Biomarker Discovery In Relapsed Acute Myeloid Leukaemia And The Impact Of CD109

Jenny Ashforth

A thesis submitted in partial fulfilment of the requirements of Nottingham Trent University for the degree of Doctor of Philosophy

September 2021

Copyright statement

Acknowledgements

I would like to thank my supervisory team; Prof. Sergio Rutella, Prof. Graham Pockley, Prof. Graham Ball and Dr. Sarah Warren, for giving me the opportunity to work on this project. Your support and guidance have been invaluable. I would also like to thank Dr. Jayakumar Vadakekolathu for all your help and guidance, you have really helped me grow as scientist. Of course, I would like to thank NanoString for funding this project in partnership with Nottingham Trent University, the time I spent in Seattle was insightful and inspiring.

I would like to thank everyone in the John van Geest cancer research centre, staff and students, old and new, for creating a friendly and supportive environment. With such a diverse range of expertise there is always someone to ask and guidance to be given. I would especially like to thank Anne Schneider and Stephen Reeder for keeping us going by providing technical support and a caring ear.

Of course, I would be nowhere without my family who are always there to talk to when I needed it and proved their unwavering support. I would also like to thank my friends Melissa, Luisa, Tania and Harish for always being around to listen and advise, you have made the past four years fly by. Additionally, I would like to include special mention for the entire biosciences PhD community, you have been the driving force for change and have really pulled together though the difficulty of the pandemic.

TABLE OF FIGURES

LIST OF TABLES

ABBREVIATIONS

A	Adenine
AML	Acute Myeloid Leukaemia
ANN	Artificial neural networks
APL	Acute Promyelocytic Leukaemia
AUC	Area Under the Curve
BLR	Binary Logistic Regression
C	Cytosine
DNA	Deoxyribonucleic acid
ECM	Extracellular matrix
ELN	European LeukaemiaNet
FAB	French-American-British
FDA	Food and Drug Administration
FISH	Fluorescence in situ hybridization
G	Guanine
GO	Gemtuzumab Ozogamicin
HSC	Haematopoietic stem cells
lncRNA	Long Non-coding RNA
LB	Lauria Broth
MRD	Minimal Residual Disease
NICE	National Institute for Health and Care Excellence
RNA	Ribonucleic acid
ROC	Receiver Operator Characteristic
T	Thymine
TGFβ1	Transforming Growth Factor Beta 1
TGFBR1	Transforming Growth Factor Receptor Beta 1
TK	Tyrosine kinase
WHO	World Health Organisation

ABSTRACT

Acute myeloid leukaemia (AML) is a blood cancer which has an overall survival rate of around 30%. As a patient is treated for AML, about 50% will achieve complete remission, of those 50% will relapse within three years, often this leads to death due to a developed treatment resistance of the cancerous cells. To facilitate better clinical outcomes in AML a prognostic score was developed to predict which patients are at a higher risk of relapse and those that respond well to current treatment options.

With the aim of developing a robust workflow of biomarker discovery, several bioinformatics approaches were implemented with the objective of identifying novel biomarkers in AML. Gene expression datasets of AML patients were used to trial a variety of workflows to discover a suitable relapse prognostic score. The workflows trialled involved both machine learning and statistical approaches of biomarker discovery. Overall a panel of 9 biomarkers were discovered and used to predict relapse in AML.

The gene CD109 was identified in the process of biomarker discovery and was able to independently predict relapse in AML, where high expression of CD109 was associated with relapse in AML and had previously been reported in other cancers as associated with progression and treatment resistance. The CD109 protein is a co-receptor for TGFBR1, and facilitates its internalisation and degradation, thus disrupting the TGFβ1 signalling pathway.

As the CD109 gene was highly associated with relapse, this study aimed to identify its functional relevance in AML cell lines using shRNA mediated gene silencing. The cells with a reduced expression of CD109 were treated with a combination of cytarabine and TGFβ1 to determine the response with altered gene expression. Some gene expression changes were observed depending on the treatment condition, indicating different pathways are activated in response to the treatment.

The prognostic score created with this study has the potential to predict relapse before it happens, allowing patients to be monitored more closely and empowering clinicians to use alternative or targeted treatments. By using this prognostic score, the relapse rate in AML could be reduced, and patients who do relapse can be detected sooner. This score has the potential to improve survival rate in AML through identification of high-risk patients and increased monitoring. The workflows developed to discover biomarkers can be applied to a magnitude of settings allowing clinical tools to be developed and the identification of key genes that are indicative of a condition.

Chapter 1 - Introduction

1.1. Cancer

1.1.1. Causes of cancer

1.1.1.1. Carcinogens and mutagens

Carcinogens are a substance capable of causing cancer, and mutagens are agents which can cause genetic mutations. Some mutagens cannot be avoided but risks can be mitigated through modified behaviour. An example of an unavoidable mutagen is UV exposure, the leading cause of skin cancer[1], by reducing exposure to UV through use of sun-cream, covering up and staying in the shade a person's risk of UV related cancers will be reduced. Environmental pollutants, often caused by combustion engines, contain known mutagens and carcinogens[2,3] which can be difficult to avoid in urban areas apart from changing routes to areas with lower pollution levels, avoiding main roads when exercising, and encouraging more people to switch to less polluting modes of transport. Certain carcinogens and mutagens, like X-rays in a medical setting, are a necessary exposure where the need outweighs the risk. In these instances, the only option is to modify behaviour to reduce the impact of environmental factors which can cause cancer.

1.1.1.2. Lifestyle and cancer

Preventable cancers account for around 40% of all diagnosis, of which, the largest lifestyle cause is smoking, accounting for 15% of preventable cancers[4]. Smoking is directly related to lung cancer, but also increases the risk of other cancer including mouth and bladder cancer[5] as tobacco smoke contains a diverse range of known carcinogens[6]. The second leading preventable cause of cancer is weight, it is estimated 60% of people in the UK[7] are overweight or obese, which increases the risk a person may develop cancers including oesophageal, stomach, bowel, liver, breast and kidney to name a few[7]. Obesity can lead to cancer through increased inflammation and the accumulation of growth and sex hormones[8]. By maintaining a healthy active lifestyle with a balanced diet can reduce the risk of cancers.

1.1.1.3. Molecular changes

Transcription and translation are highly regulated processes and are essential for cells to function and proliferate appropriately. Mutations in the cell genome can cause changes in gene and protein expression which in turn can cause dysregulation of cellular processes. Mutations in the genome can be large and affect the chromosomes or small and just affect one or more base pairs. Mutations can be due to DNA replication error or exposure to mutagens.

DNA is made up of 4 different nucleobases adenine (A), cytosine (C), guanine (G) and thymine (T), which form the genetic sequence. DNA is located in the nucleus of the cell and is organised into chromosomes, which contain introns and exons. Introns are highly variable and unique, they are often thought to be "junk" code, but likely serve a purpose we are not yet aware of. Exons are where genes are found, these are critical to growth and survival. DNA is transcribed into RNA which can exit the nucleus to be translated into protein. The RNA is read as three bases at a time, known as a codon, which corresponds to one of 20 amino acids that form proteins. There is a large redundancy in the triplicate base sequence, with several codons corresponding to the same amino acid. The triplicate code is degenerate, if a base is substituted it will likely still correspond to the correct amino acid. This is an evolutionary approach to combat small point mutations which occur in the genetic sequence.

Large mutations that occur can trigger cell death, if the cell is not able to survive or rectify the mutation[9]. If the mutation is survivable the error is copied during DNA replication and therefore transferred to the daughter cells during cell division. Large changes in the genome that cause dysregulation in cellular processors will continue in daughter cells, if these changes accelerate the cell cycle or remove checkpoints this can cause cancer.

1.1.1.3.1. Small mutations

Small mutations in the DNA sequence can result in three different outcomes: missense mutation, nonsense mutation or a silent mutation. When one nucleotide within a protein coding region is changed, an alternative triplicate is presented at translation. In missense mutations, an alternative amino acid is added in the protein sequence, which can change the structure or function of the protein. When the alternative RNA triplicate translates to a stop codon the translation process is terminated, this is known as a non-sense mutation, resulting in a truncated protein. Due to the redundancy in codon triplicates coding for amino acids, the altered codon may not impact translation and the amino acid at this position remains unaltered, this is known as a silent mutation.

Small changes in the DNA sequence that are not in protein coding regions can still affect gene and protein expression. If a mutation in the DNA sequence sits within a gene control region, this may cause inappropriate quantities of RNA to be transcribed, through increased or decreased gene transcription, with further impact on protein quantity.

