you for becoming my family, when I had no one else, and for being the reason I stayed in Sweden when nothing else was working out for me! Thanks for the two amazing kids we have together, and for taking care of them when I had so much work to do. During the last few months, and specifically those 6 intense weeks when I wrote my entire thesis, you did everything at home and took care of the kids every day and night and even weekends, so that I could finish my thesis. You even helped me with proof-reading of my thesis after kids went to bed! It is beyond a shadow of a doubt that I couldn't have finished my PhD without your help and support. It has been a really hard year, without any help from our family and friends since the pandemic started, but we managed it together, and we will always and forever. I am sorry you had to deal with a stressed and bad-tempered me for many months! But it is all over, and we can enjoy life much more from now on! I love you so much, and I am looking forward to the rest of our lives together!

If you have read my thesis, all the way through,
do not forget to find, my 'hidden alien' too!

6 REFERENCES

1. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 2008;455(7216):1061-1068.
2. Adams, M.D.*, et al.* Complementary DNA sequencing: expressed sequence tags and human genome project. *Science (New York, N.Y.)* 1991;252(5013):1651-1656.
3. Adamson, E.D. Oncogenes in development. *Development* 1987;99(4):449-471.
4. Anderson, S. Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic acids research* 1981;9(13):3015-3027.
5. Arteaga, C.L. Epidermal growth factor receptor dependence in human tumors: more than just expression? *Oncologist* 2002;7 Suppl 4:31-39.
6. Ashouri, A.*, et al.* Pan-cancer transcriptomic analysis associates long non-coding RNAs with key mutational driver events. *Nature Communications* 2016;0.
7. Auton, A.*, et al.* A global reference for human genetic variation. *Nature* 2015;526(7571):68-74.
8. Baker, S.G. A cancer theory kerfuffle can lead to new lines of research. *J Natl Cancer Inst* 2015;107(2).
9. Balas, M.M. and Johnson, A.M. Exploring the mechanisms behind long non-coding RNAs and cancer. *Non-coding RNA Res* 2018;3(3):108-117.
10. Berget, S.M., Moore, C. and Sharp, P.A. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences of the United States of America* 1977;74(8):3171-3175.
11. Beroukhim, R.*, et al.* Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proceedings of the National Academy of Sciences of the United States of America* 2007;104(50):20007-20012.
12. Birben, E.*, et al.* Oxidative stress and antioxidant defense. *World Allergy Organ J* 2012;5(1):9-19.
13. Birney, E.*, et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007;447(7146):799-816.

14. Bister, K. Discovery of oncogenes: The advent of molecular cancer research. *Proceedings of the National Academy of Sciences of the United States of America* 2015;112(50):15259-15260.

15. Brannan, C.I.*, et al.* The product of the H19 gene may function as an RNA. *Molecular and cellular biology* 1990;10(1):28-36.

16. Braslavsky, I.*, et al.* Sequence information can be obtained from single DNA molecules. *Proceedings of the National Academy of Sciences of the United States of America* 2003;100(7):3960-3964.

17. Brown, C.J.*, et al.* A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature* 1991;349(6304):38-44.

18. Brown, C.J.*, et al.* The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell* 1992;71(3):527-542.

19. Busch, H.*, et al.* SnRNAs, SnRNPs, and RNA processing. *Annu Rev Biochem* 1982;51:617-654.

20. Bussemakers, M.J.*, et al.* DD3: a new prostate-specific gene, highly overexpressed in prostate cancer. *Cancer research* 1999;59(23):5975-5979.

21. Carpenter, R.L. and Gökmen-Polar, Y. HSF1 as a Cancer Biomarker and Therapeutic Target. *Curr Cancer Drug Targets* 2019;19(7):515-524.

22. Casamassimi, A.*, et al.* Transcriptome Profiling in Human Diseases: New Advances and Perspectives. *Int J Mol Sci* 2017;18(8).

23. Cavenee, W.K.*, et al.* Prediction of familial predisposition to retinoblastoma. *N Engl J Med* 1986;314(19):1201-1207.

