Gene fusions and microRNAs in cancer

Hafstað, Völundur

2024

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

Gene fusions and microRNAs in cancer

VÖLUNDUR HAFSTAÐ | DEPARTMENT OF CLINICAL SCIENCES, LUND | FACULTY OF MEDICINE | LUND UNIVERSITY

Printed by Media-Tryck, Lund 2024 NORDIC SWAN ECOLABEL 3041 0903

LUND UNIVERSITY

FACULTY OF MEDICINE

Department of Clinical Sciences, Lund

Lund University, Faculty of Medicine
Doctoral Dissertation Series 2024:83
ISBN 978-91-8021-578-7
ISSN 1652-8220

Gene fusions and microRNAs in cancer

Gene fusions and microRNAs in cancer

Völundur Hafstað

LUND
UNIVERSITY

DOCTORAL DISSERTATION

Doctoral dissertation for the degree of Doctor of Philosophy (PhD)
at the Faculty of Medicine at Lund University to be publicly defended
on 13th of June 2024 at 09.00, Belfragesalen, BMC D15, Lund

Faculty opponent
Professor Albin Sandelin
Department of Biology & BRIC, Copenhagen University

Organization: LUND UNIVERSITY

Document name: Doctoral dissertation

Date of issue: 2024-06-13

Author(s): Völundur Hafstað

Sponsoring organization:

Title and subtitle: Gene fusions and microRNAs in cancer

Abstract:

The genome of cancer cells is unstable. Flaws in the DNA repair mechanisms of these cells can lead to the creation of gene fusions, where parts of two different genes are erroneously combined. Additionally, the expression of microRNAs (miRNAs), small regulatory molecules that control gene activity, is often dysregulated in cancer. In this thesis, we investigate miRNAs, gene fusions, and the interplay between the two in cancer using bioinformatic approaches. We found that miRNA host genes are common in gene fusions and may provide an alternative mechanism to dysregulate their expression. Since gene fusion detection methods are prone to errors, we developed a method to validate fusion transcripts at the genomic level using matched whole-genome sequencing data. Utilizing information on validated fusion events from 910 tumors in The Cancer Genome Atlas, we trained a machine learning classifier to predict which fusion event are real, and demonstrated that this approach can improve the quality of fusion detection. Finally, we investigated the function of the ERBB2-encoded mir-4728 in breast cancer at the transcriptional and translational level, and found that it impacts the level of aromatase and other genes involved in estrogen biosynthesis. These findings contribute to a growing understanding of the complex nature of the cancer genome. The papers in this thesis lay a groundwork for further exploration of the multifaceted roles of both miRNAs and gene fusions in cancer, underscoring the importance of continued investigation into their roles in cancer initiation, progression, and therapeutic response.

Key words: MicroRNAs, Gene fusions, Cancer, Molecular genetics, Bioinformatics

Language English

ISSN and key title: 1652-8220

ISBN: 978-91-8021-578-7

Recipient's notes

Number of pages: 84

Gene fusions and microRNAs in cancer

Völundur Hafstað

LUND
UNIVERSITY

Coverphoto by Völundur Hafstað

Faculty of Medicine

Department of Clinical Sciences, Lund

ISBN 978-91-8021-578-7

ISSN 1652-8220

Lund University, Faculty of Medicine Doctoral Dissertation Series 2024:83

Printed in Sweden by Media-Tryck, Lund University, Lund 2024

"We who cut mere stones must always be envisioning cathedrals.
— Quarry worker's creed"
Andrew Hunt, The Pragmatic Programmer:
From Journeyman to Master

Table of Contents

Abstract

The genome of cancer cells is unstable. Flaws in the DNA repair mechanisms of these cells can lead to the creation of gene fusions, where parts of two different genes are erroneously combined. Additionally, the expression of microRNAs (miRNAs), small regulatory molecules that control gene activity, is often dysregulated in cancer. In this thesis, we investigate miRNAs, gene fusions, and the interplay between the two in cancer using bioinformatic approaches. We found that miRNA host genes are common in gene fusions and may provide an alternative mechanism to dysregulate their expression. Since gene fusion detection methods are prone to errors, we developed a method to validate fusion transcripts at the genomic level using matched whole-genome sequencing data. Utilizing information on validated fusion events from 910 tumors in The Cancer Genome Atlas, we trained a machine learning classifier to predict which fusion event are real, and demonstrated that this approach can improve the quality of fusion detection. Finally, we investigated the function of the ERBB2-encoded mir-4728 in breast cancer at the transcriptional and translational level, and found that it impacts the level of aromatase and other genes involved in estrogen biosynthesis. These findings contribute to a growing understanding of the complex nature of the cancer genome. The papers in this thesis lay a groundwork for further exploration of the multifaceted roles of both miRNAs and gene fusions in cancer, underscoring the importance of continued investigation into their roles in cancer initiation, progression, and therapeutic response.

Popular scientific summary

Cancers develop due to errors in their genomes. These can range from single typos in genes to larger mix-ups where entire segments of the genome get rearranged, duplicated, deleted, or otherwise modified. Mistakes in the DNA repair mechanisms of the cells sometimes lead to the formation of gene fusions, where parts of two different genes are physically joined. These gene fusions can have properties from both genes and contribute to cancer progression. Our study explored these gene fusions and another key player in cancer: microRNAs (miRNAs). These tiny RNA molecules act like dimmer switches, regulating the activity of other genes. Abnormal levels of miRNAs in the cell can also lead to cancer progression. Interestingly, miRNAs are often nested inside larger genes in the genome. These larger genes that encompass miRNAs are often called miRNA hosts.

In paper I we investigated how gene fusions might affect miRNAs in the cell. We found that fusion events involving miRNA host genes are surprisingly common, occurring more often than is expected by chance alone. We show that fusion events involving specific miRNAs linked to cancer development generally caused an increase in their levels. Overall, our work suggests that gene fusions may provide an alternative way to regulate miRNA levels in the cell.

Detecting gene fusions accurately is crucial for cancer research, but current methods have limitations. We typically look for fusions at the RNA level, but many of the events that we see there are artifacts that are the result of the noisiness of RNA sequencing data. In paper II, we developed a method to confirm the presence of fusions directly in the DNA. The idea behind this pipeline is that if we see a fusion in both RNA and DNA it is very unlikely to be a false positive. We evaluated the pipeline against known gene fusions, and found that it is both faster and more sensitive than similar tools. Building on this, in paper III we utilized our pipeline to detect and validate fusions in over 900 tumor samples from various cancer types. This information was then used to train a machine learning classifier to distinguish real fusions from errors. We show that this machine learning approach can outperform other standard ways to filter out false positive fusions. The work in papers II and III paved the way for more accurate detection of gene fusions, which is crucial if they are to be utilized in a clinical setting.

In paper IV we shifted our focus back to miRNAs. One subtype of breast cancer is associated with high levels of a specific gene called ERBB2, and this subtype is generally associated with more aggressive tumors. Our research group previously discovered a miRNA called mir-4728 located inside the ERBB2 gene. In this paper we study the effects of this miRNA in breast cancer on both the RNA and protein level. We found that mir-4728 impacts the production of estrogen, another molecule that plays a vital role in breast cancer.

Overall, the work in this thesis expands our knowledge on gene fusions, miRNAs, and the interaction between the two in cancer. Our findings offer valuable clues, but they also reveal how much we still have to learn about cancer biology. By understanding fundamental mechanisms in cancer such as these, we pave the way for the development of novel therapies. Although much remains to be discovered, this work brings us closer to the ultimate goal: effective new treatments for patients battling this complex disease.

Populärvetenskaplig sammanfattning

Cancer utvecklas på grund av fel i genomet, cellens DNA. Dessa kan sträcka sig från enstaka stavfel i gener till större förändringar där hela segment av genomet flyttas, dupliceras, raderas eller på annat sätt modifieras. Misstag i cellens maskineri för att laga skador i DNA leder ibland till att fusionsgener, där delar av två olika gener är fysiskt sammanfogade, bildas. Fusionsgener kan ha egenskaper från båda generna och bidra till cancern utvecklas. Vi har studerat både fusionsgener och mikroRNA (miRNA), en annan nyckelspelare i cancer. Dessa små RNA-molekyler fungerar som en dimmer och kan reglera aktiviteten hos andra gener. Onormala nivåer av miRNA i cellen kan också leda till utveckling av cancer. Intressant nog finns miRNA ofta inuti större gener som då kallas värdgener för miRNA.

I artikel I undersökte vi hur fusionsgener kan påverka miRNA i cellen. Vi fann att fusioner som involverar värdgener för miRNA är förvånansvärt vanliga och att de förekommer oftare än slumpen. Vi visade också att fusioner som involverar värdgener för miRNA som har kopplats till utveckling av cancer ofta orsakade en ökning av deras nivåer. Sammantaget tyder vårt arbete på att fusionsgener kan vara ett sätt för cancerceller att förändra miRNA-nivåerna i cellen.

Att kunna upptäcka fusionsgener i cancer är viktigt både för forskning och i kliniken, men de nuvarande metoderna har begränsningar. Man letar ofta efter fusioner på RNA-nivå, men många av de möjliga fusionsgener som hittas där är istället artefakter som är ett resultat av bruset i RNA-sekvenseringsdata. I artikel II utvecklade vi därför en metod för att bekräfta närvaron av fusioner i DNA från samma prov. Tanken bakom denna pipeline är att det är mycket osannolikt att en fusionsgen är falsk om vi hittar den i både RNA och DNA. Vi utvärderade vår metod med kända fusionsgener och fann att den var både snabbare och känsligare än liknande verktyg. Med utgångspunkt i detta använde vi i artikel III vår pipeline för att detektera och validera fusioner i över 900 tumörprover från olika cancertyper. Informationen användes sedan för att med maskininlärning träna en klassificerare till att kunna skilja verkliga fusionsgener från falska. Vi visade att en metod som baseras på maskininlärning kan överträffa andra vanliga sätt att filtrera bort falska fusioner. Arbetet i artiklarna II och III banar väg för förbättrade analyser av fusionsgener, vilket är viktigt för kliniska tillämpningar.

I artikel IV fokuserade vi återigen på miRNA. En typ av bröstcancer karaktäriseras av höga nivåer av ett protein som kallas ERBB2 och dessa tumörer är ofta också mer aggressiva. Vår forskargrupp har tidigare upptäckt ett miRNA som heter mir-4728 och som ligger inuti genen för ERBB2. I artikeln studerar vi vilka effekter detta miRNA har i bröstcancerceller. Vi fann att mir-4728 påverkade produktionen av det kvinnliga könshormonet östrogen, en annan molekyl som är viktig i bröstcancer. Sammanfattningsvis så har arbetet i denna avhandling ökat vår kunskap om fusionsgener, miRNA och interaktionen mellan de två i cancer. Medan våra resultat ger värdefulla insikter, belyser de också det stora, ofullständigt utforskade landskap som tumörbiologin utgör. Genom att förstå grundläggande mekanismer som dessa i cancer banar vi vägen för utveckling av nya behandlingsmetoder. Även om mycket återstår att upptäcka, för detta arbete oss närmare det slutliga målet: effektiva nya behandlingar för patienter som kämpar mot denna komplexa sjukdom.

Vísindaleg samantekt

Mörg krabbamein þróast vegna villna í erfðamengi þeirra. Þetta getur verið allt frá stökum prentvillum í basaröðum gena til stærri viðburða þar sem heilir bútar erfðamengisins eru afritaðir, þeim endurraðað, eytt eða breytt á annan hátt. Mistök í DNA viðgerðarkerfum frumna leiða stundum til genasamruna, þar sem hlutar af tveimur genum eru tengdir saman. Þessir samrunar geta haft eiginleika frá báðum genum og stuðlað að framgangi krabbameins. Rannsóknir okkar beindust að þessum genasamrunum og annari sameindategund sem gegnir lykilhlutverki í krabbameini: míkróRNA (miRNA). Þessar smáu RNA sameindir fínstilla starfsemi annarra gena. Óeðlilegt magn af miRNA í frumunni getur einnig leitt til framvindu krabbameins. Athyglisvert er að miRNA eru oft staðsett inni í stærri genum í erfðaefninu. Þessi stærri gen sem umlykja miRNA eru oft kölluð miRNA-hýslar.

Í grein eitt könnuðum við hvernig genasamruni gæti haft áhrif á starfsemi miRNA í krabbameinsfrumum. Við komumst að því að samrunatilvik sem tengjast miRNA hýsilgenum eru furðulega algeng. Við sýndum einnig fram á að samrunatilvik einstakra miRNA sem tengjast krabbameinsþróun ollu almennt aukningu á magni þeirra. Á heildina litið bendir vinna okkar til þess að genasamruni geti verið önnur leið til að stjórna magni miRNA í frumunni.

Að greina samruna gena nákvæmlega er mikilvægt í krabbameinsrannsóknum, en núverandi aðferðir til að gera það eru takmarkaðar. Við leitum venjulega að genasamrunum á RNA stigi, en margir atburðir sem við sjáum þar eru ekki raunverulegir, heldur eru afleiðing af lágum gæðum gagna RNA-raðgreiningar. Í grein tvö þróuðum við aðferð til að staðfesta tilvist samruna beint í DNA. Hugmyndin á bak við þessa aðferð er sú að ef við sjáum genasamruna í bæði RNA og DNA er mjög líklegt að þetta sé raunverulegur viðburður. Við lögðum mat á aðferð okkar með því að skoða þekkta genasamruna og komumst að því að hún er bæði næmari og hraðari en svipaðar aðferðir. Í grein þrjú notuðum við aðferðina okkar til að greina og sannreyna genasamruna í yfir 900 æxlissýnum frá ýmsum krabbameinstegundum. Þessar upplýsingar voru síðan notaðar til að þjálfa vélnámsflokkara til að greina raunverulegan samruna frá villum. Við sýnum að þessi vélanámsaðferð getur skilað betri niðurstöðum en aðrar staðlaðar leiðir til að sía í burtu fölsk samrunatilvik. Vinnan í greinum II og

III greiðir leiðina fyrir nákvæmari greiningu á genasamrunum, sem skiptir sköpum ef þeir eiga að nýtast í klínísku umhverfi.

Í grein fjögur beindum við athygli okkar aftur að miRNA. Einn undirflokkur brjóstakrabbameins er tengur háu stigi á ákveðnu geni sem kallast *ERBB2*, og þessi flokkur er almennt tengdur verri æxlum. Rannsóknarhópur okkar uppgötvaði miRNA sem fékk heitið mir-4728 og er staðsett inni í *ERBB2*-geninu. Í þessari grein rannsökum við hvernig þetta miRNA getur haft áhrif á bæði RNA og prótein í brjóstakrabbameinsfrumum. Við komumst að því að mir-4728 hefur áhrif á framleiðslu estrógens, annarrar sameindar sem leikur lykilhlutverk í brjóstakrabbameini.

Rannsóknir okkar í þessari ritgerð auka þekkingu okkar á genasamrunum, miRNA, og samspili þeirra í krabbameini. Niðurstöður okkar veita innsýn í ákveðna sameindaerfðafræðilega atburði krabbameina, en þær sýna einnig hversu mikið er enn óuppgötvað í krabbameinslíffræði. Aukinn skilningur á líffræði krabbameins er eitt það helsta sem gerir okkur kleyft að þróa nýjar meðferðir gegn þessum afar flókna sjúkdómi.

List of publications

Paper I

Hafstað, V., Søkilde, R., Häkkinen, J., Larsson, M., Vallon-Christersson, J., Rovira, C., Persson, H., 2022. Regulatory networks and 5′ partner usage of miRNA host gene fusions in breast cancer. International Journal of Cancer 151, 95–106.

Paper II

Hafstað, V., Häkkinen, J., Persson, H., 2023. Fast and sensitive validation of fusion transcripts in whole-genome sequencing data. BMC Bioinformatics 24, 359.

Paper III

Hafstað, V., Häkkinen, J., Larsson, M., Staaf, J., Vallon-Christersson, J., Persson, H., 2023. Improved detection of clinically relevant fusion transcripts in cancer by machine learning classification. BMC Genomics 24, 783.

Paper IV

Hafstað V.**,** Albrecht J, Han E & Persson H. The ERBB2-encoded miRNA miR-4728-3p regulates estrogen signaling in SK-BR-3 cells. *Manuscript*.

Abbreviations

BAM	binary alignment map
cDNA	complementary DNA
dNTP	deoxyribonucleotide triphosphate
EFB	exclusive feature bundling
EMT	epithelial-mesenchymal-transition
ER	estrogen receptor
GOSS	gradient-based one-side sampling
GSEA	gene set enrichment analysis
HER2	human epidermal growth factor receptor 2
IHC	immunohistochemistry
NGS	next-generation sequencing
miRNA	microRNA
mRNA	messenger RNA
nt	nucleotide
ORA	overrepresentation analysis
PARP	poly(ADP-ribose) polymerase
PgR	progesterone receptor
PR	precision-recall
pre-miRNA	precursor miRNA
pri-miRNA	primary miRNA
RISC	RNA-induced silencing complex
RNA-Seq	RNA sequencing

ROC	receiver operator characteristic
SAM	sequence alignment map
SCAN-B	Sweden Cancerome Analysis Network – Breast
sncRNA	small non-coding RNA
snoRNA	small nucleolar RNA
TCGA	The Cancer Genome Atlas
TKI	tyrosine kinase inhibitor
TNBC	triple-negative breast cancer
UTR	untranslated region
WGS	whole-genome sequencing

Introduction

Cancer

What is cancer?

More than one in every six deaths worldwide is caused by cancer[1]. But this statistic alone does not capture how truly formidable this disease is, due to its extreme complexity and relentless adaptability. Cancer can be thought of not as a single disease, but a collection of hundreds of different diseases, each with its own characteristics, causes, prognoses, and treatments. Understanding this diversity is crucial, as it holds the key to unlocking effective treatments and ultimately saving lives.

While the fight against this disease has gained significant public momentum recently, cancer has a long-documented history. Fossil records show evidence of osteosarcoma in a 240-million-year-old reptile, and the first description of cancer in humans was in ancient Egypt 5000 years ago – around the same time as Stonehenge was built[2,3]. The terms "cancer" and "carcinoma" were created by Hippocrates in ancient Greece, and discussions of the disease have continued through historical medical texts since then. However, it was not until the mid-20th century that we truly began to understand what cancer is and how to properly treat it[4]. Since then, advancements have been rapid, with treatment methods today being significantly different from even those just a few decades ago. While much progress has been made in understanding and treating cancer, it still remains a significant public health challenge, demanding continued research to further our understanding and to develop more effective treatment strategies.

Cancer is a disease of the genome, characterized by uncontrolled growth of transformed cells[5]. These cells often develop the ability to spread to other parts of the body in a process known as metastasis. This transformation from healthy to malignant cells is generally caused by the accumulation of random mutations in the genome that are perpetuated through subsequent cell divisions. In a process similar to Darwinian evolution, some of these mutations give the cells that inherit them an advantage to survive and multiply, leading to cancer development.

Cancer can be classified based on the origin and type of the malignant tissue. For instance, carcinomas originate from epithelial tissue such as the skin or lining of organs, sarcomas develop in connective tissues such as bones and muscles, leukemia affects blood cells, and lymphoma originates from the lymphatic system. The organ that the cancer originates from adds another layer of detail; breast cancer and lung cancer are both carcinomas, but their organs of origin dramatically impact their biology and treatment approaches[6].

In addition to classifying cancer based on tissue and organ, they can be further categorized into subtypes based on various characteristics. These subtypes, however, still remain heterogeneous at the genetic level – no two tumors are ever identical, and cancer cells can continue to evolve even within a single tumor, creating further heterogeneity. Although there are over a hundred different types and subtypes of cancer, they generally all share a set of fundamental principles that characterizes this disease (Figure 1). These principles - known as the hallmarks of cancer - are capabilities that the cancer acquires over time and enables it grow and invade[7–9].