There are several different types of small mutation that can occur with differing effects on the amino acid sequence, examples of which are given in Table 1.1. Depending where the change occurs in an amino acid sequence will determine the impact on the resulting protein. Changes that affect the outer surface of the protein will be less likely to affect the structure or function, however if a hydrophilic amino acid was changed to a hydrophobic amino acid this could affect the structure and function greatly[10]. Changes in the amino acid sequence in a binding site or active site, are more likely to affect the protein function, as with changes that affect the structure and protein folding.

Mutation Type	Sequence								
original	AUG	UAU	CUU	CUU	ACC	GCA	CAU		
	Met	Tyr	Leu	Leu	Thr	Ala	His		
Silent mutation	AUG	UAU	CUU	CUC	ACC	GCA	CAU		
	Met	Tyr	Leu	Leu	Thr	Ala	His		
Point mutation	AUG	UAU	CUU	CCU	ACC	GCA	CAU		
	Met	Tyr	Leu	Pro	Thr	Ala	His		
Frame shift	AUG	UAU	CUU	ACU	UAC	CGC	ACA	U	
	Met	Tyr	Leu	Thr	Tyr	Arg	Thr		
Deletions	AUG	UAU	CUU	~~CUU~~	ACC	GCA	CAU		
	Met	Tyr	Leu	~~Leu~~	Thr	Ala	His		
Insertions	AUG	UAU	CUA	UAU	CUU	ACC	GCA	CAU	
	Met	Tyr	Leu	Tyr	Leu	Thr	Ala	His	
Duplication	AUG	UAU	CUU	CUU	CUU	CUU	ACC	GCA	CAU
	Met	Tyr	Leu	Leu	Leu	Leu	Thr	Ala	His

TABLE 1.1. SMALL MUTATION EXAMPLES. THERE ARE MANY TYPES OF SMALL MUTATIONS IN THE DNA SEQUENCE, HERE ARE EXAMPLES OF EACH TYPE OF MUTATION IN A HYPOTHETICAL SEQUENCE. THE TRANSCRIBED RNA SEQUENCE IS SHOWN ABOVE, WITH THE CORRESPONDING AMINO ACID BELOW, INDICATED BY ITS THREE-LETTER CODE. THE ALTERATIONS IN THE RNA SEQUENCE AND ANY CORRESPONDING CHANGES IN THE AMINO ACID SEQUENCE ARE INDICATED IN RED

Point mutations alter one base in the sequence, exchanging one for another. Point mutations have the potential to change the corresponding amino acid, however due to codon redundancy, the alteration may have no overall impact. In Table 1.1 the example given for point mutation shows a change to the corresponding amino acid from Leucine to Proline, however the example of a silent mutation, the base substitution was instead at the end of the codon, and as such the redundancy in the code ensured the corresponding amino acid was still Leucine.

Insertions and deletions of base-pairs in a multiple of three will increase or decrease the number of amino acids in the protein sequence, as shown in table 1.1. These insertions also have the potential to change the amino acid sequence depending where the insertion or deletion occurs. If an insertion of three base pairs is between codons of an open reading frame, this will simply insert another amino acid into the sequence. If an insertion is within a codon this could insert an amino acid and have no effect on the subsequent sequence (table 1.1), however, it can shift the sequence resulting in the subsequent amino acids changing, causing a missense or non-sense mutation. If the number of base-pairs inserted or deleted is not a multiple of three a frame-shift mutation occurs. Frame-shift mutations often result in missense or nonsense mutations as the open reading frame of a gene is altered, this will give an entirely different amino acid sequence of a truncated protein. The example of a frame shift mutation in table 1.1 shows an insertion of one base which then changes every subsequent amino acid in the sequence.

1.1.1.3.2. Chromosomal abnormalities

Larger mutations in the genome can affect the number, composition, or arrangement of the chromosomes. A normal karyotype of a somatic cell is 23 pairs of chromosomes, 46 in total, in cancer we often see altered karyotypes, due to the genetic instability. Larger mutations can alter the number of chromosomes, the copy number of genes, delete genes entirely, alter the sequence, fuse genes together or even fuse chromosomes together.

During cell division the chromosomes align down the centre of the cell, the spindle fibres from the centrioles attach to the kinetochore associated with the centromeres of sister chromatids. The sister chromatids then separate and are pulled into the two daughter cells. Disruptions in this process result in non-disjunction, where the sister chromatids do not separate, resulting aneuploidy of the daughter cells. Non-disjunction can occur when the cohesion complex holding centromeres of sister chromatids together is not cleaved, or when spindle fibres do not attach firmly to centromere[11,12].

Chromosomal rearrangement is often the result for one or more breakages of a chromosome. When a break in a chromosome occurs the cell machinery will try to repair the damage by reattaching the broken ends. Sometimes the DNA repair mechanisms incorrectly reattach chromosomes, either upside down resulting in an Inversion of that segment or to another chromosome known as a translocation.

Short segments of broken chromosomes can be lost during mitosis, known as a chromosomal deletion, but the large section which contains the centromere is retained. If two or more chromosomes break, non-reciprocal translocations can occur when the short section of chromosome is lost, but a large section is retained and re-joined on the wrong chromosome. Translocations can be reciprocal between non-homologous chromosomes, where two breakages occur, they can be erroneously re-joined to the incorrect chromosome.

When chromosomes are re-joined incorrectly, the genes at either end can become fusion genes. The genes may retain their functionality when transcribed into protein however, low expressing proteins fused into a high expressing protein can cause problems for cell metabolism. A common chromosomal translocation is t(15;17) which is characteristic of acute promyelocytic leukaemia, this causes the fusion gene PML-RARA. The PML-RARA fusion gene is transcribed into a protein which retains the functionality of both elements, although the distribution of the fusion protein is wider and there is an acquisition of additional functionality from the RARA portion of the protein[13,14]. Both the constituting proteins control the proliferation and differentiation of hematopoietic cells, therefor the disruption to these two proteins cases increase proliferation and reduced differentiation[14].

Fragile sites in the chromosome are areas which consist of two or three nucleotides in tandem repetitions. When a chromosome is stained with Giemsa there are pale bands of euchromatin, which is “open” chromatin and is actively transcribed by the cell. The pale bands contain larger proportions of CG baes pairs, which is often where fragile site occur and present as constricted sections of the chromosomes. The repeating sequence in these fragile sites can lead to slippage during DNA replication causing spontaneous increases in the number repeats, further increasing the fragility of the site. The fragile sites break more readily than other points along the chromosome, some fragile sites are common to most people and the associated translocations are seen frequently in cancer[15,16] and other diseases[17]. Together a combination of large mutations and small mutations are the driving force behind oncogenesis when cells are exposed to carcinogens and mutagens in the environment.

1.1.2. Hallmarks of cancer

Cancer is a broad term used to describe many complex diseases with common characteristics of uncontrolled cell proliferation, and in the case of solid tumours, the ability to metastasise. There are at least 100 different types of cancer, many of which also have subtypes, and are often distinct in their presentation. As cancer develops and metastasises the cells invade organs and disrupt the normal functions of the body, eventually the organs shut down and cause death. For a cancer to develop there are six physiological alterations which are required[18] to overcome the anti-cancer defences hardwired into every cell's DNA, which explains why cancer is relatively rare over the course of a person's life. The way cancer presents, responds and progresses, can be due to person-to-person variation as well as the different cancer type; there has been much research into cancer, yet further research is needed to fully understand the variation and overcome these obstacles.

The six physiological alterations within a cell which are required for the development of cancer[18] include autocrine growth signalling, insensitivity to antigrowth signalling, evasion of apoptosis, unlimited potential for replication, angiogenesis, as well as tissue invasion and metastasis. Each of these changes creates a crack in the anticancer defence mechanisms encoded in each cell, when all six changes occur the anticancer mechanisms are overwhelmed, and cancer can thrive. These hall marks are based on the changes in solid cancers which don't accurately describe haematological malignancies due to its non-epithelian origin. It is suggested additional mutations are required for solid tumours to become metastatic [18], which is not the case in blood cancers[19]. Although on occasion blood cancers can cause chloroma, which is a collect of precursor cells in the tissue most often the lymph nodes, this is not the same as metastasis. In addition, there is no evidence angiogenesis occurs in blood cancers as the tissue of origin is already vascularised and the motile cells are free to enter the blood stream due to their origin inherent properties.

In addition to the six physiological changes, there are also two main enabling characteristics which make these changes more likely[20]. The first enabling characteristics is genomic Instability, which allows the acquisition of mutations through large and small genetic aberrations, as well as epigenetic changes. The acquired genetic changes can alter cell function, including the pathways leading to one of the six physiological changes. The second enabling characteristics tumour-Promoting Inflammation which encourages immune cells to present tumour antigens to the dendritic cells, which dampen any positive anti-tumour immune responses using various immune evasive and adaptive resistance mechanisms.