24. Chamary, J.V., Parmley, J.L. and Hurst, L.D. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nature reviews. Genetics* 2006;7(2):98-108.

25. Chen, Q., Kang, J. and Fu, C. The independence of and associations among apoptosis, autophagy, and necrosis. *Signal Transduct Target Ther* 2018;3:18.

26. Chen, Y.*, et al.* Oncogenic mutations of ALK kinase in neuroblastoma. *Nature* 2008;455(7215):971-974.

27. Chial, H. Tumor Suppressor (TS) Genes and the Two-Hit Hypothesis. *Nature Education* 2008;1(1):177.

28. Chinnery, P.F.*, et al.* Accumulation of mitochondrial DNA mutations in ageing, cancer, and mitochondrial disease: is there a common mechanism? *Lancet* 2002;360(9342):1323-1325.

29. Chow, L.T.*, et al.* An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* 1977;12(1):1-8.

30. Chu, A.*, et al.* Large-scale profiling of microRNAs for The Cancer Genome Atlas. *Nucleic acids research* 2016;44(1):e3.

31. Chu, D. and Wei, L. Nonsynonymous, synonymous and nonsense mutations in human cancer-related genes undergo stronger purifying selections than expectation. *BMC Cancer* 2019;19(1):359.

32. Colaprico, A.*, et al.* Interpreting pathways to discover cancer driver genes with Moonlight. *Nature Communications* 2020;11(1):69.

33. Cooper, G.M. The Cell: A Molecular Approach. 2nd edition. *Sunderland (MA)* 2000.

34. Dahm, R. Friedrich Miescher and the discovery of DNA. *Dev Biol* 2005;278(2):274-288.

35. Darnell, J.E., Jr. Transcription factors as targets for cancer therapy. *Nature reviews. Cancer* 2002;2(10):740-749.

36. Daxinger, L. and Whitelaw, E. Understanding transgenerational epigenetic inheritance via the gametes in mammals. *Nature reviews. Genetics* 2012;13(3):153-162.

37. Deamer, D. Nanopore analysis of nucleic acids bound to exonucleases and polymerases. *Annu Rev Biophys* 2010;39:79-90.

38. Deininger, P. Genetic instability in cancer: caretaker and gatekeeper genes. *Ochsner J* 1999;1(4):206-209.

39. Dey, B.*, et al.* DNA-protein interactions: methods for detection and analysis. *Mol Cell Biochem* 2012;365(1-2):279-299.

40. Diederichs, S. The four dimensions of non-coding RNA conservation. *Trends Genet* 2014;30(4):121-123.

41. Dietlein, F.*, et al.* Identification of cancer driver genes based on nucleotide context. *Nature genetics* 2020;52(2):208-218.

42. Dinkova-Kostova, A.T., Holtzclaw, W.D. and Kensler, T.W. The role of Keap1 in cellular protective responses. *Chem Res Toxicol* 2005;18(12):1779-1791.

43. Djebali, S.*, et al.* Landscape of transcription in human cells. *Nature* 2012;489(7414):101-108.

44. Eddy, S.R. Non-coding RNA genes and the modern RNA world. *Nature reviews. Genetics* 2001;2(12):919-929.

45. Eid, J.*, et al.* Real-time DNA sequencing from single polymerase molecules. *Science (New York, N.Y.)* 2009;323(5910):133-138.

46. Eldra Solomon, C.M., Diana W. Martin, Linda R. Berg. Biology, 11th Edition. 2019.

47. Feng, J., *et al.* The RNA component of human telomerase. *Science (New York, N.Y.)* 1995;269(5228):1236-1241.

48. Fiers, W., *et al.* Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* 1976;260(5551):500-507.

49. Flockhart, R.J., *et al.* BRAFV600E remodels the melanocyte transcriptome and induces BANCR to regulate melanoma cell migration. *Genome research* 2012;22(6):1006-1014.