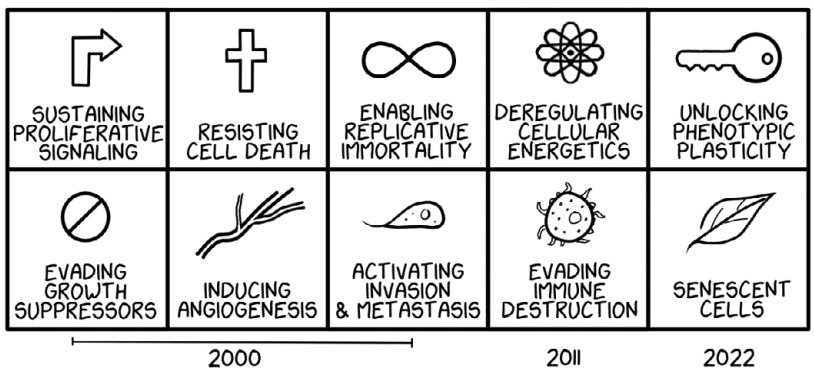

Figure 1
The hallmarks of cancer as proposed by Hanahan and Weinberg[7–9]. These are fundamental capabilities acquired by cells during tumorigenesis. The activation of the hallmarks of cancer disrupts cellular homeostasis and promotes cancer development and progression.

For a tumor to be classified as cancerous, it must have the ability to spread, either locally or to other parts of the body. This hallmark, known as *activating invasion and metastasis*, is the process where a developmental program known as epithelial-mesenchymal-transition (EMT) is enabled in cancer cells to gain invasive properties[10]. Activating this program triggers a cascade of changes within the cancer cells. They lose their epithelial characteristics, such as polarity and the expression of cell-adhesion molecules that normally keep them bound to their neighbors. These morphological changes and loss

of adhesion grant the cells motility and allow them to disseminate into surrounding tissue. They can also enter the blood stream or lymphatic system, causing them to spread to distant organs. There they can reattach and form a metastasis.

Beyond the ability to spread, cancer cells possess the ability to relentlessly grow and divide. Two hallmarks emphasize this ability: *sustaining proliferative signaling* and *evading growth suppressors*. Healthy cells tightly regulate their growth through various signaling pathways. These pathways interact and overlap, creating a complicated network of signals within the cell. Specific molecules, such as growth factors, send signals through these pathways instructing the cell to promote or inhibit proliferation. In cancer, these systems become dysregulated though various means. For example, amplification or overexpression of genes encoding growth factor receptors can lead to an overabundance of those receptors in the cell. This can lead to a constant "on" signal for the pathways that these molecules are associated with, resulting in uncontrolled growth[11]. Healthy cells possess mechanisms that regulate growth and proliferation, acting as the brakes of the cellular machinery and ensuring that genes do not divide uncontrollably. Genes involved in these mechanisms encode proteins that can arrest cell cycle progression. Cancer cells can evolve to bypass these normal controls, allowing for continued proliferation.

Normal cells have a limited lifespan, dictated by the continual shortening of the structures at the ends of their chromosomes called telomeres. Telomeres act as protective caps on the chromosomes, and with each cell division they become progressively shorter. Once they reach a critically short length, the cell can no longer divide and enters a state of permanent growth arrest or undergoes apoptosis. However, cancer cells overcome this limitation and acquire the hallmark of *enabling replicative immortality*, allowing them to divide indefinitely[12]. One of the primary mechanisms by which a cancer cell achieves this is through the activation of an enzyme called telomerase. Telomerase is naturally active in stem cells and some specialized cell types, where it maintains telomere length during cell division. However, in most adult somatic cells, telomerase activity is extremely low or absent. In cancer cells, mutations or epigenetic changes can lead to the abnormal upregulation of telomerase activity. This reactivated enzyme can then synthesize new telomeric DNA sequences, effectively lengthening the telomeres and resetting the cellular clock. By maintaining telomere length, cancer cells bypass the natural replicative barrier and gain the ability to divide indefinitely[8].

In addition to continuous and relentless proliferation, cancer cells also *resist cell death,* another natural process that eliminates unwanted or damaged cells via mechanisms such as apoptosis. Cells can enter apoptosis as a response to particularly harsh physiological stress, DNA damage, or signals indicating a loss of growth control. This self-destruct

mechanism plays a vital role in maintaining tissue homeostasis and preventing the accumulation of potentially harmful mutations. Apoptosis is a tightly regulated process orchestrated by a network of signaling pathways. Cancer cells can disrupt this network at various points, effectively disarming the cell's self-destruct program.

To sustain unlimited growth, a tumor must be able to supply the cancer cells with oxygen and nutrients. Cancer cells achieve this through the hallmark of *inducing angiogenesis*, a process where they stimulate the growth of new blood vessels towards the tumor. Without a steady supply of oxygen and nutrients, cancer cells at the core of the tumor would become starved and die. Angiogenesis plays a critical role not only in promoting tumor growth but also in facilitating metastasis, the spread of cancer to distant organs[13].

Since their initial description in 2000 by Hanahan and Weinberg[7], the hallmarks of cancer have been refined to reflect the significant advancements in our understanding of this disease. Two subsequent publications have expanded on the original framework to include four new hallmarks[8,9]. We now know that the immune system plays a vital role in preventing cancer by identifying and destroying abnormal cells. However, cancer cells can acquire the hallmark of *avoiding immune destruction*, allowing them to camouflage themselves and escape the immune system's attack**.** Cancer cells can also *reprogram their cellular metabolism*, allowing them to adapt to their rapid growth, even in unfavorable environments[14,15]. *Cellular senescence*, despite being usually thought of as a protective mechanism, has recently been shown to stimulate tumor development[16]. It is still not clear how or to what extent senescent cells contribute to tumor development, and as such this has been labeled as an emerging hallmark. Another emerging hallmark, and the last of the current hallmarks of cancer, is *unlocking phenotypic plasticity*. This encompasses the idea that cells that have gone down a specific path of cellular differentiation are able to escape from this normally terminal state, and that this characteristic is important for cancer progression[17].

Causes of cancer

Building upon the understanding of what defines cancer, we can now ask the question: what triggers the transformation of a healthy cell into these aggressive and deadly entities? While phrases such as “smoking causes cancer” highlight specific risk factors, to truly understand this transformation we must look at the underlying molecular mechanisms that result in malignancy. As previously mentioned, accumulating mutations in the cell genomes is a large factor, but it only represents one of several *enabling characteristics* that pave the way for cancer development (Figure 2).

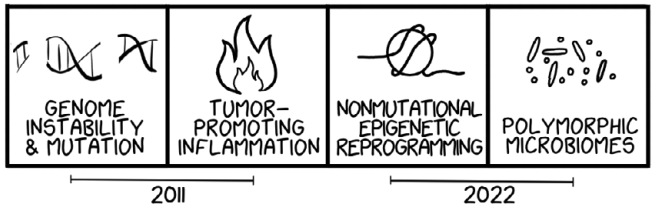

Figure 2
Enabling characteristics of cancer as proposed by Hanahan and Weinberg[7-9]. These characteristics create a permissive environment for the acquisition of the hallmarks of cancer, facilitating tumorigenesis.

Changes in the genome can be diverse, ranging from single nucleotide substitutions to the deletion of entire chromosome segments. Mutations are generally random and infrequent, but various agents in the environment, encompassing everything from UV radiation to chemical toxins, can increase the cells susceptibility to such changes. They achieve this by either directly damaging the DNA or by disrupting the mechanisms that maintain its integrity, such as DNA repair and replication fidelity[18]. A key player in maintaining genomic integrity is the p53 protein. This protein acts as a master regulator, dictating how the cell responds to various stress signals by initiating processes such as apoptosis, cell cycle arrest, DNA repair, or changes in metabolism. Through these diverse functions, p53 safeguards the integrity of the genome and prevents uncontrolled cell growth[19,20]. Genes encoding for proteins like p53, that protect the integrity of the genome and prevent unwanted cell growth, are collectively known as "tumor suppressors". Mutations that cause tumor suppressors to lose some or all of their function are commonly seen in cancer, as these mutations allow the cell to divide unchecked and give them an evolutionary advantage compared to healthy cells. This concept is explained by the two-hit hypothesis, which proposes that both alleles of a tumor suppressor gene need to be inactivated for a phenotypic change to occur[21]. This inactivation can arise through mutations that build up over an individual's lifetime. A person can also be born with inherited (germline) mutations in tumor suppressor genes, making them predisposed to developing cancer later in life. Like everything else in biology, there are of course exceptions to the two-hit hypothesis, and several tumor suppressors have been identified that require both alleles to be in-tact[22]. Mutations in the genome may also cause the activation of genes known as "oncogenes", that are involved in various functions closely related to the hallmarks of cancer such as stimulating cell growth and survival. Unlike the two-hit hypothesis of tumor suppressor genes, activating mutations in a single copy of an oncogene is usually enough to produce a phenotypic change to the cell.

In addition to changes in the genome itself, other changes can occur that do not affect DNA sequences, but rather epigenetic traits such as chromatin structure and methylation. These can be caused not only by mutations in the genome, but the tumor microenvironment can also impart epigenetic changes due to e.g., hypoxia[23,24]. Inflammation can in some cases also promote tumor progression. Although the immune system generally acts to prevent cancer, inflammation can supply a tumor with a variety of growth factors that contribute to angiogenesis, proliferation, and other hallmarks[25]. The human body harbors multiple trillions of microorganisms that together make up the microbiome. Advances in sequencing technologies have led to the discovery that many tissues that once were thought to be sterile actually contain their own microecologies[26]. Studies into tumor microbiomes indicate that the microorganisms that are present in or around the tumor can also contribute to oncogenesis, but this field is at an early stage[26–28].

Breast cancer

Introduction

Breast cancer is the second-most diagnosed cancer worldwide behind lung cancer[1]. It is by far the most commonly diagnosed cancer among women, accounting for one in every four cases and one in every six cancer related deaths[29]. According to the National Quality Register for Breast Cancer, 9491 new cases were diagnosed in Sweden in 2022[30].

Despite significant advances having been made in treating breast cancer, the incidence is increasing globally. By 2040 it is projected that the number of new cases of breast cancer will have grown by 40% and number of deaths will increase by 50%[29]. Most women diagnosed with breast cancer are over 50 years old, but the age distribution of cases and mortality rates vary significantly across the globe. In less developed countries, these metrics tend to be skewed towards a younger age, and they generally correlate with the human development index of the country. The incidence is highest in industrialized countries, and might be due to lifestyle-related risk factors such as diet, weight, stress, alcohol consumption and little physical activity. Other factors such as age, early menarche, number of children, age at first pregnancy, and late menopause also increase the risk of developing breast cancer[31,32]. Family history is also an important risk factor, with germline mutations in important tumor suppressor genes such as *BRCA1* and *BRCA2* accounting for 5% of breast cancer cases[33,34].

Early detection remains a cornerstone in the fight against breast cancer. Screening allows for the identification of the disease at its earliest stages, often before any noticeable symptoms arise. This early detection window increases the success rate of treatments, as smaller tumors are generally easier to eradicate. Early-stage cancers often qualify for less invasive procedures and lower radiation doses, minimizing the impact on the patient. In Sweden, women aged 40 to 75 are invited to participate in a breast cancer screening program every 1.5 to 2 years. This program utilizes mammography, an X-ray imaging technique that captures two to three images of each breast. These images are then examined by a radiologist for the early detection of breast cancer. Systematic screening programs such as this have been shown to reduce the the number of breast-cancer related deaths by approximately 20%[35]. Mammography screening programs have also been criticized for overdiagnosis. Studies have shown that the cumulative risk for a false positive screening in women aged 50 to 69 is between 8% and 21%. However, positive screening tests are followed up with a non-invasive assessment, minimizing the number of women that undergo an invasive biopsy or surgery that do not need it[36].

Classifying breast cancer

Cancer classification helps physicians understand the specific type of cancer a patient has and is important for determining the best course of treatment. Like most other cancers, breast cancer classification starts with the organ where the tumor originates (Figure 3). There are also several additional factors considered to further categorize them, such as their location within the breast, molecular characteristics, and growth rate.

Histopathological classification gives information on specific morphological features of the tumor. The vast majority of breast cancer tumors are derived from epithelial tissue lining the mammary ducts and lobules, making them carcinomas. These can be split into *carcinoma in-situ* or *invasive carcinoma*, based on whether the tumor has penetrated the basal layer of the epithelium. Although there are many histopathological classes of breast cancer, between 70%-80% of all tumors fall into either invasive lobular carcinoma or invasive breast carcinoma of no special type (previously known as invasive ductal carcinoma). This classification is therefore limited in that it does not accurately reflect the heterogeneity of this disease[37–39].

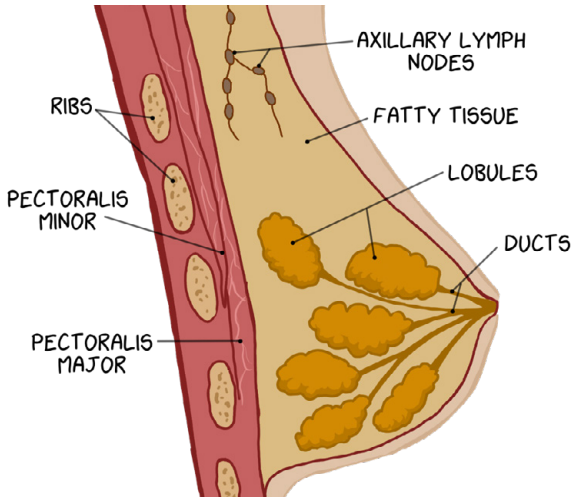

Figure 3
Anatomy of the breast.

Breast cancer cases can be further characterized by grade and stage. Grade represents how closely the cancer cells resemble normal breast cancer cells. There are several ways to measure this, but the recommended grading system by the WHO (the Nottingham grading system) looks at tubule formation, nuclear polymorphism, and mitotic count. Each characteristic is given a score from 1-3, making the final Nottingham score a range from 3-9[40]. Tumors with a high grade are called less differentiated, and a higher tumor grade is associated with a more aggressive behavior. Although tumor grade is a prognostic factor, it is not used to guide treatment decisions[41].

Stage is a measure of how much cancer is in the body. The most widely used staging system for solid tumor cancers – the TNM system – looks at three factors and scores each with a number. **T** is a measure of the primary tumor and its size, **N** refers to the number of regional lymph notes that the cancer has spread to, and **M** indicates whether the cancer has metastasized[42]. Like grade, the stage of a cancer is a prognostic factor, and can assist medical professionals in planning treatment.

While the previously mentioned classification systems provide valuable information about the cancer in question, they are primarily based on observable features and traditional clinical practices. With advances in molecular biology, it has been discovered that specific molecular markers can provide deeper insights and give a more personalized diagnosis. These markers are not only associated with prognosis, but can also indicate which treatments will be most effective for individual patients. This has led to the development of new classification systems that are rooted in molecular biology.

The most commonly examined molecular markers in breast cancer are the estrogen receptor (ER), progesterone receptor (PgR), and the human epidermal growth factor receptor (HER2, gene symbol *ERBB2*). These receptors play a vital role in the development and progression of certain breast cancers and are frequently overexpressed and/or amplified in tumor cells. In a clinical setting, they are detected using immunohistochemistry (IHC), a method that utilizes antibodies that bind to these proteins to visualize them on a microscope slide. Using the status of these three proteins as a reference, breast cancers can be broadly divided into three clinical groups: ER+, HER2+, and triple-negative breast cancer (TNBC)[43]. These are often combined with other markers such as the proliferation marker Ki67[44].

The estrogen receptor alpha, encoded by the gene *ESR1,* is the most prevalent molecular marker for breast cancer. It is overexpressed in approximately 70% of all breast cancers and encodes a transcription factor that is activated by estrogens. Once active, the estrogen receptor dimerizes and translocates to the nucleus, where it binds to estrogen response elements on target gene promoters to initiate transcriptional programs that lead to cell proliferation and other processes essential for tumor growth[45].

The progesterone receptor is a steroid hormone receptor that is primarily expressed in female reproductive tissue. In breast cancer, the expression of PgR is closely linked to the transcriptional programs activated by the estrogen receptor[46]. Despite ER and PgR being closely linked, PgR status in a tumor provides additional prognostic information beyond ER status alone. PgR positivity is generally associated with slower-growing tumors and a more favorable prognosis[47].

The *ERBB2* oncogene is overexpressed or amplified in approximately 20% of all breast cancers. This gene encodes a transmembrane receptor tyrosine kinase that contributes to aggressive tumor behavior. The ERBB2 protein has no known ligand, instead it activates by heterodimerizing with other ligand-bound members of the HER family: EGFR, ERBB3 or ERBB4. This leads to phosphorylation of tyrosine kinase residues that are present in the cytoplasmic domain of the receptor, and causes a signaling cascade that activates the phosphatidylinositol triphosphate kinase (PI3K) and mitogen-activated protein kinase (MAPK) signaling pathways. Activation of these pathways in turn causes cell cycle progression and proliferation[48].

Built upon advances in molecular biology, a more nuanced approach to classifying breast cancer tumors has emerged: intrinsic molecular subtyping. This method examines the expression patterns of specific genes to group tumors into subtypes with distinct clinical behaviors. The most notable of these is the PAM50 molecular subtypes first described two decades ago, and classifies breast cancer tumors into five main subtypes: luminal A, luminal B, HER2-enriched, basal-like and normal-like[49]. The

PAM50 subtypes are based on the expression signatures of 50 different genes, and they have been shown to have differences in incidence, survival, and treatment response. In addition, the intrinsic molecular subtypes do not reflect the standard receptor status, but complement and expand on them[50]. The PAM50 subtypes have influenced clinicopathological subtyping, which now tries to approximate the intrinsic subtypes by classifying samples into luminal A-like, luminal B-like, HER2-positive, and triple-negative subtypes based on IHC staining for ER, PgR, HER2 and Ki67[51].

Treating breast cancer

Despite massive advances in our understanding of breast cancer, surgery has been the primary method to deal with this disease since the late 18th century[52]. Breast cancer surgeries used to be mainly mastectomy, i.e., the removal of the entire breast. With medical advancements over the past decades, mastectomy has largely been replaced with lumpectomy, or breast-conserving surgeries, as they are less deforming and do not impact overall survival[53]. In Sweden, over 84% of all breast cancer surgeries were breast-conserving in 2021, up from only 7% in the 1980s[54]. Depending on the tumor stage, surgery may also be used to remove axillary lymph nodes.

Preceding or following surgery, additional treatment options like radiation therapy and chemotherapy may be recommended. Chemotherapy as a cancer treatment was pioneered during the second world war[55], and radiation therapy has been used for over 100 years[56]. Both therapies have undergone significant advancements since their inception, and continue to play a vital role in treating breast cancer today.

Radiation therapy uses ionizing radiation to damage the DNA of cancerous tissue, leading to cell death. Advances in the field now enable physicians to target a tumor with much greater precision, largely bypassing the normal side effects of heart and lung damage. Radiation therapy has evolved over the years to involve fractionated doses and to increase the precision of targeting the tumor itself[57]. Chemotherapy involves the use of different cytotoxic drugs that interfere with cell division, either by inhibiting mitosis or causing DNA damage. Today, chemotherapy is often given as a neoadjuvant (before surgery) treatment in combination with other treatment methods in effort to reduce the size of the tumor and to prevent it from spreading. Chemotherapy is also commonly prescribed after surgery, but the use of this treatment method depends on many factors such a stage and other clinical characteristics[58]. Normal cells in the body are often affected by chemotherapy, and as such this treatment method has many potential side effects. The bone marrow, hair follicles, and cells lining the intestines are particularly sensitive to chemotherapy, and the typical side effects are closely related to the functions of these cells[59].

Advancements in our understanding of cancer have led to the development of new therapies that are more tailored to the characteristics of each tumor. Targeted therapies for breast cancer are typically directed against the molecules that define the clinicopathological breast cancer subtypes: the hormone receptors and ERBB2.