1.1.2.1. Physiological changes

1.1.2.1.1. Self-sufficiency in signalling

Typically, cells require several external growth signals to exit the G0 stage of the cell cycle and enter the G1 stage[21]. Throughout the cell cycle there are several checkpoints[22], in cancer the genes associated[23] with these checkpoints are altered allowing unchecked cell cycle progression and proliferation. Cancer cells are able to overcome external cell cycle checkpoint inhibitors, by producing or acquiring higher amounts of positive growth-signals compared to negative growth signals[18,20]. Cancer cells become self-sufficient by increasing the production of growth factors by the cancerous cells, over expressing or mutating the growth signalling receptors or constitutively activating internal signalling pathways[18,20].

1.1.2.1.2. Insensitivity to antigrowth signalling

There are two mechanisms in which exogenous factors prevent proliferation, the first is promotion to the quiescent (G0) state of the cell cycle, which is a reversible process[24], the second process is irreversible, by entering a post mitotic state associated with differentiation[18,20]. Typically, after cell division, cells become quiescent meaning a sufficient number of cells has been achieved to replenish old or dyeing cells, therefore the cell is not required to divide[24]. This quiescence is achieved through anti-proliferative cell signalling[20], indicating no growth is required. Normal cells in quiescence continually monitor the external environment for pro and anti-proliferative signals which promote the maintenance of quiescence, differentiation or proliferation[18]. For a cancer to develop the cell must overcome both quiescence and differentiation pathways of anti-growth signalling[25]. Antigrowth signalling happens through both soluble and immobilised inhibitors in the extracellular matrix (ECM) and surface of nearby cells, both are important for anti-growth pathways to function as they work together to transduce external signals through the cell membrane to begin intracellular signalling cascades[18]. In cancer the cells do not respond to anti-growth signals and continue to proliferate in an uncontrolled manner which encourages the accumulation of mutations[26] and in turn increases the dysregulation of cellular checkpoints.

1.1.2.1.3. Evasion of apoptosis

There are two main types of cell death, necrosis and apoptosis[27]; necrosis is uncontrolled cell death which typically causes inflammation and is the result of trauma through injury or hypoxia, apoptosis is known as programmed cell death, it is the active process of removing damaged cells, old cells, or to create key structures during development. Apoptosis can be

triggered through DNA damage, reduced growth cytokines, infection and hypoxia, all consequentially activate intrinsic and extrinsic apoptosis pathways. There is a balance between anti-apoptotic and pro-apoptotic signalling, the extracellular and intracellular environment are constantly under surveillance by the cell to identify pro-apoptotic signals, which will trigger the apoptotic pathway, by downregulating the expression of anti-apoptotic genes and increasing pro-apoptotic genes.

Cancer cells acquire resistance to apoptosis through insensitivity to pro-apoptotic signals, by upregulation of anti-apoptotic genes and proteins[20,28], and the downregulation of pro-apoptotic genes and proteins[29]. Mutations within pro-apoptotic genes can downregulate their expression or prevent their ability to signal[18]. A common mutation in cancer occurs in the p53 tumour suppressor[20,30] gene which inactivates the protein, this mutation is seen in over 50% of human tumours[18,30]. Mutations within anti-apoptotic genes can upregulate the expression or cause constitutively active proteins, a known example in cancer is the Ras family of proteins, particularly k-ras which is mutated in 22% of cancers[31]. Changes in crucial apoptotic regulators can drastically shift the balance in favour of the anti-apoptotic mechanisms within a cell[18]. The BCL2 protein is over expressed in many cases of AML, although sometimes associated with a translocation[32] the increase expression can also occur randomly[33]. BCL2 blocks apoptosis by sequestering pro-apoptotic signalling molecules, therefore preventing mitochondrial permeabilization[34]. To combat this upregulation of BCL2, the drug Ventrolax was FDA approved[34–36] and has been shown to improve overall outcome in patients[36]. Although mutations in TP53 are relatively infrequent in AML[37–40] compared to other cancers, dysfunction of this tumour suppressor is relatively common[38,41]. Dysfunctional TP53 is associated with poor overall survival[37,41] and poor response to treatment[39,40]in AML.

1.1.2.1.4. Unlimited potential for replication

Each cell has a pre-determined replication limit, which restricts the number of cell divisions a cell can undergo through intrinsic, autonomous sensors. Once the limit for the cell has been reached the cell enters a state of permanent senescence[18,20,42] and eventually apoptosis[37,43]. Telomeres are repeating sequences of bases at the end of the chromosomes, the length of which are correlated to the replication limit of the cell[18,20]. During each replication some of the base pairs are lost from the end of the chromosome as DNA polymerase is unable to fully replicate the 3' end[18]. Telomeres function as a protective cap on the chromosome which is degraded before the genetic sequence. Once the telomeres have reached a critical length this symbolises the replication limit of the cell.

For cancers to become immortal the cells must overcome the intrinsic replication limit. Cancerous cells maintain telomeres through the upregulation of telomerase[18,20,42], an enzyme with DNA-synthesis capabilities, specifically the ability to synthesise the telomerase and increase their length. Cancerous cells can also activate cellular mechanisms which prevent the shortening of the telomeres through chromosomal exchange of sequences[18,20]. Both mechanisms maintain the length of the telomere to prevent end to end chromosomal fusion[18,20] therefor avoiding senescence and cell death.

1.1.2.1.5. Angiogenesis

Angiogenesis is the formation of new blood vessels which is a highly regulated process in normal tissue development and growth[18,20], which is required to supply the tissue with oxygen and nutrients, as well as deliver immune cells to the tissues. Angiogenesis allows a cancer to use the bodies resources, enabling it to grow, expand and metastasise[18,20]. In order for angiogenesis to occur the cancerous cells secrete pro-angiogenic factors[44], such as VEGF[45,46] and HIF1A[47], and inhibit the production of anti-angiogenic factors[18,20]. This hallmark is only relevant to solid masses, where haematological malignancies[19] already receive the privilege of a rich bone marrow niche which is highly vascularised and nutrient rich.

1.1.2.1.6. Tissue invasion and metastasis

A sign of a progressive tumour is its ability to invade neighbouring normal tissue, this makes excision difficult during surgery. Another critical point in tumorigenesis is the ability for cancerous cells to metastasis to different locations around the body[18,20]. There are several steps in the process of invasion and metastasis, involving cells from the main tumour detach from the mass and enter the blood or lymphatic systems, which carries the cells to distant tissues where they exit the circulatory system and begin to for a new mass[20]. although this process is rarely successful it is responsible for 90% of cancer related deaths[18], which is why it is crucial to find cancer at an early stage which is easier to treat. This is only applicable to solid tumours as haematological malignancies are privileged[19] by their tissue of origin and can enter the blood stream readily.

1.1.2.2. Enabling characteristics

1.1.2.2.1. Genomic Instability

Molecular changes can occur in the genome of a cell, there are many cellular mechanisms[48] to repair incorrectly transcribed portions of DNA or broken strands of DNA[20], which readies the cell for mitosis. On occasion the mechanisms incorrectly repair the genetic sequence

which can lead to permanently transcribed mutations, most of the time this leads to apoptosis however occasionally the cell cycle progresses, and the mutation is passed to the daughter cells. If a mutation occurs in the mechanisms involved in detecting and inactivating mutagenic compounds, detecting mutations and activating repair, or the DNA repair machinery[20], the cellular components preventing tumorigenesis are effectively null. Mutations can increase the speed of the cell cycle and lessens the hold on the checkpoints, the DNA replication speed is increase and the proofreading mechanism is degraded and may increase the number of mistakes in the replication.

1.1.2.2.2. Tumour-Promoting Inflammation

The tumour microenvironment plays a large part in immune regulation and tumour promotion, through the secretion of growth factors, cytokines and reactive oxygen species the tumours[49] can encourage immune cell infiltration[18,20]. Tumours are often infiltrated by immune cells of both the innate and adaptive immune system in varying capacities[20]. The infiltrating immune cells inadvertently support the cancer[50] through the provision of small bioactive molecules, including those effective in growth, survival, pro-angiogenesis, and extracellular matrix modification[20].

1.1.2.3. Emerging hallmarks

In 2011, Hanahan and Weinberg revisited their publication to review and update the critical theory[20]. In order to explain further crucial factors that promote cancer growth and metastasis they included two additional hallmarks.

1.1.2.3.1. Evasion of immune destruction

The immune system looks for danger in the body in the form of pathogens, foreign bodies, and inflammation. There are immune cells residing within the tissue which constantly survey their environment to identify danger, in addition, there are circulatory and regulatory immune cells, all play a crucial role in defending the body. Somatic cells present MHCI to cells of the immune system, if an immune cell recognises a cellular defect, they can communicate this to other immune cells. Often in cancer, the immune cell recognises an abnormality and presents these antigens to dendritic cells, which identifies the antigen as “self” and promotes quiescence, effectively terminating the immune response[51]. The promotion of quiescence for the “self-reactive”, tumour recognising, immune cells weakens the immune response and the tumour is able to continue growing. In addition to cellular quiescence on the presentation of tumour antigens, the tumours can actively recruit T-regulatory cells[20,52] to

the site which dulls the immune response therefore allowing the cancer to grow with impunity.