50. Fulda, S., *et al.* Cellular stress responses: cell survival and cell death. *Int J Cell Biol* 2010;2010:214074.

51. Futreal, P.A., *et al.* A census of human cancer genes. *Nature reviews. Cancer* 2004;4(3):177-183.

52. Gilmour, D.S. and Lis, J.T. Detecting protein-DNA interactions in vivo: distribution of RNA polymerase on specific bacterial genes. *Proceedings of the National Academy of Sciences of the United States of America* 1984;81(14):4275-4279.

53. Greaves, M. and Maley, C.C. Clonal evolution in cancer. *Nature* 2012;481(7381):306-313.

54. Greenleaf, W.J. and Sidow, A. The future of sequencing: convergence of intelligent design and market Darwinism. *Genome biology* 2014;15(3):303.

55. Guan, J., *et al.* Clinical response of the novel activating ALK-I1171T mutation in neuroblastoma to the ALK inhibitor ceritinib. *Cold Spring Harb Mol Case Stud* 2018;4(4).

56. Hanahan, D. and Weinberg, R.A. The hallmarks of cancer. *Cell* 2000;100(1):57-70.

57. Hanahan, D. and Weinberg, R.A. Hallmarks of cancer: the next generation. *Cell* 2011;144(5):646-674.

58. Heath, J.A., *et al.* Good clinical response to alectinib, a second generation ALK inhibitor, in refractory neuroblastoma. *Pediatr Blood Cancer* 2018;65(7):e27055.

59. Heather, J.M. and Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* 2016;107(1):1-8.

60. Heid, C.A., *et al.* Real time quantitative PCR. *Genome research* 1996;6(10):986-994.

61. Heldin, C.H. and Westermark, B. Mechanism of action and in vivo role of platelet-derived growth factor. *Physiol Rev* 1999;79(4):1283-1316.

62. Hofseth, L.J., Hussain, S.P. and Harris, C.C. p53: 25 years after its discovery. *Trends Pharmacol Sci* 2004;25(4):177-181.

63. Holla, V.R.*, et al.* ALK: a tyrosine kinase target for cancer therapy. *Cold Spring Harb Mol Case Stud* 2017;3(1):a001115.

64. Holley, R.W.*, et al.* STRUCTURE OF A RIBONUCLEIC ACID. *Science (New York, N.Y.)* 1965;147(3664):1462-1465.

65. Holmström, K.M.*, et al.* Nrf2 impacts cellular bioenergetics by controlling substrate availability for mitochondrial respiration. *Biol Open* 2013;2(8):761-770.

66. Hon, C.C.*, et al.* An atlas of human long non-coding RNAs with accurate 5' ends. *Nature* 2017;543(7644):199-204.

67. Huarte, M. The emerging role of lncRNAs in cancer. *Nat Med* 2015;21(11):1253-1261.

68. Huarte, M.*, et al.* A large intergenic non-coding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* 2010;142(3):409-419.

69. Hung, T.*, et al.* Extensive and coordinated transcription of non-coding RNAs within cell-cycle promoters. *Nature genetics* 2011;43(7):621-629.

70. Hutter, C. and Zenklusen, J.C. The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. *Cell* 2018;173(2):283-285.

71. Ishikawa, K.*, et al.* ROS-generating mitochondrial DNA mutations can regulate tumor cell metastasis. *Science (New York, N.Y.)* 2008;320(5876):661-664.

72. Iyer, M.K.*, et al.* The landscape of long non-coding RNAs in the human transcriptome. *Nature genetics* 2015;47(3):199-208.

73. Jackson, S.P. and Bartek, J. The DNA-damage response in human biology and disease. *Nature* 2009;461(7267):1071-1078.

74. Jaramillo, M.C. and Zhang, D.D. The emerging role of the Nrf2-Keap1 signaling pathway in cancer. *Genes & development* 2013;27(20):2179-2191.

75. Ji, P.*, et al.* MALAT-1, a novel non-coding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* 2003;22(39):8031-8041.

76. Jiménez-Morales, S.*, et al.* Overview of mitochondrial germline variants and mutations in human disease: Focus on breast cancer (Review). *Int J Oncol* 2018;53(3):923-936.