Endocrine therapy, or hormonal therapy, is used to treat tumors that are hormone receptor-positive. Their function is to prevent the ER from exerting its effects on the cell, such as sustaining proliferation and evading growth suppressors. Endocrine therapies either target the receptor directly or disrupt the synthesis of estradiol – the ligand that is required to activate ER[57]. Drugs that target ER generally act as estrogen antagonists, competing for binding and affect the receptor in various ways such as preventing dimerization or blocking co-factor binding[60]. Inhibition of estradiol synthesis is achieved using aromatase inhibitors – these drugs target the enzyme aromatase that is responsible for converting testosterone into estradiol, and represents the rate-limited step in the estrogen biosynthesis pathway[61].

Breast tumors that are ERBB2-positive are usually treated with humanized monoclonal antibodies that target the ERBB2 protein on the surface of the cancer cells. The first and most well-known example is trastuzumab, developed in the early 1990s[62]. Interestingly, the mechanism of action for this drug is still unclear. It has been hypothesized that trastuzumab can prevent ERBB2 dimerization, block cleavage of the extracellular domain, cause endocytosis of the receptor, and recruit immune effector cells to destroy the cells that have this receptor on their surface[63]. Other antibodies have also been developed for the same purpose, including pertuzumab that prevents the heterodimerization of ERBB2 with other members of the HER family. The introduction of monoclonal antibody therapies against ERBB2 has dramatically improved the prognosis for patients with ERBB2-positive breast cancer, making them a cornerstone treatment for this subtype. Other drugs can be attached to antibodies, forming antibody-drug conjugates that have multiple mechanisms of action[64]. Unfortunately, a large portion of tumors develop resistance to these monoclonal antibodies, especially in the metastatic setting[65]. These drugs are therefore often given together and in combination with other treatments such as chemotherapy.

Tyrosine kinase inhibitors (TKIs) offer another treatment option for ERBB2-positive breast cancer patients. Several TKIs have demonstrated promising results either as monotherapy or in combination with chemotherapy and/or anti-ERBB2 antibodies. TKIs can be particularly beneficial for patients who develop resistance to antibody therapies. It is important to note, however, that most TKIs are currently only approved for the metastatic setting. Only one TKI, lapatinib, has received approval for use in early-stage breast cancer as an adjuvant therapy following treatment with trastuzumab[66].

Other drugs are used or are being developed for other groups of breast tumors. Approximately 5% of breast cancer patients have so-called homologous recombination deficiency, usually harboring loss of function mutations in the *BRCA1* and *BRCA2* genes. These genes play a vital role in repairing double-strand DNA breaks, and women with these mutations have as high as a 72% cumulative risk to develop breast cancer during their lifetime[67]. These patients can be treated with poly(ADP-ribose) polymerase (PARP) inhibitors. When administered to cells with HRD mutations, PARP inhibitors block the repair of single-strand breaks, ultimately leading to the formation of double-stranded breaks during DNA synthesis. These cells, lacking the ability to properly repair double-stranded DNA breaks, instead resort to the more error-prone non-homologous end-joining pathway. This results in replication errors and the eventual death of the cell, offering a targeted treatment strategy for this group of patients[68]. Patients with metastatic HR-positive breast cancer can be treated with cyclin-dependent kinase inhibitors. These drugs block the activity of the CDK4 and CDK6 proteins, which are responsible for cell cycle entry.

Cancer cells are constantly evolving, and one of the major challenges in treating them is their ability to develop resistance to virtually any therapy. There is a selective pressure on cancer cells undergoing treatment to evolve mechanisms that allow them to survive their new unfavorable environment, and when we treat a cancer we are inadvertently selecting for the most resilient cells. For example, a subpopulation of cancer cells within a tumor may survive a specific treatment and continue to grow, leading to recurrence. This new, resistant tumor can be more difficult to treat, as it may no longer respond to the original therapy. Treating cancer therefore often involves a combination of multiple drugs in addition to surgery and/or radiotherapy, and choosing a treatment involves the careful weighing of the potential benefits and side effects. Despite the challenges of treating this disease, the past few decades have witnessed remarkable advancements in our understanding of cancer biology and treatment options. Ongoing clinical and preclinical research still holds much promise for the future of cancer treatments[57,69].

MicroRNAs

Introduction

In December 1993, the field of molecular biology took a leap forward with the discovery of microRNAs (miRNAs) – a class of non-coding RNAs whose role is to regulate gene expression. Two research groups, led by Ambros and Ruvkun respectively, published back-to-back articles in Cell that studied a gene called *lin-4* in the organism

C. elegans[70,71]. It was known that the *lin-4* gene regulated the levels of another gene, *lin-14*, although the mechanism remained unknown[72]. Here, a surprising discovery was made: the *lin-4* gene does not encode a protein. Instead, the researchers identified two short transcripts of 22 nucleotides (nt) and 61 nt that had complementary sequences to the 3' untranslated region (UTR) of *lin-14*, and that base-pairing between the lin-4 and lin-14 RNAs is the mechanism that regulates the levels of *lin-14*.

Seven years later, it was discovered that another gene in *C. elegans*, *let-7*, displayed similar properties to that of *lin-4*[73]. This gene encodes a 21 nt RNA molecule that has complementary sequences to the 3' UTR of several other genes. Following these discoveries, researchers found that *lin-4* and *let-7* represented a large class of abundant RNA molecules – miRNAs, with orthologs found in *D. melanogaster* and even humans[74–76]. This also paved the way for the identification of thousands of new miRNAs across the plant and animal kingdoms[77–79]. Today there are 1917 identified miRNAs in the human genome that are listed in the miRBase miRNA database[80].

Biogenesis

The miRNA genes in the mammalian genome are typically much longer than their ~22 nt mature RNA product. Initially, miRNAs are transcribed by RNA polymerase II as large transcripts called primary-miRNAs (pri-miRNAs). These transcripts are usually several thousand nt, and contain distinct secondary structures called hairpins that are recognized and processed in the nucleus by the RNase III enzyme DROSHA together with its cofactor DGCR8[81–84]. These two proteins form the microprocessor complex, which with the help of several other proteins, cleaves a 60-70 nt stem-loop-containing segment out of the pri-miRNA[81,84,85]. This segment is known as the precursor miRNA (pre-miRNA). Interestingly, even though hairpin structures are abundant throughout the transcriptome, the microprocessor complex acts with remarkable precision, exclusively processing pri-miRNAs. The unique features that distinguish pri-miRNAs from other hairpin-containing transcripts are still poorly understood, but modern computational approaches offer promising insights[86].

Once it has been cleaved from the pri-miRNA, the pre-miRNA is transported into the cytoplasm by Exportin-5 and then further processed by the protein DICER. Similar to DROSHA, DICER is an RNase III endonuclease enzyme that cleaves the double-stranded hairpin of the pre-miRNA to generate a ~22 nt miRNA duplex with a 3' overhang of 2 nt[87,88]. After being processed by DICER, the miRNA duplex is loaded onto a member of the Argonaute protein family to form the RNA-induced silencing complex (RISC). This loading process is complicated and in humans requires at least five chaperone proteins for correct assembly[89]. One strand in the duplex, known as the

passenger strand, is then unwound, and discarded from the complex. It is still not clear how this is achieved, but evidence points towards the N-domain of AGO driving this process[90]. The remaining strand, known as the guide strand, is used to direct the RISC complex to mRNAs to mediate silencing. Several factors determine which strand in the duplex is selected to be the guide strand, most importantly the thermodynamic stability of each strand and sequence bias at the 5' end (Figure 4)[91].

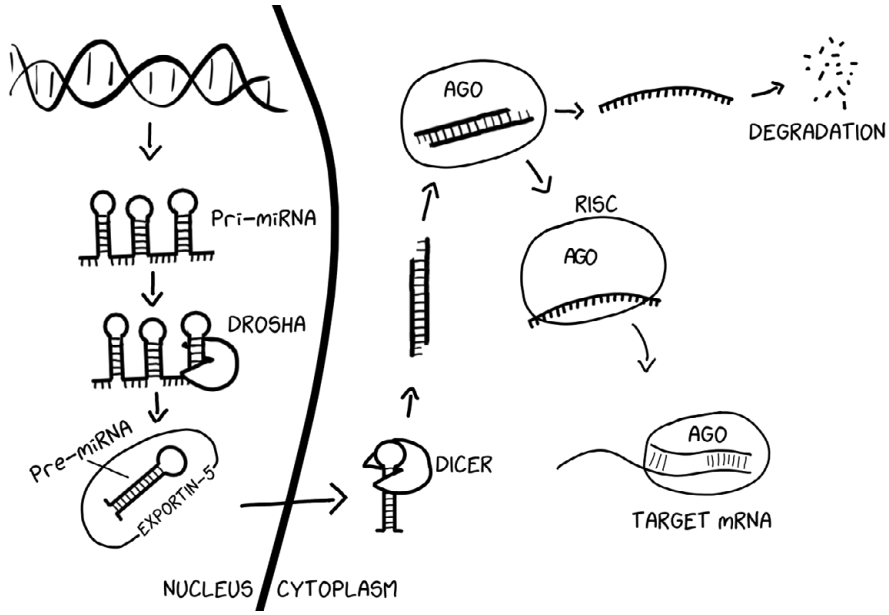

Figure 4
Canonical miRNA biogenesis. RNA polymerase II transcribes pri-miRNA from a DNA sequence. The pri-miRNA is processed into pre-miRNA in the nucleus by the enzyme DROSHA, before being exported to the cytoplasm by Exportin-5. In the cytoplasm, the pre-miRNA is processed into a mature miRNA duplex by DICER and is incorporated into an AGO protein complex. One strand of the duplex is degraded, and the remaining mature miRNA guides the complex to target mRNAs, regulating gene expression.

Mechanisms of miRNA-mediated gene silencing

The target sites for RISC are usually located in the 3' UTR of mRNAs, and target recognition is determined by nt 2-7 of the miRNA, known as the miRNA seed region[92]. The mature RISC complex exposes nt 2-5 of the seed region of the miRNA and changes its conformation so that it can easily base pair to complementary RNA targets[93,94]. When a target is found, a conformation change in AGO enables additional base pairing up to nt 8, and allows for supplementary pairing in nt 13-17 that stabilizes the target binding[95]. The binding of RISC to a target mRNA facilitates post-transcriptional regulation of the target transcript via the recruitment of other effector proteins. The

current model of miRNA-mediated silencing in animals suggests that first there is a translational repression step that is then followed by mRNA degradation[96,97]. It is still unclear how much each process contributes to miRNA-mediated silencing, although mRNA decay likely mediates the majority of the silencing effect[97].

The mechanisms behind miRNA-mediated translational repression are still poorly understood. There have been several different mechanisms proposed, but the extent of which each of them contributes to overall translational repression remains unknown. These mechanisms involve the recruitment of translational inhibitors via RISC and associated proteins[98] and breaking the mRNA loop structure that is required to start translation[99].

The effects of miRNA-mediated mRNA decay are generally better understood than their effect on translational repression. The GW182 protein interacts with AGO and recruits several other proteins to the target mRNA, including a poly(A)-binding protein, deadenylase complexes and decapping complexes[100–103]. Together, these proteins cause the deadenylation and subsequent degradation of the target mRNA poly(A) tail. RISC has also been shown to recruit decapping factors to further facilitate the degradation of its target[104].

Non-canonical miRNA biogenesis

There are several ways that miRNAs can deviate from the canonical biogenesis that is described above. Non-canonical miRNA biogenesis can be divided into two categories: DROSHA/DGCR8 independent and DICER -independent[105]. The introns of some protein-coding genes can encode non-canonical pri-miRNAs called mirtrons. During mRNA processing they are spliced out of the transcript using the normal splicing machinery. The pri-miRNAs of mirtrons contain stem-loop structures that are different from those found in canonical pri-miRNAs, and are not recognized by DROSHA. Instead, these RNA molecules are processed by the enzyme DBR1 to form pre-miRNAs, that are then processed normally exportin-5 and DICER[106,107]. The second way that miRNAs can deviate from the normal biogenesis is for their pre-miRNA to be processed independently of DICER. This is very rare, and has in fact only been described for a single miRNA – miR-451. In this case, the stem-loops of the pre-mir-451 RNA are too short to be recognized by DICER and are instead processed directly by AGO2[108,109].

Function

A substantial portion of the human miRNAs are located in genomic clusters that contain multiple miRNAs. Clusters of two or more miRNAs have been identified in both introns and intergenic regions, and a single pri-miRNA can contain multiple miRNA hairpins[110,111]. Nearly half of all miRNAs found in the human genome are encoded within introns of protein-coding genes, and these miRNAs can be categorized into two groups: those that are transcribed independently of their host gene and those that are not[112,113]. The intronic miRNAs that are transcribed together with their hosts have been suggested to play an autoregulatory role to maintain homeostasis[114,115].

The seed region is the primary determinant for which mRNAs a miRNA targets, but even the seed does not require perfect complementarity for the miRNA to mediate targeted silencing. As few as 6 nt are needed for a canonical miRNA target site, with the silencing efficacy increasing with additional base pairing (Figure 5). Additional base pairing in the 3' end of the miRNA can further facilitate target binding[94,111].

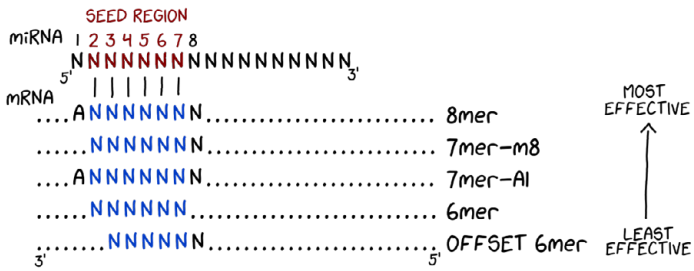

Figure 5
Seed interactions between miRNAs and mRNAs. This interaction is typically 6-7 Watson-Crick base pairings starting from nucleotide 2 of the miRNA. Several variations of this interaction exist with varying degrees of efficacy.

Due to the short length of the seed region and the sheer number of unique miRNAs found in humans, most human protein-coding genes contain conserved miRNA binding sites in their 3' UTR[116]. One mRNA transcript can be regulated by many different miRNAs, and one miRNA can target the 3' UTR of multiple different mRNAs, creating a complicated post-transcriptional regulatory network of gene expression.

The exact role of most miRNAs in the human genome is still unknown. Mouse experiments with miRNA knockouts have resulted in a wide variety of phenotypes, including embryonic lethality due to developmental defects, infertility, altered immune

response, reduced lifespan, and decreased blood glucose levels[117–120]. It can be difficult to study the effect of a single miRNA in a given setting - the many-to-many relationship between miRNAs and mRNA naturally leads to some redundancies that can be challenging to overcome.

Role in cancer

Due to their potential to regulate the expression of so many genes in the human genome, it is perhaps no wonder that miRNAs are often deregulated in cancer[121]. Many miRNAs can act as either tumor suppressors or oncogenes (sometimes referred to as oncomiRs) depending on their target genes, and their deregulation can influence many of the hallmarks of cancer including sustaining proliferative signaling, resisting cell death, angiogenesis and metastasis[122–126]. There are many ways that miRNA expression and function can become deregulated in a cancer cell: Amplification or deletion of miRNA genes, changes in the biosynthesis machinery, altered transcription of miRNAs or their hosts, and epigenetic changes.

Because miRNAs can target multiple genes, the effects that they have in cancer are context dependent. A specific mRNA might play a crucial role in one cancer type, and alterations in a miRNA targeting that mRNA could in turn also affect its development. On the other hand, this mRNA may be entirely absent in a different cancer, causing the miRNA targeting it to also have no effect.

This is, of course, a simplification. The targets of a miRNA compete with each other for binding, and the effect that a miRNA has will depend on the levels of each mRNA target in that particular cell. As a theoretical example, a tumor suppressor mRNA may be present in high amounts in a cell, and this could cause a miRNA targeting it to be saturated, causing fewer copies of that miRNA to bind to other targets. Should the expression of the tumor suppressor suddenly decrease, the miRNA is now able to regulate other targets more effectively. If these other targets are oncogenes, the miRNA's role can shift from oncogenic (by targeting the tumor suppressor), to tumor suppressive (by targeting the oncogenes), all due to changes in the expression of a single target.

One of the best examples of an oncogenic miRNA is miR-21-5p. This miRNA is overexpressed in many cancer types and is known to target important tumor suppressors, including *PTEN*, *PDCD4*, *TPM1*, and *HIF1A*[127,128]. The overexpression of miR-21-5p therefore contributes to several of the hallmarks of cancer, most notably resisting cell death and cell migration. Elevated levels of mir-21 have also been linked to other diseases, such as diabetes, hepatitis, and cardiac disease[129–131].

The machinery that controls miRNA biogenesis can also be disrupted in cancer. Overexpression of DICER, DROSHA, DGCR8, AGO1 and AGO2 has been observed in some cancer types[132,133]. Defects of the miRNA biogenesis machinery have also been observed in poorly differentiated tumors, causing a global downregulation in miRNA production[134]. Mutations can also affect the function of miRNAs. A single nucleotide variant in the miRNA binding site of an mRNA can disrupt the miRNA:mRNA interaction and lead to altered levels of the mRNA product. Similarly, mutations causing changes in the seed region itself can cause a miRNA to affect different miRNA targets[135].

Clinical potential of miRNAs

Similar to gene expression patterns, miRNA profiles in cancer can be used to predict patient survival and treatment response, and to define clinically relevant subtypes[134,136]. MicroRNAs are generally stable compared to mRNA molecules, and the miRNA profiles in formalin fixed, paraffin-embedded samples correlate better with fresh-frozen counterparts than mRNA profiles do[137]. The miRNA profiles of poorly differentiated tumors can also predict the cellular origin of the tumor with greater accuracy than mRNA profiles[134].

MicroRNAs also have the potential to be used as biomarkers, but there are a few caveats: In order to be useful as a biomarker, the monitoring of the molecule needs to be minimally invasive such as by being present in the blood stream. Several miRNAs have been identified in blood and other biological fluids, either as free-circulating molecules, bound to proteins, or inside small vesicles called exosomes[138,139]. Some miRNAs, such as mir-21, are expressed in many different cancer types and are therefore not suitable to be used as biomarkers for specific diseases. Using miRNAs as biomarkers therefore involves a panel of several miRNAs that together have a greater predictive power than any single miRNA. Such panels have been suggested for various cancer types, including breast cancer, hepatocellular carcinoma, and gastric cancer[140–142].

Due to their ability to target multiple deregulated mRNAs, miRNA-based therapeutics are a potential and unique approach to treating cancer. These therapies come in two forms: miRNA mimics and antimiRs. Both fall under the umbrella of RNA-based therapeutics, a field that has received much attention lately[143]. Both therapies also involve the delivery of artificial RNA molecules to their target tumor. These artificial RNAs usually have a modified phosphate backbone to increase stability, and can be delivered via various systems that allow them to pass into the cell[144].

MicroRNA mimics essentially mimic the function of a tumor-suppressive miRNA. They act as replacements or as an artificial way to boost the levels of certain miRNAs to restore their tumor-suppressive function. A drug mimicking the tumor-suppressive mir-34a reached phase I of a clinical trial, and although the trial was closed early due to immune-related toxicities, this first-in-human trial provides proof-of-concept of treating cancer using miRNA mimics[144]. A second miRNA mimic, this time targeting mir-16, successfully completed a phase I trial in 2017[145].

AntimiRs, on the other hand, target oncogenic miRNAs directly. These therapies use oligonucleotides that are complementary to the target miRNA, and once in the cell the antimiRs bind to and neutralize their target. An antimiR drug targeting mir-155 in cutaneous T-cell lymphoma completed a phase I trial and was undergoing phase II before being terminated early due to business reasons[146].