Cancerous cells downregulate the production of immune signalling molecules, including MHC[53], which limits the ability of immune cells to react to them. Tumours have high levels of immune suppressing cytokines, including TGFβ1, which modulate the immune reaction to the tumour[20].

1.1.2.3.2. Reprogramming of energy metabolism

As cancer cells divide at such a rate, they require larger amounts of energy compared to normal tissue. Aerobic respiration is the preferred method of respiration by normal cells, although anaerobic respiration is a secondary mechanism which allows continued respiration even when the cells are deprived of oxygen, for example during exercise. The first step in aerobic respiration involves the cleavage of glucose in the glycolysis reaction, this is the same as anaerobic respiration. The glycolysis reaction produces far less ATP compared to aerobic respiration, and yet it is the preferred method of respiration for cancerous cells[20,54]. Despite anaerobic respiration being less efficient it produces bioactive intermediates[54], which are used by the cancer cells for the generation of nuclear bases and amino acids[20,54], these are crucial for the construction of organelles and macromolecules in cell division. The increased requirement for glucose by cancerous cells is facilitated through increased expression[55] of glucose transporters[56,57] on the surface of their cells[20]. Continued glycolysis is associated with the upregulation of oncogenes and downregulation of tumour suppressor genes[45,49,58].

1.1.3. Cancer statistics

Cancer affects around 367,000 people each year in the UK[59], although this number is increasing year on year, around 50% of the population will be diagnosed with cancer in their lifetime. In general, the number of cancer diagnosis are increasing across all cancer types, however some cancers have a larger increase than others. For example, thyroid cancer diagnosis has increased 68% in males and 69% in females, but leukaemia has increased 7% in males and 8% in females[60]. There is large disparity between the increases of different cancers however the general trend is more cases are diagnosed each year. Overall, 50% of people diagnosed with cancer will die from it, which accounts for 28% of all deaths in the UK[61]. In general mortality in cancer is decreasing[61] as screening improves, diagnosis happens earlier when cancer is more treatable, in addition, advancements in treatment have improved cancer outcomes. It is important to recognise 38% of cancer cases could have been

prevented[62] through lifestyle changes, smoking is the largest cause of preventable cancers followed by obesity.

1.2. Acute Myeloid Leukaemia

1.2.1. Haematopoiesis

Haematopoiesis is the production of all the cellular components of the blood, where haematopoietic stem cells (HSC) differentiate into one of two lineages (figure 1.1). Haematopoietic stem cells are multi-potent, owing to their ability to self-renew and differentiate into all blood cells[63], whereas myeloid and lymphoid progenitor cells are Oligo-potent due to the commitment to their respective lineages[63].

The differentiation from HSC to the lymphoid and myeloid lineage relies heavily on the cytokines in the bone marrow niche[63]. Typically, only the terminal cells are fully functional where progenitor cells lack crucial functionality, and act predominantly as precursors to maintain healthy cell populations.

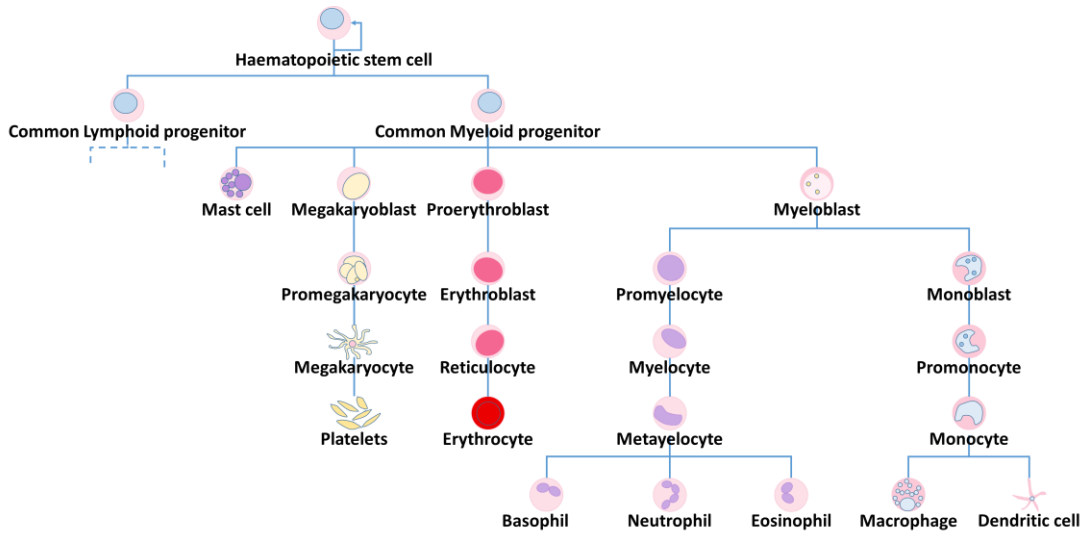

FIGURE 1.1. BLOOD CELL DIFFERENTIATION. THERE ARE TWO LINEAGES OF BLOOD CELL DIFFERENTIATION, LYMPHOID AND MYELOID, THE LATTER IS SHOWN HERE.

1.2.2. Haematological Malignancies

There are four main types of leukaemia as shown in table 1.2, the nomenclature used is dependent on the haematopoietic lineage and differentiation of the cells involved in the cancer. The myeloid and lymphoid lineages differentiate early in haematopoiesis as demonstrated in figure 1.1, meaning the cancers are distinct and separate. Chronic

leukaemia develops relatively slowly, and symptoms are slow to become apparent, compared to acute leukaemia where symptom onset and cancer development is rapid. Each type of leukaemia has a different prevalence, where CLL is the most prevalent but the mortality rate for AML is the much higher.

	Myeloid		Lymphoid	
	Incidence	Mortality	Incidence	Mortality
Acute	3102	2601	794	253
Chronic	763	219	3789	1008

TABLE 1.2. INCIDENCE AND MORTALITY OF DIFFERENT LEUKAEMIA'S. THE INCIDENCE OF EACH TYPE OF LEUKAEMIA DIFFERS, IN THE UK CLL IS MOST COMMON AND CML IS THE LEAST COMMON[64–67]. OUT OF THE FOUR MAIN TYPES OF LEUKAEMIA, AML HAS THE HIGHEST MORTALITY RATE, OF 83.9%, WHEREAS ALL, CLL AND CML HAVE A MORTALITY RATE OF 31.8%, 26.6% AND 28.7% RESPECTIVELY.

1.2.3. AML an overview

AML is a cancer affecting the cells of the myeloid lineage of blood cell differentiation (figure 1.1) and is characterised by the accumulation of immature blood cells within the bone marrow. within leukaemia, AML is the largest cause of patient death (table 1.2) owing to the fact the peak incidence rate is between the ages of 85 and 89[64], at which time treatment options are limited due to co-morbidities and treatment intensity. As with most cancers the number of AML cases each year is increasing over time[64], despite this, treatment options remained unchanged for the last 40 years. However, over the last 10 years several therapies have been developed and there have been multiple FDA approvals for new AML treatments, including targeted therapies like Gemtuzumab Ozogamicin for CD33 positive[68] AML and Enasidenib for IDH2[69] mutated AML.

AML is a heterogeneous cancer with 5234 Known driver mutations[70], some of the common mutations are used in the World Health Organisation (WHO) classification system, where there are 23 groups[71]. As the WHO classification system has many groups, it is difficult to assign risk, but looking directly at mutations can give an indication of which targeted treatments may be available. An alternative classification method is the European LeukaemiaNet[72] (ELN) classification, which consolidates the common mutations into three risk groups (table 1.3), eliminating the issue of too many sub-groups in the WHO system.