77. Johnson, T.W., *et al.* Discovery of (10R)-7-amino-12-fluoro-2,10,16-trimethyl-15-oxo-10,15,16,17-tetrahydro-2H-8,4-(metheno)pyrazolo[4,3-h][2,5,11]-benzoxadiazacyclotetradecine-3-carbonitrile (PF-06463922), a macrocyclic inhibitor of anaplastic lymphoma kinase (ALK) and c-ros oncogene 1 (ROS1) with preclinical brain exposure and broad-spectrum potency against ALK-resistant mutations. *J Med Chem* 2014;57(11):4720-4744.

78. Johnsson, P., *et al.* Evolutionary conservation of long non-coding RNAs; sequence, structure, function. *Biochimica et biophysica acta* 2014;1840(3):1063-1071.

79. Jopling, C.L. Stop that nonsense! *Elife* 2014;3:e04300.

80. Kandoth, C., *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* 2013;502(7471):333-339.

81. Kaziro, Y., *et al.* Structure and function of signal-transducing GTP-binding proteins. *Annu Rev Biochem* 1991;60:349-400.

82. Kennedy, D., *et al.* Regulation of apoptosis by heat shock proteins. *IUBMB Life* 2014;66(5):327-338.

83. Kerins, M.J. and Ooi, A. A catalogue of somatic NRF2 gain-of-function mutations in cancer. *Sci Rep* 2018;8(1):12846.

84. Kim, T., *et al.* MYC-repressed long non-coding RNAs antagonize MYC-induced cell proliferation and cell cycle progression. *Oncotarget* 2015;6(22):18780-18789.

85. Kimura, M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* 1977;267(5608):275-276.

86. Kinzler, K.W. and Vogelstein, B. Lessons from hereditary colorectal cancer. *Cell* 1996;87(2):159-170.

87. Kirkman, H.N., *et al.* Mechanisms of protection of catalase by NADPH. Kinetics and stoichiometry. *The Journal of biological chemistry* 1999;274(20):13908-13914.

88. Knudson, A.G., Jr. Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences of the United States of America* 1971;68(4):820-823.

89. Kwak, M.K., *et al.* Enhanced expression of the transcription factor Nrf2 by cancer chemopreventive agents: role of antioxidant response element-like sequences in the nrf2 promoter. *Molecular and cellular biology* 2002;22(9):2883-2892.

90. Lacher, S.E. and Slattery, M. Gene regulatory effects of disease-associated variation in the NRF2 network. *Curr Opin Toxicol* 2016;1:71-79.

91. Lander, E.S., *et al.* Initial sequencing and analysis of the human genome. *Nature* 2001;409(6822):860-921.

92. Lawrence, M.S., *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 2013;499(7457):214-218.

93. Lee, S. and Hu, L. Nrf2 activation through the inhibition of Keap1-Nrf2 protein-protein interaction. *Med Chem Res* 2020;29(5):846-867.

94. Leveille, N., *et al.* Genome-wide profiling of p53-regulated enhancer RNAs uncovers a subset of enhancers controlled by a lncRNA. *Nature Communications* 2015;6:6520.

95. Li, J. and Liu, C. Coding or Non-coding, the Converging Concepts of RNAs. *Front Genet* 2019;10:496.

96. Li, Y., Paonessa, J.D. and Zhang, Y. Mechanism of chemical activation of Nrf2. *PloS one* 2012;7(4):e35122.

97. Liu, S.J., *et al.* Single-cell analysis of long non-coding RNAs in the developing human neocortex. *Genome biology* 2016;17:67.

98. Lockhart, D.J. and Winzeler, E.A. Genomics, gene expression and DNA arrays. *Nature* 2000;405(6788):827-836.

99. Lodish H, B.A., Zipursky SL, et al. Molecular Cell Biology. 4th edition.; 2000.

100. Loewe, L. Genetic mutation. . *Nature Education* 2008;1(1):113.

101. Lord, C.J. and Ashworth, A. The DNA damage response and cancer therapy. *Nature* 2012;481(7381):287-294.

102. Ma, Q., *et al.* Induction of murine NAD(P)H:quinone oxidoreductase by 2,3,7,8-tetrachlorodibenzo-p-dioxin requires the CNC (cap 'n' collar) basic leucine zipper transcription factor Nrf2 (nuclear factor erythroid 2-related factor 2): cross-interaction between AhR (aryl hydrocarbon receptor) and Nrf2 signal transduction. *Biochem J* 2004;377(Pt 1):205-213.