Despite their promise, both miRNA mimics and antimiRs both face significant challenges. Delivering these modified RNA molecules to their target tumor tissue remains a hurdle. They are also susceptible to degradation by enzymes (RNases) in the bloodstream and within cells, and efficient delivery methods are still under development. Additionally, off-target effects are a concern. These can occur in two ways: the therapeutic molecules might reach unintended cells or interact with similar sequences of other miRNAs, leading to unintended consequences. We will undoubtedly see more clinical trials for miRNA-based therapies in the future, but this field is still at an early stage.

mir-4728

Our understanding of the *ERBB2* oncogene took an unexpected turn in 2011, when we reported the discovery of a miRNA named mir-4728 encoded within one of its introns (Figure 6)[147]. The sequence for miR-4728-3p is encoded directly upstream of the 5' boundary of exon 24, making it a classic example of a mirtron. The levels of mir-4728 correlate well with the expression of *ERBB2*, with the 3p strand being present in much greater levels than the 5p strand, indicating that miR-4728-3p is the primary guide strand of this miRNA. Its location and strong correlation with *ERBB2* expression make miR-4728-3p an interesting subject of research. It presents a relatively unknown facet of the *ERBB2* locus: its ability to produce not only a well-characterized receptor protein but also a potentially co-regulatory miRNA.

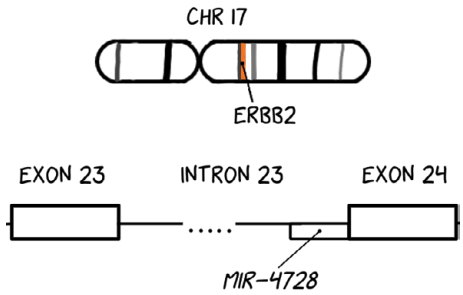

Figure 6
The miRNA mir-4728, located in the 24th intron of ERBB2 on chromosome 17.

All current therapies targeting *ERBB2* are directed against the function of the receptor protein itself, leaving the miRNA produced from the same locus intact. Given that a substantial portion of patients either do not respond to anti-ERBB2 therapies or relapse, investigating the role of this miRNA becomes even more critical[65]. Understanding the function of all genetic elements within this oncogenic locus, including non-coding elements like miRNAs, is needed for developing more comprehensive therapeutic strategies and understanding the mechanisms that lie behind drug resistance.

Over the years, we and other groups have investigated the function of miR-4728-3p, particularly in the context of ERBB2-positive breast cancer. Our research has revealed a link between miR-4728-3p and the oncogenic miR-21-5p. We found that miR-4728-3p is associated with a decrease in the poly(A) polymerase TENT4B which in turn is responsible for marking miR-21-5p for degradation via the poly(A)-specific ribonuclease PARN[148]. These results are particularly interesting because miR-21-5p targets the tumor suppressor PTEN, and this interaction has been shown to contribute to resistance to trastuzumab, a monoclonal antibody used as an anti-ERBB2 therapy[149]. This suggests that the ERBB2 locus itself encodes a molecule that could contribute to resistance to anti-ERBB2 therapy.

In addition to being linked to the oncogenic mir-21, we have shown that miR-4728-3p can also regulate the levels of the estrogen receptor alpha, *ESR1*[150]. This observation has been independently reproduced, and suggests another significant function of this miRNA with clear clinical implications[151,152]. The levels of ERBB2 and ESR1 tend to be inversely correlated in breast cancer, particularly in ERBB2-positive tumors[153]. This inverse correlation may partly be explained by ERBB2 overexpression leading to increased miR-4728-3p levels, which in turn negatively regulate ESR1. It is important to note that ERBB2 can also influence ESR1 levels via the PI3K/AKT signaling

pathway[154], and miR-4728-3p may present a potential alternative or complementary mechanism for ERBB2-mediated ESR1 regulation. Notably, ERBB2 amplification is associated with a poor outcome to endocrine therapies, particularly to the ESR1-targeting drug tamoxifen[155].

These results all suggest that miR-4728-3p has an oncogenic effect, but other studies have been performed that conflict with this statement. This miRNA was reported to exert tumor-suppressive effects in colorectal cancer by regulating key targets involved in focal adhesion signaling. Focal adhesions transmit regulatory signals between the cell and the extracellular matrix, and can contribute to cell migration and invasion[156]. Tumor-suppressive effects of mir-4728 have also been demonstrated in papillary thyroid carcinoma[157] and Burkitt lymphoma[158]. Importantly, the findings that show a tumor-suppressive effect of miR-4728-3p are performed in different cancer types. It is possible that miR-4728 does indeed have opposite effects depending on the cellular environment it is expressed in, and this goes to show the complicated nature of miRNAs in general.

Gene fusions

Origin

Genomic instability is an enabling characteristic of cancer cells that allows them to develop the hallmarks of cancer[7]. Mutations in the DNA repair machinery that build up over time can result in catastrophic and sweeping changes in the genome, where entire segments of a chromosome are deleted, amplified, or even joined with segments from other chromosomes. These genomic rearrangements are the result of double-stranded DNA breaks, and sometimes these breakpoints occur inside or closely adjacent to genes. This can result in an event known as a gene fusion – where parts of two distinct genes are erroneously joined together to form a new genetic element that is a hybrid of the two gene partners.

The first observed recurrent genomic abnormality in cancer was the so-called Philadelphia chromosome. It was first described in 1960 as a "minute chromosome", with "no other frequent or regular chromosome change" observed[159]. With the advent of cytobanding in the 1970s, researchers were able to identify changes in the cancer genome with greater precision, including the origin of the Philadelphia chromosome. It was found that this stubbed chromosome was the result of a translocation between chromosomes 22 and 9, and was present in 90-95% of all chronic myeloid leukemia

patients[160]. In the 1980s it was discovered that the genomic breakpoints of the Philadelphia chromosome translocation were located inside the genes *BCR* and *ABL1*, and that this new chromosome produced a chimeric RNA transcript that was translated into a protein (Figure 7)[161]. Because the Philadelphia chromosome is accompanied by few additional changes to the genome, and because the translocation involves a known oncogene, it was hypothesized early on that this genomic rearrangement was driver, not a result, of cancer progression.

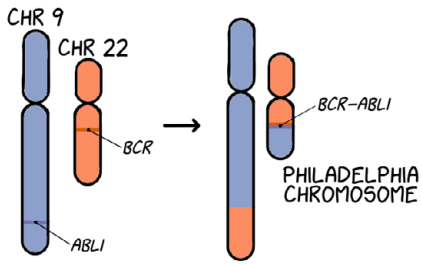

Figure 7
The Philadelphia chromosome is the result of a translocation between chromosomes 9 and 22. The genomic breakpoints are located inside two protein-coding genes, ABL1 and BCR. The Philadelphia chromosome produces a chimeric protein that is a driver of cancer progression.

Around the same time that the role of the Philadelphia chromosome in leukemia was described, the importance of other genomic rearrangements was discovered in other cancer types. Researchers discovered translocations in Burkitt's lymphoma that juxtaposes the oncogene *MYC* with parts of the immunoglobulin genes *IGH*, *IGK,* or *IGL*. Unlike the Philadelphia chromosome fusion, this translocation does not result in a fusion protein, instead it places an immunoglobulin enhancer element next to *MYC*, causing overexpression of the oncogene and thereby driving cancer progression[162].

We now know that gene fusions are common in many cancer types, including solid tissue tumors. Fusions can arise from a variety of chromosomal rearrangements, and balanced rearrangements have the potential to produce two reciprocal gene fusions (Figure 8). Next-generation sequencing (NGS) enables us to perform an unbiased search for fusion events across the whole genome or transcriptome. In most cases, fusions are now detected in the form of fusion transcripts found in RNA sequencing (RNA-Seq) data. The advent of NGS has reshaped our view of gene fusions, and studies of large pan-cancer patient cohorts have revealed that both recurrent and non-recurrent fusion transcripts can be found in most cancer types[163].

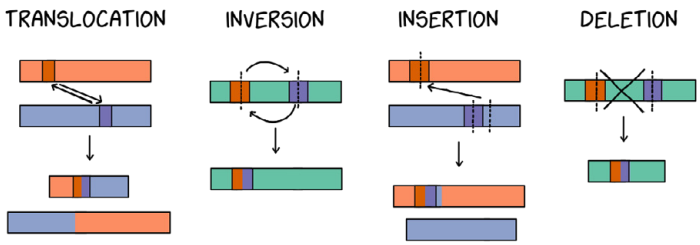

Figure 8
Genomic rearrangements can lead to gene fusions in several ways. Rearrangements can be both balanced (translocations, inversions, and insertions) and unbalanced (deletions).

While it is generally accepted that most gene fusions occur at the genomic level, other mechanisms have been proposed that may result in fusions, namely transcriptional read-through and trans-splicing[164,165]. However, these types of events have not been extensively investigated, and they are rare compared to genomic fusions[163,166].

Functional consequences in cancer

Gene fusions represent a diverse group of genetic aberrations with varying functional consequences. While some fusions create chimeric proteins with novel oncogenic properties, others disrupt normal gene regulation or lead to gene silencing altogether. Interestingly, the number of detectable fusion events within a tumor often correlates with the underlying genomic instability of the cancer[163]. There does not seem to be a rule as to which genes are involved in most fusion events, although they are generally thought to be associated with open chromatin structure[167]. Some fusions are also associated with locations in the genome that have been dubbed fragile sites[168].

Perhaps the most interesting way that fusions can impact a tumor is through the generation of a fusion protein that possesses functional domains from both fusion partners. This is the case for the BCR-ABL1 fusion protein that is the product of the Philadelphia chromosome. There, a self-regulating domain at the N-terminal of the ABL1 protein is replaced with a BCR domain that allows adapter proteins to bind and activate various downstream signaling pathways[169]. For a fusion event to be able to generate a chimeric protein, the breakpoints of the two genes generally need to align so that they generate an in-frame mRNA, and these events are of particular interest when detected.

As discussed previously with the translocation between *MYC* and immunoglobulin genes, gene fusion events do not necessarily result in a chimeric protein. This example represents another class of fusions, where the regulatory elements of one gene are

replaced with those of another gene. This can result in altered expression levels of the two fusion partners without changing the protein products themselves. In the case of the *MYC* fusions in Burkitt's lymphoma, the translocation places *MYC* under the control of powerful immunoglobulin enhancers, leading to constitutive overexpression and promoting uncontrolled cell growth[162]. Gene fusions such as this can be difficult to detect in an unbiased manner, as swapping out promoters or enhancer elements will not lead to altered mRNA sequences, causing these events to go undetected in RNA-Seq experiments.

Not all gene fusions are functionally relevant to cancer development. Unbiased approaches to detecting gene fusions have led us to discover that most fusion events found in cancer are non-recurrent and unlikely to be major drivers of tumor progression. Instead, these fusion events arise as passenger events, occurring coincidentally during tumorigenesis and without impacting cellular function or fitness. Distinguishing driver fusions from passenger events remains an ongoing challenge, hindered in part by the high error rate of fusion detection[170].

Lastly, it is important to note that the discussion of gene fusions in cancer typically focuses on rearrangements leading at the very least to functional chimeric transcripts, and not all rearrangements involving two protein-coding genes fall under this category. Some rearrangements can disrupt coding sequences or remove promoters, causing frameshift mutations, gene truncation, or rendering them incapable of being transcribed into RNA. These events arise through similar mechanisms as other fusions, but may contribute to cancer development by inactivating genes rather than creating novel functionalities[171].

Clinical relevance

Recurrent in-frame gene fusions are of particular interest in clinical practice. The presence of recurring partner genes in multiple fusion events indicates a strong selection pressure for that combination of genes. These fusions often involve a tyrosine kinase as one partner and result in the constitutive activation of its kinase domain. In some cases, these fusion events can be a defining characteristic for a specific cancer type or subtype[160,172–175]. A single gene can also be recurrent in fusion events but have multiple different partner genes. These fusion genes generally serve an important function in the tumors they are found in, and suggest a lesser role for the specific partner involved. Examples of this include *ESR1* fusions in breast cancer[176], *ALK* fusions in lung cancer[177], *EWSR1* fusions in Ewing sarcoma[178], and *RET* fusions in thyroid carcinoma[179]. Fusion events and specific fusion genes can also be recurrent across multiple cancer types. These fusions have the potential to be the target of basket clinical trials where patients

are grouped together based on their common gene fusions rather than the histological origin of their tumor[180,181].

The most clinically relevant gene fusions typically involve tyrosine kinases. Many such fusions are known to be recurrent in specific cancer types and are often associated with few other genetic abnormalities, indicating that they are strong oncogenic drivers[177,182]. These gene fusions are typically targeted with TKIs directed against domains encoded by the kinase fusion partner, rather than the chimeric protein itself. This approach arises from the high heterogeneity of gene fusion events. Even fusions involving the same two genes can have different breakpoints, potentially leading to diverse mRNAs or proteins with varying levels of activity. Targeting the common kinase domain with TKIs offers a broader solution to address this heterogeneity and effectively inhibits the oncogenic signaling driven by the fusion, regardless of the specific protein structure[182].

The prevalence of gene fusions in cancer is tightly linked to the underlying genomic instability of these malignancies[163]. While fusion transcripts have also been observed in benign tumors[183,184] and even other diseases[185], their occurrence is exceedingly rare compared to their frequency in cancer. This specificity makes them attractive candidates as biomarkers for cancer detection. However, the immense heterogeneity of fusions makes it difficult to predict the origin of a gene fusion detected in a liquid biopsy. Once a fusion has been detected, it can potentially be used to monitor treatment response. The use of gene fusions as a diagnostic tool has also been studied, and this usually involves detecting known recurrent fusions and are limited to specific cancer types[186,187].

Beyond targeting constitutively active kinase fusions with TKIs, significant knowledge gaps remain regarding the functional consequences of most gene fusions. While some tools have been developed to predict the oncogenic potential of a fusion, their utility is primarily limited to detecting fusions that retain kinase domains, neglecting the broader spectrum of potentially oncogenic fusions[188,189]. The highly heterogeneous nature of fusions further complicates the development of standardized testing methods for clinical use, but advancements in high-throughput sequencing are poised to improve their detection in clinical settings. Addressing false positive fusion events identified through RNA-Seq data analysis remains a challenge, in addition to the problem of how to separate functional fusions from passenger events.

Aims of this thesis

Overall aims

The studies in this thesis explore several knowledge gaps in cancer biology relating to gene fusions, miRNAs, and the combination of the two. We wanted to explore the role of miRNA hosts in gene fusion transcripts detected in RNA-Seq data, as these events may exert functional effects on the tumor even though they do not have protein-coding potential. A major problem in gene fusion research is the detection of false positive fusion events. Here, we developed a bioinformatic pipeline to validate the presence of fusion transcripts using matched WGS data. We also demonstrated that fusion prediction can be improved by applying this bioinformatic pipeline to large patient cohorts and constructing a machine learning classifier that can predict whether a fusion transcript is real or not. Finally, we explore the global function of the ERBB2-encoded miRNA miR-4728-3p in a breast cancer cell line.

Specific aims

Paper I

In a previous publication we observed that miRNA host-genes are over-represented in fusion transcripts in the breast cancer cohort SCAN-B[190]. Here, we aimed to reproduce these observations in a different cohort, and to explore the genes involved in these fusion events. Specifically, we were interested in seeing what genes were being used as 5' fusion partners in fusion events where the 3' gene was a miRNA host, as the 5' partner and its promoter control the rate of transcription of the fusion. We hypothesized that there would be some selection for these 5' partners, and that the genes involved in the fusion events would reflect the tumor phenotype.

Paper II

A challenge in gene fusion research is the accurate identification of these events. Fusions are typically detected as fusion transcripts in RNA-Seq data, but the performance of software tools for fusion transcript detection varies significantly, leading to inconsistencies and potential inaccuracies. In this study, we aimed to develop a method to validate fusion transcripts using matched WGS data.

Paper III

Building upon the fusion validation pipeline we developed in paper II, we expand on how we can improve gene fusion validation. Our aims for this paper were to apply our pipeline on data from a large patient cohort and to study the validated fusions found there. In addition, we aimed to improve fusion validation using machine learning using the information from these validated fusions. This would allow us to predict whether an observed fusion is real or not, based only on information obtained from RNA-sequencing.

Paper IV

In 2011 the research group published a paper describing the miRNA mir-4728 encoded in an intron of the oncogenic *ERBB2*. In this manuscript, we aimed to explore the global effect of this miRNA on both gene expression and translation. Using a method called polysome fractionation, we attempted to identify which genes and pathways are affected by miR-4728-3p.

Materials and methods

Cohorts

The studies in this thesis utilize sequencing data from patient tumors that come from two main sources: The Cancer Genome Atlas (TCGA) program and the Sweden Cancerome Analysis Network – Breast (SCAN-B) initiative.

TCGA is a landmark project funded by the National Cancer Institute and the National Human Genome Research Institute in the United States. The program focuses on many cancer types, each represented by a distinct cohort. At the time of writing, TCGA repository contains approximately 10,000 cases encompassing twenty different cohorts. The publicly available TCGA data that we used in our projects were expression matrices for both protein-coding genes and miRNAs, reverse phase protein arrays for protein quantification, methylation arrays, and patient metadata information such as receptor status for breast cancer patients. We also used unprocessed RNA-Seq and WGS data to analyze fusion sequences. As these are sensitive data, access was obtained via a project application and the data were stored on a GDPR-compliant high performance computing cluster.

The SCAN-B initiative was launched in 2010 as a population-based observational study between seven hospital centers in southern Sweden. The aim of this initiative is to improve the understanding of breast cancer biology through molecular profiling, and to create a population-based material of breast cancer that includes most of the new cases that occur in southern Sweden. All patients with a newly diagnosed breast cancer case are given the chance to participate in the study, which involves providing a blood sample and a piece of the tumor. RNA-Seq is then performed on the tumor sample and its molecular landscape profiled. In our projects, we used both raw RNA-Seq data and gene expression matrices, as well as WGS data from a subset of triple-negative breast cancer patients.

Cell lines

Besides using tumor samples from patient cohorts, we also utilize human cell lines for both *in silico* and *in vitro* analysis. These are cells derived from tumors and have been immortalized, meaning they can keep multiplying indefinitely without entering cellular senescence. Human cell lines are convenient to work with in a laboratory setting and are widely used in cancer research. This means that many commonly used cell lines are extensively characterized, and multiple different types of data are available for each one. In our work, we utilized publicly available WGS and RNA-Seq data for several cell lines, provided by the Cancer Cell Line Encyclopedia[191]. We also used published data that describes the fusion landscape of two breast cancer cell lines, BT-474 and MCF7[192–195].

Next-generation sequencing

Next-generation sequencing (NGS), encompassing both RNA-Seq and WGS, has emerged as a cornerstone technology in modern biology and medicine. Since its commercial debut in 2005, NGS has experienced dramatic cost reductions and increased efficiency, fueling its adoption in basic and translational research. The most common form of NGS uses a method called sequencing by synthesis that involves the incorporation and detection of fluorescently labeled deoxyribonucleotide triphosphates (dNTPs) to a DNA template strand. This process is performed on millions of template strands simultaneously, allowing for massive amounts of genetic material to be rapidly sequenced. Sequencing by synthesis always uses DNA as a template. To sequence RNA molecules, they are first converted to cDNA prior to sequencing.

The first step of a typical Illumina NGS workflow is library preparation. DNA or RNA molecules in the sample to be sequenced are randomly fragmented, and adapter oligonucleotides are ligated to the ends of each fragment. These adapters contain the sequencing primer site, a barcode sequence that uniquely identifies the sample that the fragment belongs to, and capture sequences so that the fragment can bind to the flow cell. For RNA sequencing, the fragmented RNA molecules are first converted into cDNA before adapter ligation. Following library preparation, the fragmented DNA molecules are loaded onto the flow cell – a glass or plastic panel with lanes that allow sequencing reagents to flow through it. Each flow cell is coated with two types of oligonucleotide probes that are complementary to the capture sequences of the adapters, and when the library is loaded onto the flow cell each molecule hybridized to one of these probes.

Bridge amplification

Once the library has been loaded onto the flow cell it is amplified via a process called bridge amplification. A polymerase synthesizes a sequence that complements the hybridized DNA strand. The resulting double stranded DNA is then denatured, causing the original template to be removed, and leaving the complementary sequence that extends from the probe bound to the flow cell. This allows the remaining sequence to hybridize to a nearby probe that is complementary to the capture sequence of the other end, creating a single stranded oligonucleotide bridge between the two probes. A polymerase synthesizes the complementary strand, and the double stranded bridge is denatured leaving two complementary DNA strands, both of whom are bound to the flow cell. This process is known as bridge amplification and continues until a cluster of strands forms on the flow cell for each of the original library molecules bound. These clusters originally contain both the forward and reverse strands of the amplification process, but then the reverse strands are removed, so that the final clusters are each comprised of identical, forward-strand DNA molecules.