ELN Risk category	Genetic abnormality
Favourable	RUNX1-RUNX1T1 t(8;21)(q22;q22.1)
	CBFB-MYH11 inv(16)(p13.1q22) or t(16;16)(p13.1;q22)
	Mutated NPM1 without FLT3-ITD or with FLT3-ITD^{low}
	Biallelic mutated CEBPA
Intermediate	Mutated NPM1 and FLT3-ITD^{high}
	Wild-type NPM1 without FLT3-ITD or with FLT3-ITD^{low}
	MLLT3-KMT2A t(9;11)(p21.3;q23.3)
	Cytogenetic abnormalities not otherwise classified
Adverse	DEK-NUP214 t(6;9)(p23;q34.1)
	KMT2A rearranged t(v;11q23.3)
	BCR-ABL1 t(9;22)(q34.1;q11.2)
	GATA2,MECOM(EVI1) inv(3)(q21.3q26.2) or t(3;3)(q21.3;q26.2)
	-5 or del(5q); -7; -17/abn(17p)
	Complex karyotype (Three or more chromosomal abnormalities)
	Monosomal karyotype
	Wild-type NPM1 with FLT3-ITD^{high}
	Mutated RUNX1
	Mutated ASXL1
	Mutated TP53

TABLE 1.3. SUMMARY OF MUTATION ADJUSTMENT OF THE ELN CLASSIFICATION. THE EUROPEAN LEUKAEMIANET[72] RISK CLASSIFICATION FOR AML IS BASED ON THE MUTATIONS PRESENT IN A PATIENT SAMPLE. THIS TABLE HAS BEEN MODIFIED FROM DIAGNOSIS AND MANAGEMENT OF AML IN ADULTS: 2017 ELN RECOMMENDATIONS FROM AN INTERNATIONAL EXPERT PANEL[73].

The baseline for all patients is intermediate risk, certain mutations increase or decrease the risk assigned, as detailed in table 1.3. Another method of classifying AML is the French-American-British (FAB) classification, which looks at cell morphology and differentiation, this was developed in the 1980's, and mutation-based classifications have overtaken its use in clinical diagnostics.

1.2.4. AML clinical presentation and diagnosis

Symptoms associated with AML[74] include fatigue, breathlessness, fever, weight loss, frequent infections, poor clotting or frequent bleeding, bruising easily, bone or joint pain. AML is diagnosed through a peripheral blood sample or bone marrow biopsy and aspirate; it is characterised by a blast percentage greater[74] than 20%. Myeloblasts are myeloid progenitor cells which have not reached terminal differentiation, these immature myeloid cells do not function normally[75]. The clonal expansion of the immature blood cell population is rapid, which crowds the bone marrow causing the normal cell population to be reduced[76], which explains some of the clinical symptoms of AML including poor clotting and susceptibility to infection.

Wherever possible a peripheral blood sample is used for diagnosis, however if there is uncertainty in the result a bone marrow biopsy will be required[74]. The necessary confirmatory tests can be performed on a peripheral blood sample where a patient has a high blast percentage, reducing the need for a bone marrow biopsy as these are invasive[74]. The sample is tested using a variety of methods including flow cytometry looking at markers of differentiation, PCR to determine mutation status, metaphase cytogenetics or fluorescence in situ hybridization (FISH) to identify karyotype and large genetic abnormalities, as well as a whole blood smear to visualise the differentiation of the cells[74]. The variety of tests used help in the classification of the AML and will thus determine which treatment option is most appropriate.

1.2.5. AML treatment

There are many different subtypes of AML according to each of the classification methods, however the standard treatment is the same for all of them. The initial treatment for AML is cytarabine for 7 days, followed by an anthracycline for 3 days[77], The type of anthracycline used in treatment depends on the overall health and comorbidities, so is selected on a case by case basis. For patients with the t(15;17) translocation, also known as Acute Promyelocytic Leukaemia (APL), there is an alternative treatment available. The treatment for APL is a

combination of arsenic trioxide and all-trans retinoic acid[13,78–80], in some cases the standard chemotherapy is also required. APL has a positive overall survival with a cure rate of 80-90%[81], compared to that of AML with a survival rate[64] of around 20%.

Remission in AML is achieved in 60-80% of adults[72,82], of those 50% will relapse[82], which is a concern for patients and clinicians as it leads to further complications and a high likelihood of adverse outcomes. One key problem with relapse AML is the increased incidence of treatment resistance. Over the last 10 years more treatments have been developed and approved for use, primarily in relapse patients but the approval is often extended to initial treatment.

There is a portion of AML patients who are CD33-positive, in this instance Gemtuzumab Ozogamicin (GO) is applicable. GO is an antibody drug conjugate which targets the CD33 molecule and delivers the drug directly to these cells. The drug was initially approved in 2010 but later withdrawn from the market due to increased mortality among newly diagnosed patients, it was then reapproved in 2017 using an alternative dosing schedule[68,83–85]. The initial approval was based on relapse and refractory patients but the approval in 2017 extended the application to newly diagnosed patients[68].

In 2017 Enasidenib[69,86] was approved, for relapse and refractory AML patients with IDH2 mutation, this gene is associated with metabolism and the mutated form is reported to block myeloid differentiation[87,88]. In addition, Ivosidenib was approved for IDH1[88] mutation in 2019[89] although was later withdrawn[90] despite evidence the drug was effective[91] on the grounds the risk outweighed the benefit to patients.

Over expression of BCL-2 causes inhibition of the apoptotic pathways, which is found in multiple cancers including AML[92]. Venetoclax, a BCL-2 inhibitor, was initially[93] FDA approved in 2018 for older patients with co-morbidities[94] meaning they were not suitable for intensive chemotherapy; this approval was extended[95] in 2020. In the UK this drug is only approved for use in patients with CLL, although it may be used with discretion for other cancers.

The drug midostaurin[96] was approved in 2018 for treatment of adult patients with newly diagnosed acute myeloid leukaemia who are FLT3 mutation positive[97,98]. The gene encodes a tyrosine kinase (TK) protein which is important for cell division and differentiation in early haematopoiesis[92]. FLT3 is mutated in 30% of AML cases[99] and is associated with a worse overall prognosis. Gilteritinib[100,101] is another TK inhibitor, also FDA[102] approved in 2018 and

approved by NICE[103] in 2020 for patients with relapse or refractory AML with the FLT3 mutation.

All the newly approved targeted treatments available are a step in the right direction after 40 years of the same chemotherapeutic schedule. Most of the initial approvals are in relapse and refractory AML is because this is often harder to treat and may not respond to standard chemotherapy, so patients are referred to clinical trials. Once the initial testing has established safety and efficacy in relapse and refractory patients, the drug is then trailed in diagnostic patient, where the approval is often extended. The treatments offer hope for a better survival rate for AML, especially in patients who are not suited to intensive chemotherapy or experience relapse or refractory AML.

1.3. Relapse

As part of a patients care plan, patients are monitored and tested regularly throughout treatment and into remission[73,104]. After a patient has completed a course of treatment a sample is taken[73]: a cancer is said to be in remission when no abnormal cells are detected in a post treatment sample. As tests for residual cancer cells focused on cell morphology in a post treatment bone marrow sample, minimal residual disease (MRD) is defined as the presence of abnormal cells after treatment at numbers below the sensitivity of routine morphology examinations[104,105]. The residual cancer cells that are undetected have the potential to begin clonal expansion, and cause a reoccurrence of the cancer, known as relapse[81,104,106]. Relapse is associated with the presence of a small sub population of cells at diagnosis or the clonal evolution of the main population, where further mutations are acquired[106], in both these cases the populations harbour mutations which give them an advantage to resist treatments.

Although molecular testing for MRD is not explicitly recommend, it is recognised by the National Comprehensive Cancer Network AML 2019 Clinical Practice Guidelines as an important area of expansion[107], but the molecular tests available have limitations that need to be adressed[104], In contrast the ELN does recommend testing be implemented in a clinical setting to improve patient care[73], despite the current limitations. The molecular tests available are quickly becoming standard practice and can detect molecular abnormalities within a sample that is not evident by morphology screening[108]. These molecular tests can show if additional treatment is required and often detect relapse sooner, leading to faster treatment and better patient outcomes.

Relapsed cancer can be difficult to treat, as a reoccurrence from clonal expansion of cells which have already survived treatment and may have treatment resistance mutations[106,109,110]. By predicting relapse in AML patients at diagnosis, the high-risk patients could be referred to an alternative treatment strategy sooner, this could include, longer initial treatments, additional rounds of treatment, more intense treatment, referral to clinical trials or a bone marrow transplant earlier in the treatment. As an added benefit to predicting relapse at diagnosis, high-risk patients could be monitored more frequently and molecular testing could be utilised to detect relapse early giving a patient the best chance of second remission. As mentioned previously, the relapse rate in AML is exceptionally high with 50% of patients who achieve remission eventually relapsing[82].

1.4. Biomarkers

1.4.1. Biomarker overview

Biomarkers are a measurable biological component that are indicative of a disease or condition[111]. Biomarkers can be measured directly from a patient like temperature and blood pressure[111]; however, some biomarkers require a patient sample to be taken. Within patient samples we focus on molecular biomarkers, including DNA[78], RNA[112], proteins[113,114] and cytokines[115,116], among others.

As already discussed the genetic changes in a cell are transcribed into RNA[112] which in turn are translated into protein. The molecular changes represent system dysregulation and the impact on function, growth, communication, and structure in the body. By looking into the molecular changes that are typical of a disease, molecular biomarkers can be discovered and utilised for treatment, monitoring and management of different diseases. Within many different diseases, biomarkers are being used as part of standard care procedures, throughout diagnosis[117], treatment[118] and monitoring[119].