103. Macleod, K. Tumor suppressor genes. *Curr Opin Genet Dev* 2000;10(1):81-93.

104. Mandal, A.K., Mitra, A. and Das, R. Sickle Cell Hemoglobin. *Subcell Biochem* 2020;94:297-322.

105. Margulies, M., *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005;437(7057):376-380.

106. Martin, G.S. The hunting of the Src. *Nature reviews. Molecular cell biology* 2001;2(6):467-475.

107. Martínez-Jiménez, F., *et al.* A compendium of mutational cancer driver genes. *Nature reviews. Cancer* 2020;20(10):555-572.

108.Martinsson, T., *et al.* Appearance of the novel activating F1174S ALK mutation in neuroblastoma correlates with aggressive tumor progression and unresponsiveness to therapy. *Cancer research* 2011;71(1):98-105.

109.Masella, R., *et al.* Novel mechanisms of natural antioxidant compounds in biological systems: involvement of glutathione and glutathione-related enzymes. *J Nutr Biochem* 2005;16(10):577-586.

110.Maxam, A.M. and Gilbert, W. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America* 1977;74(2):560-564.

111.McCarroll, S.A., *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature genetics* 2008;40(10):1166-1174.

112.McKernan, K.J., *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome research* 2009;19(9):1527-1541.

113.McMahon, M., *et al.* The Cap'n'Collar basic leucine zipper transcription factor Nrf2 (NF-E2 p45-related factor 2) controls both constitutive and inducible expression of intestinal detoxification and glutathione biosynthetic enzymes. *Cancer research* 2001;61(8):3299-3307.

114.McMahon, M., *et al.* Dimerization of substrate adaptors can facilitate cullin-mediated ubiquitylation of proteins by a "tethering" mechanism: a two-site interaction model for the Nrf2-Keap1 complex. *The Journal of biological chemistry* 2006;281(34):24756-24768.

115.Menegon, S., Columbano, A. and Giordano, S. The Dual Roles of NRF2 in Cancer. *Trends in molecular medicine* 2016;22(7):578-593.

116.Min Jou, W., *et al.* Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature* 1972;237(5350):82-88.

117.Morris, S.W., *et al.* Fusion of a kinase gene, ALK, to a nucleolar protein gene, NPM, in non-Hodgkin's lymphoma. *Science (New York, N.Y.)* 1994;263(5151):1281-1284.

118.Morris, S.W., *et al.* ALK, the chromosome 2 gene locus altered by the t(2;5) in non-Hodgkin's lymphoma, encodes a novel neural receptor tyrosine kinase that is highly related to leukocyte tyrosine kinase (LTK). *Oncogene* 1997;14(18):2175-2188.

119.Nagalakshmi, U., *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science (New York, N.Y.)* 2008;320(5881):1344-1349.

120. Nordling, C.O. A new theory on cancer-inducing mechanism. *Br J Cancer* 1953;7(1):68-72.

121. Nyrén, P. and Lundin, A. Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. *Anal Biochem* 1985;151(2):504-509.

122. Ohno, S., Smith, H. H. Evolution of genetic systems. *Gordonand Breach, New York* 1972;366.

123. Oren, M. and Rotter, V. Mutant p53 gain-of-function in cancer. *Cold Spring Harb Perspect Biol* 2010;2(2):a001107.

124. Palmer, R.H*., et al.* Anaplastic lymphoma kinase: signalling in development and disease. *Biochem J* 2009;420(3):345-361.

125. Prasad, S., Gupta, S.C. and Tyagi, A.K. Reactive oxygen species (ROS) and cancer: Role of antioxidative nutraceuticals. *Cancer Lett* 2017;387:95-105.