Sequencing by synthesis

Next, fluorescently labeled dNTPs are introduced to the flow cell. The dNTPs contain a reversible terminator modification that prevents more than one dNTP from binding to each strand at a time. At this point a light source excites the labeled dNTPs that then give off a signal that is unique to each of the four nucleotides. The sequencing instrument measures the signal in each cluster, and determines which nucleotide is present in that part of the sequence. The terminator modifications are then removed from the bound dNTPs, allowing the next one in the sequence to hybridize to the strand. This process is called sequencing by synthesis, and it continues for a predetermined number of cycles. Many modern NGS methods employ so-called paired-end sequencing. Here, after the bound DNA strand has been sequenced by synthesis, the fragments are denatured and the complementary strands are sequenced from the opposite end, generating a second sequence read for the cluster. Paired-end sequencing has several advantages over single-read sequencing. Knowing the sequences on either end of each DNA fragment makes it easier to align to the genome, especially if the fragment is in a repetitive region of the genome. It can also aid in the identification of structural changes in the genome, such as gene fusions.

Aligning NGS data to a genome

The output of an Illumina sequencing run is stored in a text-based format, typically FASTQ, that contains both the sequence of each read generated on the flow cell and a corresponding quality score for each base in the read. These sequences are then aligned to a reference genome. Aligning reads involves finding the most likely positions in the reference genome where each read originated. This process accounts for potential sequencing errors, mutations such as insertions or deletions, and repetitive or low-complexity sequences. Most aligners use heuristic algorithms to find the most likely alignments, prioritizing those with fewer mismatches and gaps compared to the reference.

The alignment process assigns each short read from the sequencer a position within the reference genome or transcriptome, creating a map of where each fragment originated. The results of this mapping are typically stored in either SAM (sequence alignment map) or BAM (binary alignment map) file formats. Both formats contain the same core information, including the reference sequence location for each read, any mismatches between the read and the reference, and insertions or deletions identified during alignment. However, SAM files present this data in a human-readable text format with tab-delimited columns. This allows researchers to visually inspect the alignments, but the large size of these files can be cumbersome for storage and analysis. In contrast, BAM files represent the same data in a compressed binary format, significantly reducing file size and making them more efficient for downstream computational analyses. While not directly readable by humans, BAM files can be readily converted back to SAM format for detailed inspection if necessary.

Fusion detection

A large part of the work that went into this thesis focused on analyzing gene fusions. We used a total of three different software to detect gene fusions: Arriba[196], STAR-Fusion[197], and FusionCatcher[198]. In principle these tools all work in a comparable way to each other: they query raw RNA-Seq data and look for reads or read pairs that support a fusion event. There are two main ways that a fusion caller processes RNA-Seq data; either through genomic alignment or via *de novo* transcript assembly. In the first case, reads from RNA-Seq data are aligned to a reference genome or transcriptome, and then a search is made for fusion-supporting reads. In the second case, full-length transcripts are assembled without genomic alignment, followed by the identification of chimeras. Mapping-based fusion callers are generally quicker and more computationally efficient, however they are sensitive to mapping errors and do poorly

at detecting complex rearrangements. Assembly-based callers, however, are not limited to existing annotations but often have higher false-positive rates due to assembly errors. All three fusion callers that we used employ a mapping-based approach, as we were mostly interested in standard gene fusion events and wanted to avoid as many false-positive events as possible. Sequencing reads that support a gene fusion event can be classified into one of two groups (Figure 9). Discordant read pairs have each read mapped on either side of the fusion junction, with no overlap of the junction itself. Chimeric reads have one member of the read pair overlap the fusion junction. Fusion callers typically report in their output files the number of both types of reads, and this can be used to assess the confidence of a given fusion transcript.

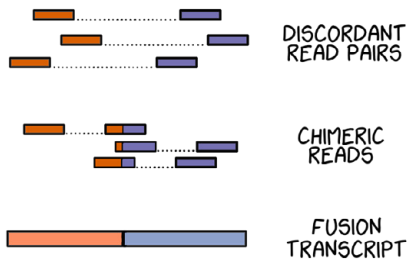

Figure 9
Discordant and chimeric sequencing reads supporting gene a gene fusion.

In practice, although the three fusion callers that we used function in a comparable way to each other, and to most other fusion callers, their output of predicted fusions can differ dramatically. This is true not only for these three tools, but for all fusion calling algorithms. Comparison studies have found that fusion callers produce many false-positive events and that there is a high degree of variance between them[199,200]. Many fusion callers, including the ones we used, have intrinsic filtering and scoring steps that are used to filter out likely false-positive events and to assign a confidence score to observed fusions based on factors such as read evidence, chimeric junction quality and alignment characteristics. Additional filters are often present, such as steps to remove known false-positive events, PCR artifacts, homologous reads, and read-through transcripts[200]. However, these steps are not foolproof and might explain a large portion of the variance that is observed between fusion callers.

Benchmarking fusion callers is also difficult, as establishing a ground truth of fusions present in a sample is almost impossible without them. Many benchmarking studies use simulated data[196,197], but it is unclear how accurately simulations can reflect the real-world complexity of a tumor sample. Other studies have used fusions that have been

experimentally validated, and while this is robust, it is often limited to cell lines due to lack of tumor material[199,201]. WGS, while offering a comprehensive view of the entire genome, also has limitations for fusion detection. Like RNA-Seq, fusion detection at the WGS level is susceptible to alignment errors. Unlike RNA-Seq, which focuses on transcribed sequences, WGS captures all DNA elements, including non-coding regions. This can lead to the identification of irrelevant fusion events that never get transcribed into RNA and lack biological significance.

Machine learning

Machine learning is a branch of artificial intelligence that focuses on creating algorithms that mimic the learning process of a human being. This involves going through an iterative learning process, where the algorithm analyzes its past performance, adjusts its internal parameters, and then applies what it has learned to new data. This process, called training, gradually improves the model's ability to perform specific tasks without being explicitly programmed for each situation.

Machine learning can be used to solve a very wide range of problems, from recognizing images and audio to classification and regression. In this chapter, we will focus on a subcategory of machine learning called supervised learning algorithms. These algorithms are constructed using data that includes both the desired input features, such as the expression of several genes, and the output, such as what cancer subtype a sample belongs to. This type of data is usually called "labeled", and a model's performance is reliant on the fact that the data are correctly labeled.

The simplest supervised machine learning model is linear regression. In this case, the algorithm models the outcome based on the input features by fitting a linear equation to the data. Another simple machine learning algorithm is the decision tree (Figure 10). This algorithm works by splitting labeled data into subsets based on what features best separate the outcome on a continuous scale or into distinct categories. At each node of the tree, a decision is made based on the value of a selected feature. This leads to recursive splitting at the node based on new features, creating branches of decisions. This process of splitting the data continues until a certain criterion is met, such as the maximum depth of the tree is reached. The nodes at the end of each branch are called leaves and represent the final prediction of the outcome.

The strength of models such as linear regression and decision trees is that they are very intuitive – it is easy to explain how the model reaches a certain conclusion based on a simple formula or by following the branches of a decision tree. The limitation of these

models also lies in their simplicity. When the data becomes complicated and high-dimensional it becomes difficult to generalize observations with a simple linear equation. Decision trees tend to overfit their training data, especially if they are allowed to grow too deep. The results from a single decision tree therefore tend to not be generalizable, and small variations in input data can result in vastly different predictions. One way that this can be addressed is via an ensemble learning approach. This approach is built on the idea that multiple weak learners can collectively make predictions that are much more robust than any single learner in the ensemble. A classic example of an ensemble learning method is the random forest model - a collection of decision trees where the output of the model is the average prediction of all the trees (Figure 10).

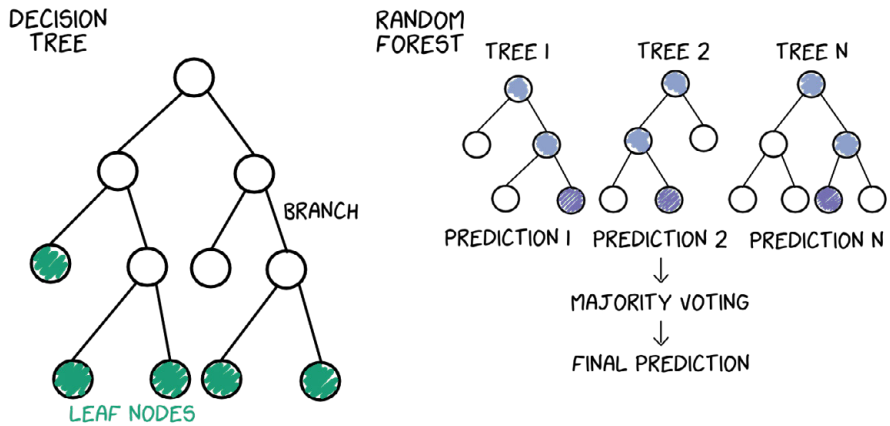

Figure 10
Decision tree and random forest.

Another example of an ensemble learning approach is a technique called gradient boosting. Similarly to random forest, this technique combines multiple weak learners (usually decision trees) into a single, stronger model. Unlike random forest, gradient boosting trains weak learners sequentially rather than all at once. Each subsequent learner focuses on correcting the errors that the previous learner made by comparing the prediction to the labeled outcome. This can result in a surprisingly strong model, with predictions that are much more robust than those from a single weak learner. Models employing gradient boosting are also typically much faster to train compared to random forests, making them very efficient for large datasets.

LightGBM

In this thesis we employed the use of a relatively obscure gradient-boosting framework called LightGBM to predict whether observed fusion transcripts are real or false. LightGBM, short for light gradient-boosting machine, is a fast, efficient, and robust machine learning framework that is built around the concept of gradient boosted decision trees. LightGBM differs from other gradient boosting frameworks in several ways, with most changes aimed to further improve speed and scalability. Tree-based algorithms usually grow their trees level-wise, meaning that each iteration of the tree-building process increases the length of every branch in the tree by one node. LightGBM instead uses a leaf-wise tree growth approach, where a single leaf is selected to be split in each iteration (Figure 11). The leaf with the largest potential reduction in loss is chosen to be split, and this results in trees that have branches of different lengths. To determine which leaf to split, LightGBM uses two novel techniques: gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB). The GOSS algorithm reduces the number of data instances that need to be examined during the tree growth process, while EFB groups features together to further improve computational efficiency[202].

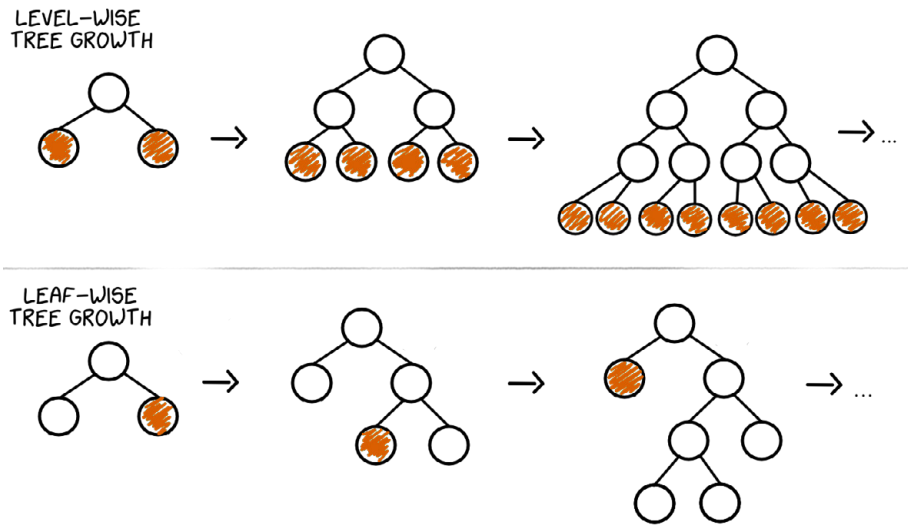

Figure 11
Level-wise tree growth and leaf-wise tree growth of decision trees.

Evaluating machine learning classifiers

One of the most important considerations when training a new machine learning model is how to properly assess its performance. While it may be tempting to focus solely on achieving perfect classification on the training dataset, the true challenge lies in creating a model that generalizes well to unseen data. This is due to the risk of overfitting - a model can capture noise or patterns that are unique to the training data, but when it is applied to new data it performs no better than random guessing. It is therefore essential to evaluate a model's ability to generalize and its overall performance on new, unseen data.

Model performance is usually assessed by partitioning the training data, often reserving 20-25% of the observations randomly chosen as testing data. This unseen portion remains untouched during model training and serves as the final assessment of the model's performance and its ability to generalize. To prevent overfitting on the training data, machine learning workflows often employ a method called resampling. This involves withholding a certain portion of the training data and using that for an initial evaluation, much like the initial splitting of the data into training and testing groups (Figure 12).

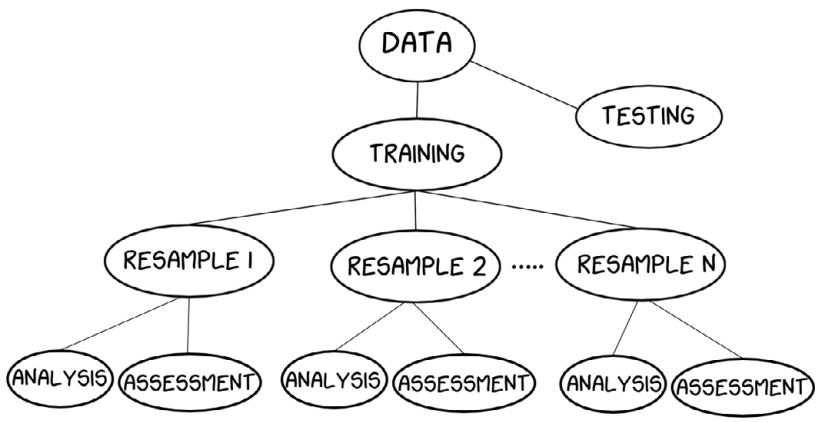

Figure 12
Data is split into training and testing sets. The testing data is untouched during training, and is only used as a final evaluation of the machine learning model. To prevent overfitting, training data data can be further split into resamples.

In our models we used a resampling technique called 10-fold cross-validation. There the data is split into 10 equally sized groups, and the model goes through 10 rounds of training. In each round, one of the groups is withheld for evaluation while the model

is trained on the remaining nine. Thus, by using a resampling approach, we will already have an idea of how the model will perform on unseen data before we ever introduce the unseen testing data. This is especially important because in practice, multiple models are typically trained at the same time, each with a different combination of parameters that control the learning process. These parameters, known as *hyperparameters*, play a significant role in determining a model's performance. It is important to identify an optimal combination of hyperparameters to achieve the best results for the task at hand. Note that *hyperparameters* are distinct from the model's *parameters*, which refer to the values inherent to the data that are being learned and adjusted during training. Examples of hyperparameters in the LightGBM framework include how many trees to construct and the maximum depth of each tree, while parameters refer to features such as which variable to select for the initial splitting of the decision tree.

The performance of a machine learning classifier can be evaluated by many different metrics. The models that we have created have all been binary classifiers, i.e., the outcome for a given observation can be one of two classes. The predicted positives (PP) and predicted negatives (PN) can therefore be compared to the actual positives (P) and actual negatives (N) in a 2x2 matrix called the confusion matrix (Figure 13). Many performance metrics can be calculated from this matrix including accuracy, recall (sensitivity, true positive rate), specificity (true negative rate) and precision (positive predictive value). The choice of which metric to use to evaluate a model is an important decision and depends on the specific use case.

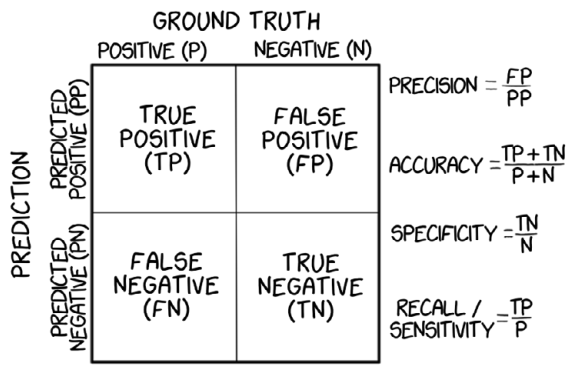

Figure 13
A confusion matrix can be used to show the performance of a machine learning classifier. For binary classifiers, the outcomes of the classification can be summed up in a 2x2 matrix by comparing them to the ground truth. Other performance metrics can be calculated from this matrix, such as precision, accuracy, specificity, and recall.

Many classification models output probabilities for each prediction that indicates the likelihood of an observation belonging to a specific class. However, to make a definitive classification, a classification threshold must be set that acts as a decision line – predictions above it are classified as one class and those below it fall into the other. Classification thresholds affect the performance metrics of a model and allow for trade-offs in performance. For example, a high threshold value would minimize false positives, but miss some true positives. Conversely, a low threshold would capture most true positive events, but also increase the rate of false positives. It is important to assess model performance over a range of thresholds, and metrics such as the receiver operating characteristic (ROC) curve and the precision-recall (PR) curve can be used to visualize these trade-offs (Figure 14). The ROC curve plots the trade-off between the true positive rate (recall / sensitivity) and the false positive rate, while the PR curve shows the trade-off between precision and recall. Calculating the area that falls under each curve summarizes the capabilities of a model over all threshold values, and these metrics are commonly used to measure performance.

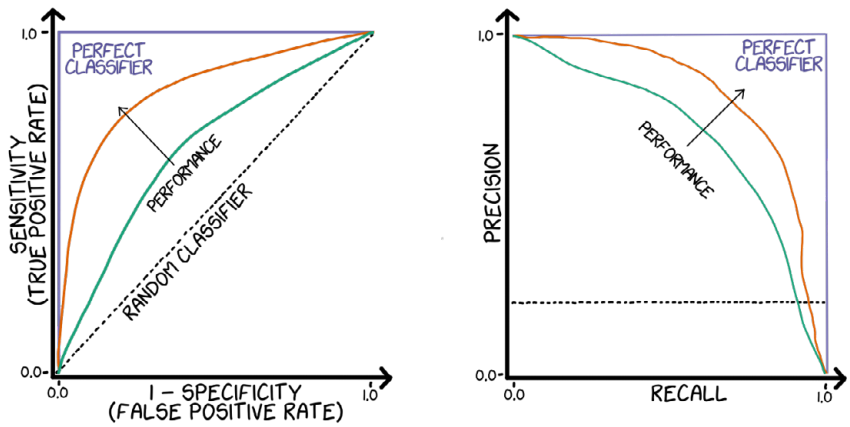

Figure 14
Receiver operating characteristic (ROC) and precision-recall (PR) curves visualize the trade-offs of common metrics on a range of classification thresholds.

Gene overrepresentation analysis

Gene overrepresentation analysis (ORA) is a bioinformatics method used to identify and characterize biological pathways, functional categories, or molecular processes that are significantly enriched with genes of interest compared to what would be expected

by chance. This approach is widely used in genomic studies to gain insights into the underlying biological mechanisms associated with a set of genes and is commonly used when analyzing transcriptomic data.

The core principle of ORA lies in comparing the overlap between a set of genes of interest (often termed query genes) and a pre-defined gene set representing a biological category, such as a metabolic pathway, Gene Ontology term, or protein complex. Statistical tests, typically the hypergeometric distribution, are then employed to assess the significance of this overlap. A statistically significant enrichment implies that genes within the pre-defined biological category are overrepresented amongst the query genes compared to what would be expected by chance. Identification of enriched categories can point towards specific biological processes that are potentially dysregulated under the investigated conditions. This information can be then used to formulate hypotheses and guide subsequent experiments.