1.4.2. Different types of biomarker

There are seven different types of biomarker[120], as shown in table 1.4, although different sources divide biomarkers into different categories[121,122]. Each type of biomarker is defined to give the overall aim and how they should be implemented, to ensure a patient receive appropriate care for their specific condition. Biomarkers are being introduced as clinical tools more often, for many different diseases, some examples in relation to cancer are shown in table 1.4.

In AML the t(15;17) translocation is often used in diagnosis of Acute Promyelocytic Leukaemia[13,78,79,123], this translocation is associated with the PML-RARA fusion gene and the patients have a much better prognosis. Patients with the t(15;17) translocation receive a different treatment strategy[78,79,123] involving arsenic-trioxide and all-trans retinoic acid. Another key mutation in AML is the FLT3 internal tandem repetition, which is a receptor tyrosine kinase that regulates haematopoiesis though the activation our pathways leading to apoptosis, proliferation or differentiation of hematopoietic cells in bone marrow[124]. Mutations in the FLT3 gene are associated with a worse overall outcome[123] and indicates tyrosine kinases inhibitors may be beneficial as a complementary therapy[96,100].

Good biomarkers can predict their determined goal with 99% sensitivity and specificity, however with many biomarkers we often see an 80-90% sensitivity or specificity, as such these are often used in addition to conventional tests; For example, a high PSA test result indicates a likelihood of prostate cancer, however due to the uncertainty[125] the patient will be referred to receive a biopsy test to confirm abnormal cells. Another example is HPV[126,127] which causes 99% of cervical cancers[128], which is the initial step for the cervical screening process which then looks for the presence of abnormal cells.

Type	Application	Example
Diagnostic	Detection of early disease and subtype	Prostate cancer: Prostate specific antigen
Prognostic	likelihood of an outcome or clinical event	Prostate cancer: Gleason score
Predictive	Prediction of effective treatment response	Brest cancer: HER2, ER, PR
Response	Identification of a response to therapy	Solid cancers: Tumour volume
Risk	Identifying those who may develop the disease	Cervical cancer: HPV
Monitoring	Assessing disease status	Solid cancers: Tumour volume and metastasis
Safety	Indicate adverse effects from therapy	All cancers: Neutropenia

TABLE 1.4. SEVEN MAIN TYPES OF BIOMARKERS. ALTHOUGH BIOMARKERS ARE AVAILABLE FOR MANY DIFFERENT DISEASES, THE EXAMPLES GIVEN ARE CANCER SPECIFIC[120].

1.4.3. Methods of biomarker discovery

There are numerous ways biomarkers can be discovered through bioinformatics approaches, involving statistics and machine learning[129,130]: used in combination these are exceptionally powerful tools. To discover biomarkers through any method, patient data is required, which can be limited in sample size, clinical annotations, and the type of molecular data available. The most attainable source of patient data is publicly available datasets, which have been collected from trials or for publication, and can be found in data repositories[131,132]. As part of the biomarker discovery process, several datasets are needed discover and validate the results obtained. Although it is important to use several datasets, different sources will have different annotations, methodologies, patient sampling methods, platforms, treatments, patient follow up, all of which will have to be considered and controlled for during biomarker discover and validation.

1.5. TGFβ and CD109 in cancer

TGFβ1 is a regulatory cytokine with a multitude of capabilities including proliferation, differentiation, survival, adhesion, as well as controlling the cellular microenvironment[18,20]. The many different rolls of TGFβ1 coincide with its tumour-suppressive effects[133], although cancer cells manipulate[134] the pathways to create a pro-tumorigenic environment. The tight regulation of cellular processes through TGFβ1 signalling can be avoided or repurposed by cancer cells[134,135], inactivation of pathway components, or alterations in downstream elements to disable the tumour-suppressive functions of this pathway are both associated with genetic mutations in cancer[136].

By disabling the tumour suppression pathway of TGFβ1 the remaining functions can be utilised to promote tumour progression and metastisis[133]. Mutations in the TGFβ receptor family are prevalent in cancer[133,136], these receptors mutations can inactivate the TGFβ1 pathway or specific components of the pathway[136]. In addition to mutations in the receptors themselves, downregulation of the TGFβ1 receptors and ligand trapping reduces the capacity of the TGFβ1 signalling pathway[133], both of these regulatory elements are performed by co-receptors. There are many co-receptors of TGFBR1, one of which is CD109[137], although not well characterised, it has been shown to be highly expressed in many different cancers[138–143]. CD109 downregulates TGFβ1 signalling through binding of TGFBR1[144,145] and promoting its internalisation and degradation[146], in addition, the soluble form of CD109 is able to bind and

sequester TGFβ1[145,147]. High CD109 has been shown to impact heavily on cancer progression[140] and clinical outcome[141,144].

TGFβ1 enables immune tolerance, tumours that produce high levels TGFβ1 may be protected from immune surveillance. A defective TGFβ1 response in immune cells can lead to chronic inflammation and the production of a pro-tumorigenic environment[133]. Tumour-derived TGFβ1 may recruit other stromal cell types such as myofibroblasts and osteoclasts enabling tumour metastisis[133] through angiogenesis and epithelial to mesenchymal transition.

1.6. Aims and objectives

There are two main aims for this investigation, the first is to develop a robust method of biomarker discovery, which will speed up the discovery pipeline in the future. Different methods of biomarker discovery will be trialled to find one that is quick and easy to understand. By finding a reliable method of biomarker discovery, the number of biomarkers available will increase and influencing the way patients are treated and monitored. To achieve this aim a relapse prognostic index for AML will serve as the trial for measuring the success of the biomarker discovery methods used. Relapse is a starting point of this investigation as it is a common complication in cancer, the rate varies according to the cancer type, but is of serious concern to AML patients who have a high rate of relapse, which in turn is difficult to treat and has a poor overall prognosis. The second aim of this investigation is to identify clinically relevant genes as they are discovered in the workflows and determine their impact in AML, with the objective of identifying their function in treatment resistance and relapse. These three aims allow for the method to be developed, create a useful predictive tool, and apply the identified genes to a range of different cancers.

Chapter 2 - Methods and materials

2.1. Datasets

2.1.1. Downloading and processing datasets

Datasets compiling RNA expression from AML patients were identified by searching online dataset repositories, including Array Express and Gene Expression Omnibus. From the AML datasets identified, those without relapse data were removed. It is essential to have multiple datasets available, one for discovery and another one or more for validation purposes

2.1.1.1.HOVON

The CEL files for the HOVON[148] dataset were downloaded from array express[149] and were merged using the "ReadAffy" function from the R package *affy*[150]. The dataset was quality checked using the *ArrayQualityMetrics* package to identify outliers, looking at the Relative Log Expression (RLE) and the Normalized Unscaled Standard Error (NUSE). Once the outliers were removed from the dataset, it was RMA normalised using the *affy*[150] package. The "collapse rows" function from the R package *WGCNA*[151] was used to select one representative probe-ID for each gene, the settings used select the probe-ID with the fewest missing values, and secondarily the probe-ID with the largest mean value to resolve any tied probes. Biomarker discovery was performed in this dataset as it is the largest, which reduced false biomarkers being identified.

2.1.1.2.TCGA-LAML

The TCGA[152] dataset was downloaded using the R package *TCGAWorkflowData*[153], all cases were downloaded from the repository as FPKM data. The dataset was quality checked using the *arrayQualityMetrics*[151] package, looking at the distance between arrays, any outliers were removed. This was used as a validation dataset.

2.1.1.3.TARGET -AML

TARGET-AML[154] was downloaded from the TARGET site in the RPKM data format. The dataset was quality checked using the *ArrayQualityMetrics*[151] package, looking at the distance between arrays. This was used as a validation dataset.

2.1.2. Patient selection

	TCGA	TARGET	HOVON
Platform	Illumina RNA-Seq	Illumina RNA-Seq	Affymetrix Array
Accession number	TCGA-LAML	TARGET-AML	E-MTAB-3444
Total patients	91	119	382
EFS	27	35	151
relapse	64	84	231
Age 0-18	0	118	13
Age 19-65	73	1	346
Age 65+	18	0	23
Male	49	63	191
Female	42	56	191
FAB M0	10	2	11
FAB M1	24	14	90
FAB M2	26	32	110
FAB M4	19	34	73
FAB M5	9	28	90
FAB M6	2	2	8
FAB M7	1	7	0
ELN risk			
Low	16	60	151
Intermediate	61	54	199
high	14	5	32
Clinical Mutations			
t(8;21)	1	20	36
inv(16)	5	24	33
NPM1	46	5	138
FLT3	58	13	103
CEBPA mutation	NA	6	22
MLL	62	22	28

TABLE 2.1. SUMMARY OF PATIENT DATASETS. RELEVANT CLINICAL INFORMATION AND PATIENT NUMBERS FOR EACH DATASET USED IN THE ANALYSIS. EFS INDICATES EVENT FREE SURVIVAL PATIENTS, ELN INDICATES EUROPEAN LEUKEMIC NET CLASSIFICATION, AND FAB INDICATES THE FRENCH-AMERICAN-BRITISH CLASSIFICATION.