126. Pushkarev, D., Neff, N.F. and Quake, S.R. Single-molecule sequencing of an individual human genome. *Nature biotechnology* 2009;27(9):847-850.

127. Rinn, J.L*., et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by non-coding RNAs. *Cell* 2007;129(7):1311-1323.

128. Robertson, G*., et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* 2007;4(8):651-657.

129. Rosenfeld, S. Are the somatic mutation and tissue organization field theories of carcinogenesis incompatible? *Cancer Inform* 2013;12:221-229.

130. Rous, P. A SARCOMA OF THE FOWL TRANSMISSIBLE BY AN AGENT SEPARABLE FROM THE TUMOR CELLS. *J Exp Med* 1911;13(4):397-411.

131. Rubin, H. Quantitative relations between causative virus and cell in the Rous no. 1 chicken sarcoma. *Virology* 1955;1(5):445-473.

132. Rushmore, T.H., Morton, M.R. and Pickett, C.B. The antioxidant responsive element. Activation by oxidative stress and identification of the DNA consensus sequence required for functional activity. *The Journal of biological chemistry* 1991;266(18):11632-11639.

133. Sabharwal, S.S. and Schumacker, P.T. Mitochondrial ROS in cancer: initiators, amplifiers or an Achilles' heel? *Nature reviews. Cancer* 2014;14(11):709-721.

134. Saiki, R.K*., et al.* Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science (New York, N.Y.)* 1985;230(4732):1350-1354.

135. Sanger, F., Brownlee, G.G. and Barrell, B.G. A two-dimensional fractionation procedure for radioactive nucleotides. *J Mol Biol* 1965;13(2):373-398.

136.Sanger, F. and Coulson, A.R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* 1975;94(3):441-448.

137.Sanger, F., Nicklen, S. and Coulson, A.R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* 1977;74(12):5463-5467.

138.Sauna, Z.E. and Kimchi-Sarfaty, C. Understanding the contribution of synonymous mutations to human disease. *Nature reviews. Genetics* 2011;12(10):683-691.

139.Schadt, E.E., Turner, S. and Kasarskis, A. A window into third-generation sequencing. *Human molecular genetics* 2010;19(R2):R227-240.

140.Schena, M*., et al.* Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (New York, N.Y.)* 1995;270(5235):467-470.

141.Schuch, A.P*., et al.* Sunlight damage to cellular DNA: Focus on oxidatively generated lesions. *Free radical biology & medicine* 2017;107:110-124.

142.Shaw, A.T*., et al.* Alectinib in ALK-positive, crizotinib-resistant, non-small-cell lung cancer: a single-group, multicentre, phase 2 trial. *Lancet Oncol* 2016;17(2):234-242.

143.Shaw, A.T*., et al.* Ceritinib in ALK-rearranged non-small-cell lung cancer. *N Engl J Med* 2014;370(13):1189-1197.

144.Shaw, A.T*., et al.* Crizotinib versus chemotherapy in advanced ALK-positive lung cancer. *N Engl J Med* 2013;368(25):2385-2394.

145.Solimini, N.L., Luo, J. and Elledge, S.J. Non-oncogene addiction and the stress phenotype of cancer cells. *Cell* 2007;130(6):986-988.

146.Sondka, Z*., et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nature reviews. Cancer* 2018;18(11):696-705.

147.Sonnenschein, C. and Soto, A.M. Over a century of cancer research: Inconvenient truths and promising leads. *PLoS Biol* 2020;18(4):e3000670.

148.Sonnenschein, C., Soto, A.M. The Society of Cells: Cancer and Control of Cell Proliferation. . *New York: Springer-Verlag;* 1999.

149.Srikantan, V*., et al.* PCGEM1, a prostate-specific gene, is overexpressed in prostate cancer. *Proceedings of the National Academy of Sciences of the United States of America* 2000;97(22):12216-12221.

150.Stehelin, D*., et al.* DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature* 1976;260(5547):170-173.

151.Stehr, H., *et al.* The structural impact of cancer-associated missense mutations in oncogenes and tumor suppressors. *Mol Cancer* 2011;10:54.