However, there are some limitations that are inherent to ORA. This method can be sensitive to the number of query genes, and the choice of which gene sets to check for overrepresentation can impact the biological interpretation of the results. In our studies, we primarily used common curated gene sets such as those from the Gene Ontology consortium[203], KEGG[204], and REACTOME[205]. ORA is also based on the presence or absence of genes within categories, and therefore may not capture the magnitude or direction of their expression changes. ORA is therefore less commonly used in differential gene expression analysis, where the change in expression is quantifiable, making the data more suitable for other techniques.

Gene set enrichment analysis

Gene set enrichment analysis (GSEA) a technique that, like ORA, is used to identify relevant biological pathways or other gene sets in a given dataset[206]. GSEA works by ranking genes based on some metric the reflects the differential expression between biological states. This is typically the $\log_2$-fold change of a gene between treatment and control conditions, but can also be adjusted e.g., by dividing this value by its corresponding adjusted p-value. Once a ranked list of genes has been made, GSEA assesses whether genes within a predefined gene set are statistically enriched at the top or the bottom of the list. This is done by calculating an enrichment score, that reflects the degree to which the genes in the set are overrepresented at the top or bottom of the ranked list. The enrichment score is computed by walking down the ranked list of genes, increasing a running-sum statistic when a gene in the set is encountered, and decreasing it when encountering a gene that is not in the set. The maximum deviation

from zero of this running-sum statistic corresponds to the enrichment score for the gene set. Enrichment at the top suggests a coordinated upregulation of the genes within the set, potentially indicating an activated pathway under that condition. Conversely, enrichment at the bottom implies a coordinated downregulation and a possible pathway suppression.

Compared to looking at the overlap within categories in ORA, GSEA offers a more nuanced analysis of the data. By considering the entire ranked list of genes and enrichment at both ends of the spectrum, GSEA captures coordinated changes within gene sets, is less sensitive to background gene set selection, and can potentially uncover subtle but biologically relevant variations in gene regulation.

Results and discussion

Paper I

The current consensus among gene fusion researchers is that fusion events can impact the cell in three ways. First, they can result in a chimeric protein that has some properties of one or both fusion partners. This usually involves the constitutive activation of a tyrosine kinase, and these events typically produce the most obvious phenotype of all gene fusion events. Second, fusion events can cause the juxtaposition of regulatory elements of two genes. This does not result in a chimeric protein, as the mRNA product of the fusion would be unchanged, but it can lead to a change in expression in one or both fusion partners. Finally, gene fusions can lead to silencing of the fusion partners. Thanks to advances in high-throughput sequencing technologies, we now know that gene fusions are relatively common events in most cancer types, and in many cases they are transcribed as a chimeric mRNA. It is thought that only a small number of gene fusion events are functional, and most of the fusion transcripts observed in RNA-Seq data are passenger events that arise as a result of the unstable nature of cancer genomes. However, the fact that many fusion events are transcribed into mRNAs has interesting implications for non-coding RNAs. In many cases, small non-coding RNAs (sncRNAs) are located inside the loci of larger protein-coding genes, and they are often processed co-transcriptionally with their hosts[207–209]. This implies that if a host-gene is involved in a fusion event, and the sncRNA lands inside the fusion transcript, the sncRNA may still be expressed and processed regardless of the protein-coding potential of the fusion transcript itself. We have previously observed that host genes of both miRNAs and small nucleolar RNAs (snoRNAs) are specifically enriched in fusion transcript events in breast cancer[190,210]. This suggests that, contrary to popular belief, many gene fusion transcripts may be impacting cancer progression by deregulating the expression of intronic miRNAs and snoRNAs. In cases where the sncRNA host is a 3' fusion partner, the fusion event would effectively be a promoter-swapping event for that sncRNA.

In paper I, we expanded on these findings, and focused on the 5' partner usage of miRNA host fusions in breast cancer. We analyzed fusion transcripts detected in a total of 2632 breast tumors from the TCGA-BRCA and SCAN-B cohorts, and confirmed

that miRNA host genes were more likely than other genes to be involved in fusion events. Interestingly, we found that the expression of 5' fusion partners was on average higher when the 3' partner was a miRNA host gene, supporting the hypothesis that fusions may be a mechanism to upregulate oncogenic miRNAs. Following these observations, we found that gene sets involving EMT, the extracellular matrix, and focal adhesion were consistently overrepresented among fusion genes, regardless of orientation or the partner gene miRNA host status. We also saw overrepresentation of pathways related to the molecular subtype of the sample, such as estrogen response in the Luminal A and B subtypes. There were also differences between fusion partners of miRNA hosts vs non-hosts. Fusion partners of miRNA hosts were more often involved in translation, with overrepresented gene sets such as eukaryotic translation initiation and elongation. Interestingly, these gene sets were overrepresented regardless of subtype, but only in fusion events where the miRNA host was the 3' fusion partner. This indicates that these gene fusion events may be controlled by different transcriptional programs. We therefore investigated which transcription factors were regulating the expression of 5' fusion partners of miRNA hosts using data from the UniBind database[211]. The fusion genes were generally regulated by transcription factors known to have oncogenic effects, such as the components of AP-1. In some cases, the fusion genes also represented target genes of subtype-specific transcription factors, such as targets of the estrogen receptor being enriched in ER-positive subtypes.

Next, we looked at specific miRNAs involved in fusion events. In our data we observed 80 miRNAs that had significantly higher expression in samples where their host gene was part of a fusion event. Among these were the oncogenic mir-21 and the mir-106b~mir-93~mir-25 miRNA cluster. In many cases, the 5' fusion partners in these events had higher expression than the miRNA host gene, indicating that these fusions may contribute to the elevated miRNA levels. We did not see a single miRNA whose expression was significantly downregulated when the host gene was involved in fusion events. While this observation may support our hypothesis that this is a mechanism to upregulate miRNA expression in cancer cells, it may also be the result of how we detect gene fusions. In this study, fusions are detected as fusion transcripts in RNA-Seq data, meaning that we are inherently only detecting those events that are transcribed into chimeric mRNAs and therefore have a biased view on how fusions can impact expression. Fusion prediction in RNA-Seq data is also error-prone, with varying levels of sensitivity and specificity depending on which software is used.

Overall, our results suggest that gene fusions in cancer may allow the cell to uncouple the expression of miRNAs from their host genes. Many fusion events that previously would have been classified as passenger events may still be contributing to cancer progression via this mechanism, regardless of the protein-coding potential of the fusion transcript itself.

Paper II

As has been previously discussed, fusion prediction from RNA-Seq data is error-prone. There are many different software available that detect fusions transcripts, but their performance is discordant when compared. In paper I we attempted to experimentally validate 11 fusion events found in SCAN-B samples using real-time quantitative RT-PCR, and although these events were chosen for having high levels of support at the RNA level, only seven of them were validated experimentally. While primer design limitations on our part might have played a role, this result highlights a significant challenge in gene fusion research: a substantial portion of fusion transcripts detected in RNA-Seq are false positive events and do not reflect true genomic rearrangements. We also performed manual validation of fusion transcripts involving the oncogenic mir-21 and its host gene *VMP1* by examining WGS data from the same samples. Using the reported fusion junctions, we were able to identify reads at the DNA level that supported all the observed *VMP1* fusion transcripts. In this project, we developed a bioinformatic pipeline that automates this process and expands on it. The pipeline takes information from fusion transcripts identified in RNA-Seq data and searches for evidence for them in WGS data from the same sample.

The concept for our pipeline is relatively simple. The observed fusion junction and the orientation of the two fusion partners gives us a relatively small window in the genome that a breakpoint can be located in, and we can search this region for reads supporting the event. Like most software that detect fusions, our validation pipeline utilizes paired-end sequencing data, and the first step of the pipeline is to search for discordant read pairs supporting the fusion. These are paired sequencing reads where one read maps to one gene involved in the fusion, and the other read maps to the other fusion partner gene. Since these reads come from the same DNA fragment, finding discordant pairs indicates that the two genes might be fused in the genome. If read pairs that support the fusion transcript are found, we can then query the regions close by to see if we can pinpoint the exact genomic breakpoint. To do so, we look for reads that have soft-clipped ends with sequences that align to the other fusion partner. This approach allows us to search a very limited region of the genome for fusion-supporting reads, minimizing the chances of interpreting noise in the data as evidence for a fusion. Another strength of our pipeline is that it is fusion-caller agnostic, with the only requirement being the identities of the two genes and the coordinates of the observed fusion junction – information that is provided by virtually all fusion prediction algorithms. Pipelines built on similar ideas have been made before, but they either involved identifying fusion transcripts from scratch or are built into larger pipelines and have not been released as standalone tools.

Benchmarking tools like this can be challenging. It is difficult to find data that has a well-documented ground truth of what gene fusions are present in the sample and has both RNA-Seq and WGS data available. In this paper we evaluated our pipeline by using experimentally validated fusions in eight different cell lines. We used FusionCatcher to generate the initial list of fusion transcripts to be validated. Our pipeline found evidence for approximately 80% of the previously reported gene fusions events for these cell lines. In addition, we also found evidence at the DNA level for 35 fusion events in MCF7 and BT-474 that had not been previously reported. We also applied our pipeline on patient samples from four TCGA cohorts to validate the presence of predicted gene fusions. In addition to being fast, our pipeline was more sensitive than other tools that could be used to detect gene fusions at the DNA level.

As an alternative to using cell line data to evaluate the pipeline, we could have used synthetic reads to simulate gene fusions in both RNA and DNA. However, we were wary in using this approach as it is unclear how well synthetic data reflects the complexities of a real tumor. Our pipeline is limited by the fusion prediction software that is used to provide the initial list of fusions to validate in WGS – if that software does not detect a fusion transcript to begin with, we will not validate it at the DNA level. Choosing a robust fusion predictor to pair with our pipeline is therefore critical, and may influence the results of downstream analysis. Sequencing depth can also be a limiting factor, both for RNA-Seq and WGS. Low depth in RNA-Seq may cause the fusion predictor to miss the event, and low WGS depth may result in a lack of evidence for the event at the DNA level.

Paper III

In paper II, we developed a bioinformatic pipeline to validate fusion transcripts detected in RNA-Seq data at the DNA level using matched WGS data. In this paper, we applied our pipeline to 910 tumors from 11 different cancer types in the TCGA cohort. This allowed us to identify over 4000 fusion transcripts that were validated at the DNA level. We also evaluated the specificity of the pipeline by applying it to fusion transcripts detected in normal tissue samples, with the expectation that they primarily represented false positive observations. Indeed, we only saw evidence for very few of these fusion events, and many of them were found in the same sample indicating either sample contamination or problems during library preparation. For most of the validated fusion events, we also identified at least one genomic breakpoint, allowing us to look at the motifs that give rise to fusions. We found that there was significantly more microhomology between the genomic breakpoints of fusion partners than one

would expect by random chance. The average sequencing depth by the genomic breakpoints was greater on the side that encompassed the fusion transcript, indicating that fusions commonly arise from amplification events.

Utilizing WGS data is a robust way to validate fusion transcripts, but it is relatively uncommon to have both WGS and RNA-Seq data available for the same tumor sample. We therefore used the gene fusion events we validated using our pipeline as a ground truth and trained a supervised machine learning classifier to predict which fusion events are real. The classifier was trained using only features obtainable from RNA-Seq data or common annotation sources, so that it is applicable on samples that do not have matched WGS data. The testing data we used to generate the final evaluation metrics for the classifier consisted of fusions detected and validated in 249 triple-negative breast cancer samples from the SCAN-B cohort. It was important to use testing data that came from an independent cohort, as this ensured that we were not overfitting our data to noise that was present in the TCGA cohorts. The results of the classifier were promising, and we showed that biological interpretations on the set of predicted gene fusions would reflect the ground truth closer than that of the unfiltered set of fusions.

It has been shown that fusion detection software differ significantly in both specificity and sensitivity, and the overlap between predicted fusion transcripts is generally low. We wanted to demonstrate that a machine learning approach could improve fusion detection regardless of which fusion detection software was used. To accomplish this, we trained a second classifier, this time using the validated results of a different fusion detection software. We found that a machine learning-based filtering approach achieved robust performance metrics on the testing data for both software. We then compared the results of the classifiers to "classical" filtering methods, such as keeping only in-frame fusions or events labeled as "high confidence" by the fusion detection software. Our machine learning-based filtering approach consistently outperformed these classical filtering methods, and is a proof-of-concept that by using machine learning we can improve fusion prediction.

Another generally accepted approach to minimize false positives in gene fusion research is to employ a so-called ensemble approach, where the results of several fusion prediction software are combined to create a list of high confidence fusions that are detected by multiple algorithms. Using the fusion events validated by our pipeline, we show that such an approach is generally suboptimal, with many true fusions being excluded and false positives being kept. The fact that fusion transcript prediction generates so many false positive fusion transcripts raises the question: is the observation that miRNA host genes are enriched in fusion events that we made in paper I still reproducible? Perhaps these results were just an artifact due to so many false positive events in the data. To test this, we applied the same logistic regression model we used

in paper I to the validated set of fusion events from all 11 cancer types in TCGA. We found that the miRNA host status of a gene still positively influenced the likelihood of it being involved in a fusion event, even after limiting the analysis to only validated events. Here we had much fewer fusion events to work with, and therefore did not split the analysis between different cancer types. In future projects it would be interesting to see if the miRNA host gene enrichment is more prominent in certain cancer types, and if the 5' partner usage of events with 3' miRNA hosts continues to reflect the tumor phenotype.

Paper IV

In 2011, our research group reported the discovery of the miRNA *mir-4728*, located within an intron of the *ERBB2* oncogene. Subsequent research, including our own, has revealed intriguing connections between this miRNA and other cancer-related factors, such as ESR1[150] and miR-21-5p[148]. In this paper we look at the global effects of miR-4728-3p – the main mature product of mir-4728.

As with other miRNAs, miR-4728-3p has a set of predicted target genes based on complementary regions in mRNA 3' UTRs. However, these predictions often fail to capture the true biological impact that a miRNA has, as it is heavily dependent on the relative abundance of all its putative targets. MicroRNAs can regulate their targets in two ways: by either catalyzing the degradation of the target mRNA or repressing its translation. We therefore wanted to study the effects of miR-4728-3p in ERBB2-positive breast cancer, both at the mRNA and protein level.

To achieve this, we blocked the activity of miR-4728-3p in the ERBB2-positive cell line SK-BR-3 using antisense oligonucleotides. Subsequently, we performed a polysome fractionation, a technique that separates mRNAs in cell lysate based on their ribosome occupancy via ultracentrifugation. The mRNA molecules with multiple ribosomes bound to them are typically associated with actively translated genes and are collectively called polysomes. Prior to ultracentrifugation, the translational processes in the cells are stopped using a chemical called cycloheximide, providing a snapshot of the translational state at the time of inhibition. Finally, RNA-Seq was performed on the isolated polysome-bound RNA, monosome-bound RNA, and total RNA to gain insights into both the transcriptional and translational landscape of the cells when miR-4728-3p was blocked.

We found that genes involved with steroid hormone biosynthesis were consistently upregulated when we blocked miR-4728-3p activity. Interestingly, most of the

differentially expressed genes in this pathway were related to estrogen synthesis, with aromatase (*CYP19A1*) showing the most significant upregulation in both the polysome fraction and total RNA. Aromatase is a critical enzyme responsible for converting testosterone into estradiol, the primary circulating estrogen hormone in humans. This conversion step is the rate-limiting step of estrogen synthesis, making aromatase a key target in breast cancer therapy. This observation was particularly interesting considering previous results, as we and others have demonstrated that miR-4728-3p can also regulate the levels of ESR1.

To investigate if the observed aromatase upregulation resulted in functional estrogen production, we designed an experiment utilizing conditioned media. We blocked miR-4728-3p activity in SK-BR-3 cells and then transferred the resulting conditioned medium to cells from the ER-positive cell line MCF7. This conditioned medium containing potentially elevated estrogen levels stimulated the proliferation of MCF7 cells, while control conditioned medium with normal miR-4728-3p activity had no such effect. Adding the aromatase inhibitor letrozole to the SK-BR-3 cells before collection of media abolished the proliferative effects on MCF7 cells, suggesting that the observed increase in MCF7 proliferation was indeed dependent on estrogen production. Finally, supplementing the conditioned medium with additional estrogen after letrozole treatment restored the proliferative effects on MCF7 cells, further confirming the role of estrogen in this process.

Overall, these results established an interesting link between the oncogenic ERBB2 and estrogen synthesis via the intronic miR-4728-3p. It is important to note, however, that these results are from a single breast cancer cell line, and using a single time point. Further studies are needed to fully explore the relationship between this miRNA and estrogen synthesis. An interesting experiment to complement our analysis could be to overexpress the miRNA in an ERBB2-negative cell line, and see if the effects mirror the results of the polysome fractionation.

Ethical considerations

The studies in this thesis utilized raw RNA-Seq and WGS data derived from patient tumors. Sequencing data derived from human beings is generally classified as being "sensitive", as it can be used to identify the patient that the sample was derived from. As such, data like this needs to be stored so that malicious third parties cannot access it. While the analysis of all sensitive data was performed on a secure GDPR-compliant High-Performance Computing cluster, ethical considerations regarding informed

consent and patient privacy remain important. These patients are entitled to privacy, and the data generated from them should be treated with respect.

Most of the sensitive data that we used in our studies comes from TCGA. The processed data in this cohort is publicly available, including gene expression and copy number variation matrices. Each patient and sample in the cohort also receives an encrypted ID, and the key to identify them is only accessible to a very few authorized personnel. Access to sensitive data, including raw RNA-Seq and WGS data, is strictly controlled through a project-by-project approval process. Researchers are granted access only for the stated purpose of their project and must delete the data upon project completion. Because their data is used for thousands of research projects, TCGA utilizes an "umbrella consent" approach. Patients enrolled in the cohort provide broad consent for their anonymized genetic data to be used in any cancer research project approved by TCGA. This approach streamlines the research process but raises ethical considerations regarding the level of specificity that the patients have in consenting to data use. Similar consent is provided by the patients enrolled in SCAN-B, the second cohort we utilized raw sequencing data from.

In addition to sequencing data derived from patient tumors, our studies also utilized data from established human cell lines. While data from these lines may not be inherently sensitive in the same way as patient data, the use of human cell lines raises distinct ethical considerations. Many of the widely used cell lines predate current informed consent standards. It is crucial to acknowledge the possibility that the initial tissue collection for the cell line may not have involved proper informed consent from the donor. The cell lines that we used are also commercially available, and it is important to consider the ethical implications of profiteering from human biological material.

Conclusions and future perspectives

The overall aims of this thesis were to study the role of gene fusions and miRNAs in cancer. To achieve this, we primarily used *in silico* analyses, analyzing sequencing data from large cancer patient cohorts or cell lines. Much of our analysis was performed in breast cancer, but in some of our gene fusion studies we expanded our efforts to include ten additional cancer types.

Our results indicate that many gene fusions that once would have been classified as silent passenger events may in fact be impacting the cell by deregulating miRNA expression. To tackle the problem of detecting false positive fusion transcripts, we created a tool to validate fusions found in RNA-Seq using matched WGS data. Using the results of this pipeline, we trained a machine learning classifier to predict if a fusion event is real or false. This classifier was then applied on samples that do not have matched WGS, and is a proof-of-concept that a machine learning-based filtering approach can improve fusion detection. Finally, we investigated the function of the *ERBB2*-encoded miR-4728-3p and established a link between it and estrogen synthesis.

Acknowledgements

First and foremost, I would like to extend my deepest gratitude to my supervisor, **Helena**. Your guidance and support have been invaluable throughout my PhD journey, and sharing an office with you for the past five years has been an intellectual privilege. I know that I will quickly begin to miss our daily conversations. Beyond your academic brilliance, your kindness and ability to offer insightful advice on both scientific and personal matters have been invaluable. Thank you for everything, Helena.

Next, I want to thank my former main supervisor, **Carlos Rovira**. Both his research and mentorship laid the foundation of this thesis. His passion for science and dedication to his students were truly inspiring. Even after his diagnosis with glioblastoma at the beginning of my PhD, he continued to share his time and expertise, fostering stimulating discussions on research and beyond. His absence is deeply felt, and his contributions to my development as a researcher will never be forgotten.