To ensure the biomarkers discovered were specific to relapse, certain filters were applied to all the cohorts. Only diagnostic samples were used, all other time points were removed from the dataset. Only event free survival (EFS) and relapse patients were used, patients who experienced different events were removed, an additional check was applied to the EFS patients, to remove patients who had died from other causes which may have been described as event free.

Other filters were also applied to improve the quality of the score developed. If the FAB or ELN classification were missing the sample was removed, due to the difference in treatment all patients with FAB-M3 and t(15;17) translocation were removed. Table 2.1 shows a summary of the patient numbers in the analyses.

2.2. Bioinformatics techniques

To identify biomarkers in large datasets a variety of bioinformatics techniques are implemented. Using bioinformatics provides a consistent and replicable way of analysing data, and doing this through machine learning or statistics programmed in R speeds up this process. Through statistics and machine learning, biomarkers with high accuracy can be identified in one dataset and validated in another.

2.2.1. Artificial neural networks

Artificial neural networks (ANN) find association between an input and the possible variables in a given dataset, it creates networks of association like those of neurones in the brain. There are two main types of ANN, binary and continuous: Binary ANN must have an input of integer values, for example 1 or 0, continuous ANN uses a continuous variable as the input.

Binary ANN was used, by classifying cases based on their relapse status, event free survival (0) or relapse (1), the output was sorted by average test error. The gene expression values identified either from the binary ANN or another statistical test, were used as continuous inputs for the continuous ANN, the output was sorted by average test error. The parameters of the ANN were set to 300 epoches, 100 windows, 50 bootstraps and 10 loops.

The data-mining algorithm comprised a three-layer multilayer perception architecture modified with a feed forward back-propagation algorithm and a sigmoidal transfer function, as previously described by Lancashire *et.al*[155].

2.2.2. Welch's T-test

T-tests were used to determine if the means of two groups were significantly different from each other[156]. The assumptions in this test are; the two samples have unequal variance or sample size; the sample is normally distributed, and the samples are unpaired. A $P<0.05$ is considered statistically significant, indicating the means of the population are significantly different. The "t.test" function in R was used to calculated P-values, this test was used to determine if the mean gene expression of relapse patients significantly differed from event free survival patients.

2.2.3. Mann–Whitney U test

Much like the t-test explained above, Mann–Whitney U tests were used to determine if the means of two groups were significantly different from each other[157], however the mean is calculated based on the sum of the rank order of the samples, which gives less weight to anomalies within the data. A $P<0.05$ is considered statistically significant, indicating the means of the population are significantly different. This test was used to determine if gene expression of relapse patients significantly differed from event free survival patients. The "wilcox.test" function in R was used to calculate the P-values, where this statistic was used independently. It was also computed in GraphPad in association with Kaplan-Meier survival curves.

2.2.4. Kaplan–Meier survival curve

Kaplan–Meier survival analysis curves, illustrate the survival of a population based on an event and time to the event[158], in this instance the event was relapse (1), the censoring event was event free survival (0), and the time used was event free survival time. It was used to compare the difference of event free survival time in populations of high or low relapse prognostic index. A Mann-Whitney u test was calculated alongside the Kaplan-Meier survival curve, which is detailed previously.

2.2.5. Receiver operator characteristic curves

Receiver operator characteristic[159] (ROC) curves illustrate the diagnostic ability of a variable, by look at the true positive rate versus the false positive rate. When analysing a ROC curve the sensitivity, true positive, and specificity, true negative, can be used to determine how well a marker predicts an outcome and decide upon an appropriate threshold. A P-value is given, which indicated if a variable can correctly discriminate between the two populations,

a P<0.05 was considered significant. The area under the curve (AUC) is equal to the probability of patients being correctly identified, the aim is to achieve an AUC as close to 1 as possible, which indicates a perfect predictor. The R package *pROC*[160] was used to perform the calculation, this was implemented to determine if a specific gene or genes can discriminate between Event free survival and relapse patients.

2.2.6. Binary logistic regression

Determines the probability of a binary event occurring, based on the independent variable or variables given[161]. The P value indicates if the variable is a good predictor, P<0.05 was considered significant. The Beta correlation coefficient describes the trend of the relationship: the further from 0 the stronger the association, a positive coefficient indicates an increase in the variable, increases the hazard and decreases time to event, a negative coefficient indicates a decrease in the variable, decreases the hazard and increases time to event. The calculations were computed using SPSS, the binary event used was the relapse status, where 1 represents relapse and 0 indicates event free survival, the independent variable was the gene expression value or values.

2.2.7. Cox regression

Cox regression proportional hazards model[162] is used to associate the time passed before an event to the independent variable or variables in the model. A variable which predicted outcome with a P-value of P<0.05 was considered statistically significant. The Beta correlation coefficient describes the trend of the relationship: the further from 0 the stronger the association, a positive coefficient indicates an increase in the variable, increases the hazard and decreases time to event, a negative coefficient indicates a decrease in the variable, decreases the hazard and increases time to event. The calculations were computed using SPSS, the binary event used was relapse status, where 1 represents relapse and 0 indicates event free survival, the independent variable was the gene expression value or values and the time variable was the event free survival time.

2.3. Type one error correction

A type one error is the false rejection of the null hypothesis, using a standard P-value of P<0.05 will give a false positive 5% of the time. Within large datasets, such as those being used in this investigation, 5% of the "significant" genes equates to a large number. To combat the false discoveries, type one error correction is required.

The method implemented in this investigation to reduce type one error, was the Benjamini & Hochberg[163] method otherwise known as false discovery rate (FDR) which is neither too stringent nor too lenient. FDR reduces the Type one error without introducing too many type two errors, the false acceptance of the null hypothesis. To obtain corrected p-values the r function "*p.adjust*" was used.

2.4. Biomarker discovery

To develop a robust method of biomarker discovery, several workflows were trialled (figure 2.1), which combine statistical and machine learning approaches. After the initial variable workflows, shown in colour in figure 2.1, the gene lists obtained were used in independent multivariate binary logistic regression, to identify genes that work synergistically to predict patient outcome.

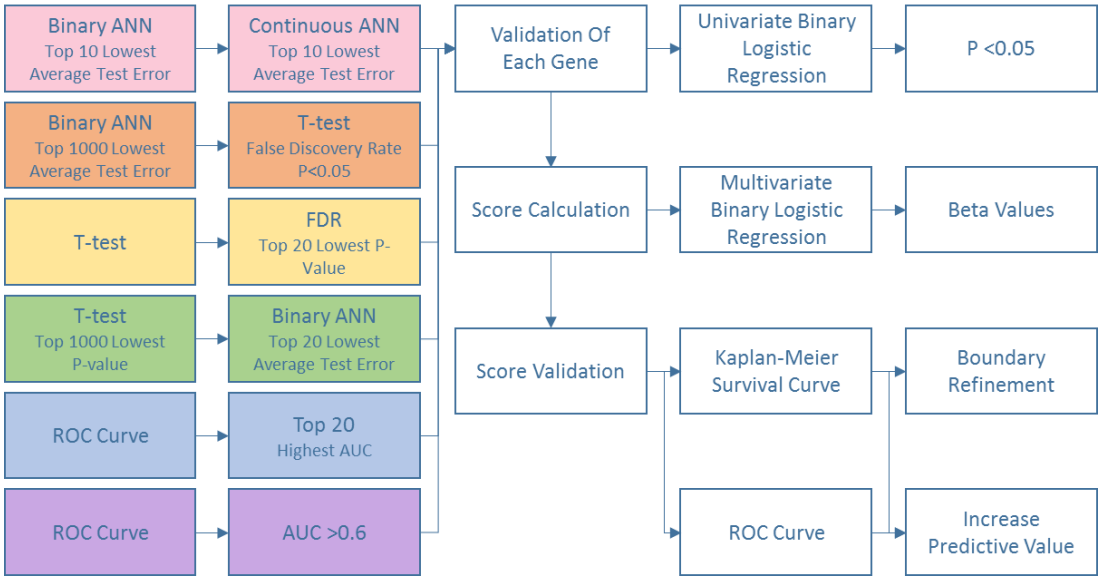

FIGURE 2.1. BIOMARKER DISCOVERY WORKFLOWS. A COMBINATION OF STATISTICS AND MACHINE LEARNING USED IN THE INITIAL BIOMARKER DISCOVERY WORKFLOW LEADS INTO A UNIFORM SCORE CALCULATION AND VALIDATION PROCEDURE, ENSURING DIRECTLY COMPARABLE SCORES. INITIAL BIOMARKER DISCOVERY IS PERFORMED IN THE HOVON DATASET WHICH IS FOLLOWED BY VALIDATION IN THE TCGA AND TARGET DATASETS.