152.Subramanian, A., *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* 2005;102(43):15545-15550.

153.Sun, Y. and Ma, L. New Insights into Long Non-Coding RNA MALAT1 in Cancer and Metastasis. *Cancers (Basel)* 2019;11(2).

154.Tebay, L.E., *et al.* Mechanisms of activation of the transcription factor Nrf2 by redox stressors, nutrient cues, and energy status and the pathways through which it attenuates degenerative disease. *Free radical biology & medicine* 2015;88(Pt B):108-146.

155.Thompson, S.L. and Compton, D.A. Chromosomes and cancer cells. *Chromosome Res* 2011;19(3):433-444.

156.Tibes, R., *et al.* Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. *Mol Cancer Ther* 2006;5(10):2512-2521.

157.Trachootham, D., Alexandre, J. and Huang, P. Targeting cancer cells by ROS-mediated mechanisms: a radical therapeutic approach? *Nat Rev Drug Discov* 2009;8(7):579-591.

158.Trachootham, D., *et al.* Redox regulation of cell survival. *Antioxid Redox Signal* 2008;10(8):1343-1374.

159.Vafa, O., *et al.* c-Myc can induce DNA damage, increase reactive oxygen species, and mitigate p53 function: a mechanism for oncogene-induced genetic instability. *Molecular cell* 2002;9(5):1031-1044.

160.Van den Eynden, J., *et al.* Phosphoproteome and gene expression profiling of ALK inhibition in neuroblastoma cell lines reveals conserved oncogenic pathways. *Sci Signal* 2018;11(557).

161.van Dijk, E.L., *et al.* Ten years of next-generation sequencing technology. *Trends Genet* 2014;30(9):418-426.

162.Vander Heiden, M.G., Cantley, L.C. and Thompson, C.B. Understanding the Warburg effect: the metabolic requirements of cell proliferation. *Science (New York, N.Y.)* 2009;324(5930):1029-1033.

163.Velculescu, V.E., *et al.* Serial analysis of gene expression. *Science (New York, N.Y.)* 1995;270(5235):484-487.

164.Venter, J.C., *et al.* The sequence of the human genome. *Science (New York, N.Y.)* 2001;291(5507):1304-1351.

165. Wang, J., *et al.* Mouse transcriptome: neutral evolution of 'non-coding' complementary DNAs. *Nature* 2004;431(7010):1 p following 757; discussion following 757.

166. Wang, K.C. and Chang, H.Y. Molecular mechanisms of long non-coding RNAs. *Molecular cell* 2011;43(6):904-914.

167. Wang, X.J., *et al.* Nrf2 enhances resistance of cancer cells to chemotherapeutic drugs, the dark side of Nrf2. *Carcinogenesis* 2008;29(6):1235-1243.

168. Webb, T.R., *et al.* Anaplastic lymphoma kinase: role in cancer pathogenesis and small-molecule inhibitor development for therapy. *Expert Rev Anticancer Ther* 2009;9(3):331-356.

169. Yamamoto, H., *et al.* Microsatellite instability in cancer: a novel landscape for diagnostic and therapeutic approach. *Arch Toxicol* 2020;94(10):3349-3357.

170. Zelko, I.N., Mariani, T.J. and Folz, R.J. Superoxide dismutase multigene family: a comparison of the CuZn-SOD (SOD1), Mn-SOD (SOD2), and EC-SOD (SOD3) gene structures, evolution, and expression. *Free radical biology & medicine* 2002;33(3):337-349.

171. Zhang, D.D., *et al.* Keap1 is a redox-regulated substrate adaptor protein for a Cul3-dependent ubiquitin ligase complex. *Molecular and cellular biology* 2004;24(24):10941-10953.

172. Zhang, Y., *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome biology* 2008;9(9):R137.

173. Zou, H.Y., *et al.* PF-06463922, an ALK/ROS1 Inhibitor, Overcomes Resistance to First and Second Generation ALK Inhibitors in Preclinical Models. *Cancer cell* 2015;28(1):70-81.

83

7 APPENDIX