My co-supervisors **Rolf Søkilde, Johan Vallon-Christersson, Mattias Höglund**, and **Åke Borg**. While I may not have sought your guidance as frequently as I would have liked, your willingness to offer insightful advice and support throughout my PhD journey was invaluable.

A warm thank you to all the past and present members of the Functional Breast Cancer Genomics group for creating a supportive and stimulating research environment. To the current members, **Juliane, Mirjam, Robin**, and **Izabela**, a special thanks for the camaraderie and countless insightful discussions during our many lunches and coffee breaks. Your positive energy has made a big difference in my PhD. To the other PhD students at MV, especially **Lennart** and **Suze**, thanks for the many fun conversations as we try to motivate each other to get through these four years.

I would like to thank the organizers and other participants of the MedBioInfo, the national research school for medical bioinformatics. While it was unfortunate that the pandemic resulted in us taking the first few courses remotely, the later ones where we met in person were fantastic. It was a super stimulating environment, and I made a lot of great friends and memories. I would also like to thank everyone who organized and attended the RaukR workshop for advanced R programming in Gotland in the summer of 2023. It was an amazing two weeks with even more amazing people.

To everyone who helped me teach and R course: **Karin, Suze, Raquel**, and **Petter**. It was such a great and fulfilling experience!

A big thanks to the organizers of the MentLife program and to my mentor **Anders Carlsson**. We had a lot of great discussions in your office, and you helped me be less worried about the future.

I would also like to thank everyone who has helped me get into bioinformatics and given me useful advice and pointers, whether it was a programming tip, an R package, or advice on how to structure my code more efficiently.

To my friends in Lund – **Charley, Fred, Rikki, Stefano, Anna**, and everyone else. We have shared a lot of great moments and experiences over the past few years, here's to many more!

A special thanks goes out to the open-source bioinformatics and data science communities, especially to everyone working at Posit (formerly R Studio), the Tidyverse and Tidymodels. Your contribution to science cannot be overstated, and you made doing my PhD a much smoother experience.

To my family, I would like to thank you for always fostering my scientific curiosity.

Finally, I would like to thank my sambo **Deborah** for all you have done for me. For the past five years you have been my companion in the R course, workshops, conferences, seminars, courses, and life in general. Life in Sweden would not be the same without you.

References

1. Cancer Today. https://gco.iarc.who.int/today/. Accessed on February 29 2024.
2. Hajdu, S. I. A note from history: Landmarks in history of cancer, part 1. *Cancer* **117**, 1097–1102 (2011).
3. Haridy, Y. *et al.* Triassic Cancer—Osteosarcoma in a 240-Million-Year-Old Stem-Turtle. *JAMA Oncology* **5**, 425–426 (2019).
4. Hajdu, S. I. & Vadmal, M. A note from history: Landmarks in history of cancer, Part 6. *Cancer* **119**, 4058–4082 (2013).
5. Brown, J. S. *et al.* Updating the Definition of Cancer. *Molecular Cancer Research* **21**, 1142–1147 (2023).
6. Cancer Classification | SEER Training. https://training.seer.cancer.gov/disease/categories/classification.html. Accessed on February 29 2024.
7. Hanahan, D. & Weinberg, R. A. The Hallmarks of Cancer. *Cell* **100**, 57–70 (2000).
8. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).
9. Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discovery* **12**, 31–46 (2022).
10. Klymkowsky, M. W. & Savagner, P. Epithelial-Mesenchymal Transition: A Cancer Researcher's Conceptual Friend and Foe. *The American Journal of Pathology* **174**, 1588–1593 (2009).
11. Perona, R. Cell signalling: growth factors and tyrosine kinase receptors. *Clin Transl Oncol* **8**, 77–82 (2006).
12. Blasco, M. A. Telomeres and human disease: ageing, cancer and beyond. *Nat Rev Genet* **6**, 611–622 (2005).
13. Bielenberg, D. R. & Zetter, B. R. The Contribution of Angiogenesis to the Process of Metastasis. *The Cancer Journal* **21**, 267 (2015).
14. Vander Heiden, M. G., Cantley, L. C. & Thompson, C. B. Understanding the Warburg effect: the metabolic requirements of cell proliferation. *Science* **324**, 1029–1033 (2009).
15. Semenza, G. L. HIF-1: upstream and downstream of cancer metabolism. *Curr Opin Genet Dev* **20**, 51 (2010).

16. Wang, B., Kohli, J. & Demaria, M. Senescent Cells in Cancer Therapy: Friends or Foes? *Trends in Cancer* **6**, 838–857 (2020).
17. Yuan, S., Norgard, R. J. & Stanger, B. Z. Cellular Plasticity in Cancer. *Cancer Discovery* **9**, 837–851 (2019).
18. Aguilera, A. & García-Muse, T. Causes of Genome Instability. *Annual Review of Genetics* **47**, 1–32 (2013).
19. Lane, D. P. p53, guardian of the genome. *Nature* **358**, 15–16 (1992).
20. Kruse, J.-P. & Gu, W. Modes of p53 Regulation. *Cell* **137**, 609–622 (2009).
21. Knudson, A. G. Mutation and Cancer: Statistical Study of Retinoblastoma. *Proc Natl Acad Sci U S A* **68**, 820–823 (1971).
22. Inoue, K. & Fry, E. A. Haploinsufficient tumor suppressor genes. *Adv Med Biol* **118**, 83–122 (2017).
23. Baylin, S. B. & Jones, P. A. Epigenetic Determinants of Cancer. *Cold Spring Harb Perspect Biol* **8**, a019505 (2016).
24. Nam, A. S., Chaligne, R. & Landau, D. A. Integrating genetic and non-genetic determinants of cancer evolution by single-cell multi-omics. *Nat Rev Genet* **22**, 3–18 (2021).
25. Grivennikov, S. I., Greten, F. R. & Karin, M. Immunity, Inflammation, and Cancer. *Cell* **140**, 883–899 (2010).
26. Cullin, N., Azevedo Antunes, C., Straussman, R., Stein-Thoeringer, C. K. & Elinav, E. Microbiome and cancer. *Cancer Cell* **39**, 1317–1341 (2021).
27. Parhi, L. *et al.* Breast cancer colonization by Fusobacterium nucleatum accelerates tumor growth and metastatic progression. *Nat Commun* **11**, 3259 (2020).
28. Pushalkar, S. *et al.* The Pancreatic Cancer Microbiome Promotes Oncogenesis by Induction of Innate and Adaptive Immune Suppression. *Cancer Discov* **8**, 403–416 (2018).
29. Arnold, M. *et al.* Current and future burden of breast cancer: Global statistics for 2020 and 2040. *The Breast* **66**, 15–23 (2022).
30. *Nationellt Kvalitetsregister för Bröstcancer (NKBC) - Sammanfattning och vägledning till den interaktiva årsrapporten för 2022*. 16 (2023).
31. Risk factors for breast cancer. https://www.cancerresearchuk.org/about-cancer/breast-cancer/risks-causes/risk-factors. Accessed on February 29 2024.
32. Sun, Y.-S. *et al.* Risk Factors and Preventions of Breast Cancer. *Int J Biol Sci* **13**, 1387–1397 (2017).
33. van der Groep, P., van der Wall, E. & van Diest, P. J. Pathology of hereditary breast cancer. *Cell Oncol.* **34**, 71–88 (2011).
34. Brewer, H. R., Jones, M. E., Schoemaker, M. J., Ashworth, A. & Swerdlow, A. J. Family history and risk of breast cancer: an analysis accounting for family structure. *Breast Cancer Res Treat* **165**, 193–200 (2017).

35. Løberg, M., Lousdal, M. L., Bretthauer, M. & Kalager, M. Benefits and harms of mammography screening. *Breast Cancer Res* **17**, 63 (2015).

36. Hofvind, S. *et al.* False-positive results in mammographic screening for breast cancer in Europe: a literature review and survey of service screening programmes. *J Med Screen* **19 Suppl 1**, 57–66 (2012).

37. Viale, G. The current state of breast cancer classification. *Annals of Oncology* **23**, x207–x210 (2012).

38. *WHO Classification of Tumours of the Breast: Reflects the Views of a Working Group That Convened for a Consensus and Editorial Meeting at the International Agency for Research on Cancer (IARC), Lyon, September 1-3, 2011*. (Internat. Agency for Research on Cancer, Lyon, 2012).

39. Cserni, G. Histological type and typing of breast carcinomas and the WHO classification changes over time. *Pathologica* **112**, 25–41 (2020).

40. Chang, J. M. *et al.* Back to Basics: Traditional Nottingham Grade Mitotic Counts Alone are Significant in Predicting Survival in Invasive Breast Carcinoma. *Ann Surg Oncol* **22**, 509–515 (2015).

41. Rakha, E. A. *et al.* Breast cancer prognostic classification in the molecular era: the role of histological grade. *Breast Cancer Res* **12**, 207 (2010).

42. Cancer Staging - NCI. https://www.cancer.gov/about-cancer/diagnosis-staging/staging (2015). Accessed on February 29 2024.

43. Nolan, E., Lindeman, G. J. & Visvader, J. E. Deciphering breast cancer: from biology to the clinic. *Cell* **186**, 1708–1728 (2023).

44. Orrantia-Borunda, E., Anchondo-Nuñez, P., Acuña-Aguilar, L. E., Gómez-Valles, F. O. & Ramírez-Valdespino, C. A. Subtypes of Breast Cancer. in *Breast Cancer* (ed. Mayrovitz, H. N.) (Exon Publications, Brisbane (AU), 2022).

45. Fuentes, N. & Silveyra, P. Estrogen receptor signaling mechanisms. *Adv Protein Chem Struct Biol* **116**, 135–170 (2019).

46. Lange, C. A. & Yee, D. Progesterone and Breast Cancer. *Womens Health (Lond Engl)* **4**, 151–162 (2008).

47. Purdie, C. A. *et al.* Progesterone receptor expression is an independent prognostic variable in early breast cancer: a population-based study. *Br J Cancer* **110**, 565–572 (2014).

48. Gutierrez, C. & Schiff, R. HER 2: Biology, Detection, and Clinical Implications. *Arch Pathol Lab Med* **135**, 55–62 (2011).

49. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).

50. Parker, J. S. *et al.* Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *J Clin Oncol* **27**, 1160–1167 (2009).

51. Goldhirsch, A. *et al.* Strategies for subtypes—dealing with the diversity of breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. *Annals of Oncology* **22**, 1736–1747 (2011).

52. Hajdu, S. I. A note from history: Landmarks in history of cancer, part 3. *Cancer* **118**, 1155–1168 (2012).

53. Fisher, B. *et al.* Twenty-Year Follow-up of a Randomized Trial Comparing Total Mastectomy, Lumpectomy, and Lumpectomy plus Irradiation for the Treatment of Invasive Breast Cancer. *New England Journal of Medicine* **347**, 1233–1241 (2002).

54. Sweden: share of breast-conserving surgeries 2015-2021. *Statista* https://www.statista.com/statistics/972132/share-of-breast-conserving-surgeries-in-sweden/. Accessed on February 29 2024.

55. Christakis, P. The Birth of Chemotherapy at Yale. *Yale J Biol Med* **84**, 169–172 (2011).

56. Zeman, E. M., Schreiber, E. C. & Tepper, J. E. 27 - Basics of Radiation Therapy. in *Abeloff's Clinical Oncology (Sixth Edition)* (eds. Niederhuber, J. E., Armitage, J. O., Kastan, M. B., Doroshow, J. H. & Tepper, J. E.) 431-460.e3 (Elsevier, Philadelphia, 2020). doi:10.1016/B978-0-323-47674-4.00027-X.

57. Burguin, A., Diorio, C. & Durocher, F. Breast Cancer Treatments: Updates and New Challenges. *Journal of Personalized Medicine* **11**, 808 (2021).

58. Nationellt vårdprogram bröstcancer - RCC Kunskapsbanken. https://kunskapsbanken.cancercentrum.se/diagnoser/brostcancer/vardprogram/. Accessed on February 29 2024.

59. Makin, G. Principles of chemotherapy. *Paediatrics and Child Health* **28**, 183–188 (2018).

60. Patel, H. K. & Bihani, T. Selective estrogen receptor modulators (SERMs) and selective estrogen receptor degraders (SERDs) in cancer treatment. *Pharmacology & Therapeutics* **186**, 1–24 (2018).

61. Chumsri, S., Howes, T., Bao, T., Sabnis, G. & Brodie, A. Aromatase, aromatase inhibitors, and breast cancer. *The Journal of Steroid Biochemistry and Molecular Biology* **125**, 13–22 (2011).

62. Carter, P. *et al.* Humanization of an anti-p185HER2 antibody for human cancer therapy. *Proceedings of the National Academy of Sciences* **89**, 4285–4289 (1992).

63. Hudis, C. A. Trastuzumab — Mechanism of Action and Use in Clinical Practice. *New England Journal of Medicine* **357**, 39–51 (2007).

64. Lambert, J. M. & Chari, R. V. J. Ado-trastuzumab Emtansine (T-DM1): An Antibody–Drug Conjugate (ADC) for HER2-Positive Breast Cancer. *J. Med. Chem.* **57**, 6949–6964 (2014).

65. Nahta, R. & Esteva, F. J. Trastuzumab: triumphs and tribulations. *Oncogene* **26**, 3637–3643 (2007).

66. Schlam, I. & Swain, S. M. HER2-positive breast cancer and tyrosine kinase inhibitors: the time is now. *npj Breast Cancer* **7**, 1–12 (2021).

67. Kuchenbaecker, K. B. *et al.* Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers. *JAMA* **317**, 2402–2416 (2017).

68. Cortesi, L., Rugo, H. S. & Jackisch, C. An Overview of PARP Inhibitors for the Treatment of Breast Cancer. *Target Oncol* **16**, 255–282 (2021).

69. Waks, A. G. & Winer, E. P. Breast Cancer Treatment: A Review. *JAMA* **321**, 288–300 (2019).

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

71. Wightman, B., Ha, I. & Ruvkun, G. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell* **75**, 855–862 (1993).

72. Ambros, V. A hierarchy of regulatory genes controls a larva-to-adult developmental switch in C. elegans. *Cell* **57**, 49–57 (1989).

73. Reinhart, B. J. *et al.* The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature* **403**, 901–906 (2000).

74. Lagos-Quintana, M., Rauhut, R., Lendeckel, W. & Tuschl, T. Identification of Novel Genes Coding for Small Expressed RNAs. *Science* **294**, 853–858 (2001).

75. Lau, N. C., Lim, L. P., Weinstein, E. G. & Bartel, D. P. An Abundant Class of Tiny RNAs with Probable Regulatory Roles in Caenorhabditis elegans. *Science* **294**, 858–862 (2001).

76. Lee, R. C. & Ambros, V. An Extensive Class of Small RNAs in Caenorhabditis elegans. *Science* **294**, 862–864 (2001).

77. Lagos-Quintana, M. *et al.* Identification of Tissue-Specific MicroRNAs from Mouse. *Current Biology* **12**, 735–739 (2002).

78. Wienholds, E. *et al.* MicroRNA Expression in Zebrafish Embryonic Development. *Science* **309**, 310–311 (2005).

79. Zhang, B., Wang, Q. & Pan, X. MicroRNAs and their regulatory roles in animals and plants. *Journal of Cellular Physiology* **210**, 279–289 (2007).

80. Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. miRBase: from microRNA sequences to function. *Nucleic Acids Res* **47**, D155–D162 (2019).

81. Lee, Y. *et al.* The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425**, 415–419 (2003).

82. Denli, A. M., Tops, B. B. J., Plasterk, R. H. A., Ketting, R. F. & Hannon, G. J. Processing of primary microRNAs by the Microprocessor complex. *Nature* **432**, 231–235 (2004).

83. Gregory, R. I. *et al.* The Microprocessor complex mediates the genesis of microRNAs. *Nature* **432**, 235–240 (2004).

84. Han, J. *et al.* Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell* **125**, 887–901 (2006).

85. Nguyen, T. A. *et al.* Functional Anatomy of the Human Microprocessor. *Cell* **161**, 1374–1387 (2015).

86. Roden, C. *et al.* Novel determinants of mammalian primary microRNA processing revealed by systematic evaluation of hairpin-containing transcripts and human genetic variation. *Genome Res.* **27**, 374–384 (2017).

87. Kuehbacher, A., Urbich, C., Zeiher, A. M. & Dimmeler, S. Role of Dicer and Drosha for Endothelial MicroRNA Expression and Angiogenesis. *Circulation Research* **101**, 59–68 (2007).

88. Vergani-Junior, C. A., Tonon-da-Silva, G., Inan, M. D. & Mori, M. A. DICER: structure, function, and regulation. *Biophys Rev* **13**, 1081–1090 (2021).

89. Naruse, K., Matsuura-Suzuki, E., Watanabe, M., Iwasaki, S. & Tomari, Y. In vitro reconstitution of chaperone-mediated human RISC assembly. *RNA* **24**, 6–11 (2018).

90. Kwak, P. B. & Tomari, Y. The N domain of Argonaute drives duplex unwinding during RISC assembly. *Nat Struct Mol Biol* **19**, 145–151 (2012).

91. Meijer, H. A., Smith, E. M. & Bushell, M. Regulation of miRNA strand selection: follow the leader? *Biochem Soc Trans* **42**, 1135–1140 (2014).

92. Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell* **136**, 215–233 (2009).

93. Chandradoss, S. D., Schirle, N. T., Szczepaniak, M., MacRae, I. J. & Joo, C. A Dynamic Search Process Underlies MicroRNA Targeting. *Cell* **162**, 96–107 (2015).

94. Iwakawa, H. & Tomari, Y. Life of RISC: Formation, action, and degradation of RNA-induced silencing complex. *Molecular Cell* **82**, 30–43 (2022).

95. Schirle, N. T., Sheu-Gruttadauria, J. & MacRae, I. J. Structural basis for microRNA targeting. *Science* **346**, 608–613 (2014).

96. Mathonnet, G. *et al.* MicroRNA Inhibition of Translation Initiation in Vitro by Targeting the Cap-Binding Complex eIF4F. *Science* **317**, 1764–1767 (2007).

97. Eichhorn, S. W. *et al.* mRNA Destabilization Is the Dominant Effect of Mammalian MicroRNAs by the Time Substantial Repression Ensues. *Molecular Cell* **56**, 104–115 (2014).

98. Cooke, A., Prigge, A. & Wickens, M. Translational Repression by Deadenylases*. *Journal of Biological Chemistry* **285**, 28506–28513 (2010).

99. Iwakawa, H.-O. & Tomari, Y. The Functions of MicroRNAs: mRNA Decay and Translational Repression. Trends Cell Biol 25, 651–665 (2015).

100. Behm-Ansmant, I. *et al.* mRNA degradation by miRNAs and GW182 requires both CCR4:NOT deadenylase and DCP1:DCP2 decapping complexes. *Genes Dev.* **20**, 1885–1898 (2006).

101. Braun, J. E., Huntzinger, E., Fauser, M. & Izaurralde, E. GW182 proteins directly recruit cytoplasmic deadenylase complexes to miRNA targets. *Mol Cell* **44**, 120–133 (2011).

102. Fabian, M. R. *et al.* Mammalian miRNA RISC recruits CAF1 and PABP to affect PABP-dependent deadenylation. *Mol Cell* **35**, 868–880 (2009).

103. Jinek, M., Fabian, M. R., Coyle, S. M., Sonenberg, N. & Doudna, J. A. Structural insights into the human GW182-PABC interaction in microRNA-mediated deadenylation. *Nat Struct Mol Biol* **17**, 238–240 (2010).

104. Nishihara, T., Zekri, L., Braun, J. E. & Izaurralde, E. miRISC recruits decapping factors to miRNA targets to enhance their degradation. *Nucleic Acids Res* **41**, 8692–8705 (2013).

105. Stavast, C. J. & Erkeland, S. J. The Non-Canonical Aspects of MicroRNAs: Many Roads to Gene Regulation. *Cells* **8**, 1465 (2019).

106. Ruby, J. G., Jan, C. H. & Bartel, D. P. Intronic microRNA precursors that bypass Drosha processing. *Nature* **448**, 83–86 (2007).

107. Berezikov, E., Chung, W.-J., Willis, J., Cuppen, E. & Lai, E. C. Mammalian mirtron genes. *Mol Cell* **28**, 328–336 (2007).

108. Cheloufi, S., Dos Santos, C. O., Chong, M. M. W. & Hannon, G. J. A dicer-independent miRNA biogenesis pathway that requires Ago catalysis. *Nature* **465**, 584–589 (2010).

109. Cifuentes, D. *et al.* A novel miRNA processing pathway independent of Dicer requires Argonaute2 catalytic activity. *Science* **328**, 1694–1698 (2010).

110. Altuvia, Y. *et al.* Clustering and conservation patterns of human microRNAs. *Nucleic Acids Res* **33**, 2697–2706 (2005).

111. Bartel, D. P. Metazoan MicroRNAs. *Cell* **173**, 20–51 (2018).

112. Steiman-Shimony, A., Shtrikman, O. & Margalit, H. Assessing the functional association of intronic miRNAs with their host genes. *RNA* **24**, 991–1004 (2018).

113. Boivin, V., Deschamps-Francoeur, G. & Scott, M. S. Protein coding genes as hosts for noncoding RNA expression. *Seminars in Cell & Developmental Biology* **75**, 3–12 (2018).

114. Hinske, L. C. G., Galante, P. A., Kuo, W. P. & Ohno-Machado, L. A potential role for intragenic miRNAs on their hosts' interactome. *BMC Genomics* **11**, 533 (2010).

115. Lutter, D., Marr, C., Krumsiek, J., Lang, E. W. & Theis, F. J. Intronic microRNAs support their host genes by mediating synergistic and antagonistic regulatory effects. *BMC Genomics* **11**, 224 (2010).

116. Friedman, R. C., Farh, K. K.-H., Burge, C. B. & Bartel, D. P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* **19**, 92–105 (2009).

117. Ahmed, K. *et al.* Loss of microRNA-7a2 induces hypogonadotropic hypogonadism and infertility. *J Clin Invest* **127**, 1061–1074 (2017).

118. Papadopoulou, A. S. *et al.* The thymic epithelial microRNA network elevates the threshold for infection-associated thymic involution via miR-29a mediated suppression of the IFN-α receptor. *Nat Immunol* **13**, 181–187 (2012).

119. Moffett, H. F. *et al.* The microRNA miR-31 inhibits CD8+ T cell function in chronic viral infection. *Nat Immunol* **18**, 791–799 (2017).

120. Tattikota, S. G. *et al.* Argonaute2 Mediates Compensatory Expansion of the Pancreatic β Cell. *Cell Metabolism* **19**, 122–134 (2014).

121. Volinia, S. *et al.* A microRNA expression signature of human solid tumors defines cancer gene targets. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 2257–2261 (2006).

122. Woods, K., Thomson, J. M. & Hammond, S. M. Direct Regulation of an Oncogenic Micro-RNA Cluster by E2F Transcription Factors *. *Journal of Biological Chemistry* **282**, 2130–2134 (2007).

123. Zhang, C.-Z. *et al.* MiR-221 and miR-222 target PUMA to induce cell survival in glioblastoma. *Mol Cancer* **9**, 229 (2010).

124. Fasanaro, P. *et al.* MicroRNA-210 modulates endothelial cell response to hypoxia and inhibits the receptor tyrosine kinase ligand Ephrin-A3. *J Biol Chem* **283**, 15878–15883 (2008).

125. Kong, W. *et al.* MicroRNA-155 is regulated by the transforming growth factor beta/Smad pathway and contributes to epithelial cell plasticity by targeting RhoA. *Mol Cell Biol* **28**, 6773–6784 (2008).

126. Peng, Y. & Croce, C. M. The role of MicroRNAs in human cancer. *Sig Transduct Target Ther* **1**, 1–9 (2016).

127. Selcuklu, S. D., Donoghue, M. T. A. & Spillane, C. miR-21 as a key regulator of oncogenic processes. *Biochemical Society Transactions* **37**, 918–925 (2009).

128. Pfeffer, S. R., Yang, C. H. & Pfeffer, L. M. The Role of miR-21 in Cancer. *Drug Development Research* **76**, 270–277 (2015).

129. Sekar, D., Venugopal, B., Sekar, P. & Ramalingam, K. Role of microRNA 21 in diabetes and associated/related diseases. *Gene* **582**, 14–18 (2016).

130. Zhang, T., Yang, Z., Kusumanchi, P., Han, S. & Liangpunsakul, S. Critical Role of microRNA-21 in the Pathogenesis of Liver Diseases. *Front Med (Lausanne)* **7**, 7 (2020).

131. Surina, S. *et al.* miR-21 in Human Cardiomyopathies. *Front Cardiovasc Med* **8**, 767064 (2021).

132. Sand, M. *et al.* Expression levels of the microRNA maturing microprocessor complex component DGCR8 and the RNA-induced silencing complex (RISC) components argonaute-1, argonaute-2, PACT, TARBP1, and TARBP2 in epithelial skin cancer. *Mol Carcinog* **51**, 916–922 (2012).

133. Papachristou, D. J. *et al.* Expression of the ribonucleases Drosha, Dicer, and Ago2 in colorectal carcinomas. *Virchows Arch* **459**, 431–440 (2011).

134. Lu, J. *et al.* MicroRNA expression profiles classify human cancers. *Nature* **435**, 834–838 (2005).

135. Lee, B. *et al.* Mapping genetic variability in mature miRNAs and miRNA binding sites in prostate cancer. *J Hum Genet* **66**, 1127–1137 (2021).

136. Dvinge, H. *et al.* The shaping and functional consequences of the microRNA landscape in breast cancer. *Nature* **497**, 378–382 (2013).

137. Liu, A. *et al.* MicroRNA expression profiling outperforms mRNA expression profiling in formalin-fixed paraffin-embedded tissues. *Int J Clin Exp Pathol* **2**, 519–527 (2009).

138. Cui, M. *et al.* Circulating MicroRNAs in Cancer: Potential and Challenge. *Front Genet* **10**, 626 (2019).

139. Cheng, G. Circulating miRNAs: Roles in cancer diagnosis, prognosis and therapy. *Advanced Drug Delivery Reviews* **81**, 75–93 (2015).

140. Zou, R. *et al.* Development and validation of a circulating microRNA panel for the early detection of breast cancer. *Br J Cancer* **126**, 472–481 (2022).

141. Lin, X.-J. *et al.* A serum microRNA classifier for early detection of hepatocellular carcinoma: a multicentre, retrospective, longitudinal biomarker identification study with a nested case-control study. *The Lancet Oncology* **16**, 804–815 (2015).

142. So, J. B. Y. *et al.* Development and validation of a serum microRNA biomarker panel for detecting gastric cancer in a high-risk population. *Gut* **70**, 829–837 (2021).

143. Zhang, C. & Zhang, B. RNA therapeutics: updates and future potential. *Sci. China Life Sci.* **66**, 12–30 (2023).

144. Rupaimoole, R. & Slack, F. J. MicroRNA therapeutics: towards a new era for the management of cancer and other diseases. *Nat Rev Drug Discov* **16**, 203–222 (2017).

145. Asbestos Diseases Research Foundation. *MesomiR 1: A Phase I Study of Intravenously Administered Epidermal Growth Factor Receptor -Targeted, EnGeneIC Delivery Vehicle (EDV)-Packaged, miR-16 Mimic (TargomiRs) for Patients With Malignant Pleural Mesothelioma (MPM) and Advanced Non-Small Cell Lung Cancer (NSCLC) Failing on Std Therapy*. https://clinicaltrials.gov/study/NCT02369198 (2017).

146. miRagen Therapeutics, Inc. *SOLAR: A Phase 2, Randomized, Open-Label, Parallel-Group, Active Comparator, Multi-Center Study to Investigate the Efficacy and Safety of Cobomarsen (MRG-106) in Subjects With Cutaneous T-Cell Lymphoma (CTCL), Mycosis Fungoides (MF) Subtype*. https://clinicaltrials.gov/study/NCT03713320 (2022).

147. Persson, H. *et al.* Identification of New MicroRNAs in Paired Normal and Tumor Breast Tissue Suggests a Dual Role for the ERBB2/Her2 Gene. *Cancer Research* **71**, 78–86 (2011).

148. Newie, I. *et al.* HER2-encoded mir-4728 forms a receptor-independent circuit with miR-21-5p through the non-canonical poly(A) polymerase PAPD5. *Sci Rep* **6**, 35664 (2016).

149. Gong, C. *et al.* Up-regulation of miR-21 Mediates Resistance to Trastuzumab Therapy for Breast Cancer *. *Journal of Biological Chemistry* **286**, 19127–19137 (2011).

150. Newie, I. *et al.* The HER2-Encoded miR-4728-3p Regulates ESR1 through a Non-Canonical Internal Seed Interaction. *PLoS ONE* **9**, e97200 (2014).

151. Floros, K. V. *et al.* Coamplification of miR-4728 protects HER2-amplified breast cancers from targeted therapy. *Proc Natl Acad Sci U S A* **115**, E2594–E2603 (2018).

152. Rui, T. *et al.* Mir-4728 is a Valuable Biomarker for Diagnostic and Prognostic Assessment of HER2-Positive Breast Cancer. *Front Mol Biosci* **9**, 818493 (2022).

153. Pinhel, I. *et al.* ER and HER2 expression are positively correlated in HER2 non-overexpressing breast cancer. *Breast Cancer Res* **14**, R46 (2012).

154. Guo, S. & Sonenshein, G. E. Forkhead Box Transcription Factor FOXO3a Regulates Estrogen Receptor Alpha Expression and Is Repressed by the Her-2/neu/Phosphatidylinositol 3-Kinase/Akt Signaling Pathway. *Mol Cell Biol* **24**, 8681–8690 (2004).

155. Wright, C. *et al.* Relationship between c-erbB-2 protein product expression and response to endocrine therapy in advanced breast cancer. *Br J Cancer* **65**, 118–121 (1992).

156. Pekow, J. *et al.* miR-4728-3p Functions as a Tumor Suppressor in Ulcerative Colitis-associated Colorectal Neoplasia Through Regulation of Focal Adhesion Signaling. *Inflammatory Bowel Diseases* **23**, 1328–1337 (2017).

157. Liu, Z., Zhang, J., Gao, J. & Li, Y. MicroRNA-4728 mediated regulation of MAPK oncogenic signaling in papillary thyroid carcinoma. *Saudi Journal of Biological Sciences* **25**, 986–990 (2018).

158. Wang, W. *et al.* MicroRNA-4728 serves as a suppressor and antagonist of oncogenic MAPK in Burkitt lymphoma. *Saudi Journal of Biological Sciences* **25**, 982–985 (2018).

159. Nowell, P., Hungerford, D. & Nowell, P. A minute chromosome in human chronic granulocytic leukemia. *Science* (1960).

160. Rowley, J. D. A New Consistent Chromosomal Abnormality in Chronic Myelogenous Leukaemia identified by Quinacrine Fluorescence and Giemsa Staining. *Nature* **243**, 290–293 (1973).

161. Shtivelman, E., Lifshitz, B., Gale, R. P. & Canaani, E. Fused transcript of abl and bcr genes in chronic myelogenous leukaemia. *Nature* **315**, 550–554 (1985).

162. Hecht, J. L. & Aster, J. C. Molecular Biology of Burkitt's Lymphoma. *JCO* **18**, 3707–3721 (2000).

163. Yoshihara, K. *et al.* The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene* **34**, 4845–4854 (2015).

164. Li, H., Wang, J., Ma, X. & Sklar, J. Gene fusions and RNA trans-splicing in normal and neoplastic human cells. *Cell Cycle* **8**, 218–222 (2009).

165. Grosso, A. R. *et al.* Pervasive transcription read-through promotes aberrant expression of oncogenes and RNA chimeras in renal carcinoma. *eLife* **4**, e09214 (2015).

166. Hu, X. *et al.* TumorFusions: an integrative resource for cancer-associated transcript fusions. *Nucleic Acids Res* **46**, D1144–D1149 (2018).

167. Hogenbirk, M. A. *et al.* Defining chromosomal translocation risks in cancer. *Proc Natl Acad Sci U S A* **113**, E3649-3656 (2016).

168. Burrow, A. A., Williams, L. E., Pierce, L. C. & Wang, Y.-H. Over half of breakpoints in gene pairs involved in cancer-specific recurrent translocations are mapped to human chromosomal fragile sites. *BMC Genomics* **10**, 59 (2009).

169. Faderl, S. *et al.* The Biology of Chronic Myeloid Leukemia. *New England Journal of Medicine* **341**, 164–172 (1999).

170. Johansson, B. *et al.* Most gene fusions in cancer are stochastic events. *Genes, Chromosomes and Cancer* **58**, 607–611 (2019).

171. Mertens, F., Johansson, B., Fioretos, T. & Mitelman, F. The emerging complexity of gene fusions in cancer. *Nat Rev Cancer* **15**, 371–381 (2015).

172. Kris, M. G. *et al.* Using Multiplexed Assays of Oncogenic Drivers in Lung Cancers to Select Targeted Drugs. *JAMA* **311**, 1998–2006 (2014).

173. Liquori, A. *et al.* Acute Promyelocytic Leukemia: A Constellation of Molecular Events around a Single PML-RARA Fusion Gene. *Cancers* **12**, 624 (2020).

174. Yu, J. S. E., Colborne, S., Hughes, C. S., Morin, G. B. & Nielsen, T. O. The FUS-DDIT3 Interactome in Myxoid Liposarcoma. *Neoplasia* **21**, 740–751 (2019).

175. Tomlins, S. A. *et al.* Role of the TMPRSS2-ERG gene fusion in prostate cancer. *Neoplasia* **10**, 177–188 (2008).

176. Veeraraghavan, J., Ma, J., Hu, Y. & Wang, X.-S. Recurrent and pathological gene fusions in breast cancer: current advances in genomic discovery and clinical implications. *Breast Cancer Res Treat* **158**, 219–232 (2016).

177. Schneider, J. L., Lin, J. J. & Shaw, A. T. ALK-positive lung cancer: a moving target. *Nat Cancer* **4**, 330–343 (2023).

178. Tsuda, Y. *et al.* The clinical heterogeneity of round cell sarcomas with EWSR1/FUS gene fusions: Impact of gene fusion type on clinical features and outcome. *Genes, Chromosomes and Cancer* **59**, 525–534 (2020).

179. Santoro, M., Moccia, M., Federico, G. & Carlomagno, F. RET Gene Fusions in Malignancies of the Thyroid and Other Tissues. *Genes* **11**, 424 (2020).

180. Subbiah, V. *et al.* Tumour-agnostic efficacy and safety of selpercatinib in patients with RET fusion-positive solid tumours other than lung or thyroid tumours (LIBRETTO-001): a phase 1/2, open-label, basket trial. *The Lancet Oncology* **23**, 1261–1273 (2022).

181. Chen, Y. & Chi, P. Basket trial of TRK inhibitors demonstrates efficacy in TRK fusion-positive cancers. *J Hematol Oncol* **11**, 78 (2018).

182. Medves, S. & Demoulin, J.-B. Tyrosine kinase gene fusions in cancer: translating mechanisms into targeted therapies. *Journal of Cellular and Molecular Medicine* **16**, 237–248 (2012).

183. Enlund, F. *et al.* Altered Notch signaling resulting from expression of a *WAMTP1-MAML2* gene fusion in mucoepidermoid carcinomas and benign Warthin's tumors. *Experimental Cell Research* **292**, 21–28 (2004).

184. Mark, J., Dahlenfors, R., Ekedahl, C. & Stenman, G. The mixed salivary gland tumor — A normally benign human neoplasm frequently showing specific chromosomal abnormalities. *Cancer Genetics and Cytogenetics* **2**, 231–241 (1980).

185. Oliver, G. R. *et al.* A tailored approach to fusion transcript identification increases diagnosis of rare inherited disease. *PLoS One* **14**, e0223337 (2019).

186. Haley, L. *et al.* Diagnostic Utility of Gene Fusion Panel to Detect Gene Fusions in Fresh and Formalin-Fixed, Paraffin-Embedded Cancer Specimens. *The Journal of Molecular Diagnostics* **23**, 1343–1358 (2021).

187. Engvall, M. *et al.* Detection of leukemia gene fusions by targeted RNA-sequencing in routine diagnostics. *BMC Med Genomics* **13**, 106 (2020).

188. Shugay, M., Ortiz de Mendíbil, I., Vizmanos, J. L. & Novo, F. J. Oncofuse: a computational framework for the prediction of the oncogenic potential of gene fusions. *Bioinformatics* **29**, 2539–2546 (2013).

189. Gaonkar, K. S. *et al.* annoFuse: an R Package to annotate, prioritize, and interactively explore putative oncogenic RNA fusions. *BMC Bioinformatics* **21**, 577 (2020).

190. Persson, H. *et al.* Frequent miRNA-convergent fusion gene events in breast cancer. *Nat Commun* **8**, 788 (2017).

191. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).

192. Asmann, Y. W. *et al.* A novel bioinformatics pipeline for identification and characterization of fusion transcripts in breast cancer and normal cell lines. *Nucleic Acids Res* **39**, e100 (2011).

193. Edgren, H. *et al.* Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome Biol* **12**, R6 (2011).

194. Kangaspeska, S. *et al.* Reanalysis of RNA-sequencing data reveals several additional fusion genes with multiple isoforms. *PLoS One* **7**, e48745 (2012).

195. Maher, C. A. *et al.* Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc Natl Acad Sci U S A* **106**, 12353–12358 (2009).

196. Uhrig, S. *et al.* Accurate and efficient detection of gene fusions from RNA sequencing data. *Genome Res.* **31**, 448–460 (2021).

197. Haas, B. J. *et al.* STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. Preprint at https://doi.org/10.1101/120295 (2017).

198. Nicorici, D. *et al.* FusionCatcher – a tool for finding somatic fusion genes in paired-end RNA-sequencing data. 011650 Preprint at https://doi.org/10.1101/011650 (2014).

199. Haas, B. J. *et al.* Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biology* **20**, 213 (2019).

200. Carrara, M. *et al.* State-of-the-Art Fusion-Finder Algorithms Sensitivity and Specificity. *BioMed Research International* **2013**, e340620 (2013).

201. Hafstað, V., Häkkinen, J. & Persson, H. Fast and sensitive validation of fusion transcripts in whole-genome sequencing data. *BMC Bioinformatics* **24**, 359 (2023).

202. Ke, G. *et al.* LightGBM: A Highly Efficient Gradient Boosting Decision Tree. in *Advances in Neural Information Processing Systems* vol. 30 (Curran Associates, Inc., 2017).

203. The Gene Ontology Consortium *et al.* The Gene Ontology knowledgebase in 2023. *Genetics* **224**, iyad031 (2023).

204. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**, 27–30 (2000).

205. Milacic, M. *et al.* The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Research* **52**, D672–D678 (2024).

206. 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* **102**, 15545–15550 (2005).

207. Liu, B., Shyr, Y., Cai, J. & Liu, Q. Interplay between miRNAs and host genes and their role in cancer. *Briefings in Functional Genomics* **18**, 255–266 (2019).

208. Fafard-Couture, É., Jacques, P.-É. & Scott, M. S. Motif conservation, stability, and host gene expression are the main drivers of snoRNA expression across vertebrates. *Genome Res.* **33**, 525–540 (2023).

209. Morlando, M. *et al.* Primary microRNA transcripts are processed co-transcriptionally. *Nat Struct Mol Biol* **15**, 902–909 (2008).

210. Persson, H. *et al.* Analysis of fusion transcripts indicates widespread deregulation of snoRNAs and their host genes in breast cancer. *International Journal of Cancer* **146**, 3343–3353 (2020).

211. Puig, R. R., Boddie, P., Khan, A., Castro-Mondragon, J. A. & Mathelier, A. UniBind: maps of high-confidence direct TF-DNA interactions across nine species. *BMC Genomics* **22**, 482 (2021).