The gene expression values used to calculate the overall score were obtained from the finalised gene list of each workflow, the β values were obtained from the multivariate binary logistic regression of the finalised gene list for each workflow. The number of elements within the score calculation can be increased or decreased according to the number of genes obtained for each method. To calculate the score the following equation was used[164], where Σ indicates the sum of, and β values were obtained though binary logistic regression:

$$\sum(gene\ expression\ value \times \beta\ value\ of\ the\ gene)$$

2.4.1. ANN

The ANN workflow is indicated in pink in figure 2.1. Using the ANN algorithm explained previously, relapse status was used as the initial binary input where 1 indicates relapse and 0 indicates event free survival. The top 10 genes by lowest average test error, from the initial binary ANN, were each used as continuous inputs for the next ANN analysis. The top 10 genes by lowest average test error from each of the continuous ANN were taken resulting in 110 genes in total, not accounting for duplicates in the results.

2.4.2. ANN-first

The ANN-first workflow is indicated in orange in figure 2.1. Using the ANN algorithm explained previously, relapse status was used as the initial binary input where 1 indicates relapse and 0 indicates event free survival. The top 1000 genes by lowest average test error were taken, and a T-test was performed, where patients were grouped by their relapse status. After a false discovery P-value correction was used, up to 20 genes with the lowest P-value were selected.

2.4.3. T-test

The T-test workflow is indicated in yellow in figure 2.1. A T-test was performed where patients were grouped by their relapse status, event free survival versus relapse. After a false discovery P-value correction was used, the top 20 genes by lowest P-value were selected.

2.4.4. ANN-Second

The ANN-second workflow is indicated in green in figure 2.1. A T-test was performed where patients were grouped by their relapse status, event free survival versus relapse. The top 1000 genes by lowest P-value were taken as a truncated dataset. Using the ANN algorithm

explained previously, relapse status was used as a binary input where 1 indicates relapse and 0 indicates event free survival. The top 20 genes by lowest average test error were selected.

2.4.5. ROC

The ROC workflow is indicated in blue in figure 2.1. The R package *pROC*[160] was used to determine the P-value and the AUC for each gene in the dataset, to identify genes which can discriminate between Event free survival and relapse patients. The genes were sorted by AUC and the top 20 highest AUC were selected.

2.4.6. BLR

The binary logistic regression workflow is indicated in purple in figure 2.1. The R package *pROC*[160] was used to determine the P-value and the AUC for each gene in the dataset, to identify genes which can discriminate between Event free survival and relapse patients. The genes were sorted by AUC, all the genes with an AUC >0.6 were selected. A binary logistic regression was performed on these genes, which were then sorted by beta-value. The genes with the 10 highest and 10 lowest beta-values were selected.

2.5. Tissue culture

2.5.1. Adherent cell lines

Cells were grown at 37°C in a humidified atmosphere with 95% air and 5 % CO_2. Cells were passaged when 80-90% confluent, before the loss of the cell monolayer, the time between passages was dependent on the specific cell line. Medium was discarded, and the flask was washed with PBS to remove any residual medium. A solution of 0.05 % trypsin mixed with 0.02 % EDTA was added to the cells, the smallest volume required to cover surface of the flask was used. Flasks were incubated at 37°C until cells had completely detached from the surface and were suspended individually. The cell suspension was removed to a falcon tube and complete growth medium was added, this was then centrifuged at 300xg for 5 minutes at room temperature. The supernatant was discarded, and cell pellets were resuspended in an appropriate volume of complete growth medium before cells were counted and seeded at the required density in a plate or flask.

Cell line	Growth requirements
HEK293	EMDM + 10% FCS + 1% L-Glutamine
MDA-MB-468	EMDM + 10% FCS + 1% L-Glutamine

TABLE 2.2. GROWTH MEDIA REQUIREMENTS OF ADHERENT CELL LINES.

2.5.2. Suspension cell lines

Cells were grown at 37°C in a humidified atmosphere with 95% air and 5 % CO_2. The cell suspension was removed to a falcon tube when the growth medium was spent or when the flask was confluent. The cell suspension was counted, as described below, before the falcon tube centrifuged at 300xg for 5 minutes at room temperature. The supernatant was discarded, and cell pellet was resuspended in an appropriate volume of complete growth medium for seeding at the required density in a plate or flask.

Cell line	Growth requirements
KG-1	IMDM + 20% FCS + 1% L-Glutamine
THP-1	RPMI-1640 +10% FCS + 1% L-Glutamine
MolM-13	RPMI-1640 +10% FCS + 1% L-Glutamine
MV4-11	IMDM + 10% FCS + 1% L-Glutamine
Kasumi-1	RPMI-1640 + 20% FCS + 1% L-Glutamine
SigM5	IMDM + 20% FCS + 1% L-Glutamine

TABLE 2.3. GROWTH MEDIA REQUIREMENTS OF SUSPENSION CELL LINES.

2.5.3. Cryopreservation and recovery

Cells were collected using the method outlined for passaging cells. Cell pellets were resuspended in freezing medium (FCS containing 10% DMSO) and aliquoted into 1mL cryovials, which were stored in a -80°C freezer.

To recover cells from storage, they were defrosted rapidly before medium was added dropwise. Cells were centrifuged at 300xg for 5 minutes, the supernatant was discarded, and pellet was resuspended medium. The cells were seeded in an appropriate plate or flask, which was then incubated according to the growth requirements of the cell line.

2.5.4. Cell counting

2.5.4.1.Trypan blue

Cell suspension was harvested using the method outlined for passaging cells. 50µL of cell suspension was mixed with 50µL of trypan blue. To count the cells, 10µL of trypan cell suspension was place onto a haemocytometer, each of the four corner squares was counted. Using the following calculation, the cell density was calculated.

$$Cell\ density\ = \frac{Average\ Number\ of\ Cells\ \times\ 10000}{Dilution\ Factor}$$

2.5.4.2.Solution-18

Cell suspension was harvested using the method outlined for passaging cells. 50µL of cell suspension was mixed with 2.5µL of solution-18. To count the cells, 10µL was place into one chamber of the A8- slide, which was then read using the NucleoView™ software on the NucleoCounter® NC-250™ machine

2.5.5. Drug titration

Cell lines were seeded at a density between $0.4x_{10}^{6}$ and $1.0x_{10}^{6}$ depending on the assay use, in a 6-well plate for cell counting, or a 96-well plate for an XTT assay. The drug was diluted to a working stock concentration, which was then used to treat the cells at a range of concentrations. The cells were counted daily using solution-18 as previously mentioned, or daily using the XTT assay. This was done for Puromycin, Daunorubicin and TGFβ1.

2.6. Reverse Transcription- Quantitative PCR

2.6.1. RNA extraction

RNA was extracted from cells using the Quiagen RNAeasy Kit, following the manufactures instruction. Buffer RLT was prepared by adding 10 µl of β-mercaptoethanol to 1 ml Buffer RLT, improving the quality of the extracted RNA. After cells had been harvested and buffer RLT had been added, following the appropriate methods detailed below, the lysate suspension was passed through a 21-gage needle to complete the lysis step.

To the cell lysate 350 µl of 70% ethanol was added and mixed by pipetting. The sample was transferred onto the spin column and centrifuged for 15 seconds at 8000xg, the flow-through was discarded. 700 µl of Buffer RW1 was added and the column was centrifuged as before,

after each centrifugation the flow-through was discarded. 500 µl of buffer RPE was added and the column centrifuged as before, this step was repeated however the centrifugation was extended to 2 minutes. The column was placed in a new collection tube and centrifuged at 14000xg for 1 minute, to completely remove the buffers. The column was placed in 1.5 ml microfuge tube, 30 µl of NFW was added to the centre of the spin-column filter, incubated at room temperature for 10 minutes and then centrifuged for 1 minute at 8000xg. The samples were read using the nanodrop and stored at -80°C until needed.

2.6.1.1. Adherent cells

Medium was removed from adherent cells and the flask was washed with PBS, all liquid was removed before 350 µl of pre-prepared buffer RLT was added directly to the plate. A plate scraper was used to collect the cells and buffer together for harvesting.

2.6.1.2. Non-adherent cells

For non-adherent cells, the cell suspension was centrifuged at 300xg for 5 minutes, the supernatant was removed, the cells were resuspended in PBS and centrifugation was repeated. After the supernatant was removed, 350 µl pre-prepared buffer RLT was added to the pellet.

2.6.2. Nanodrop

To measure the concentration and assess the quality of the extracted RNA or DNA, the NanoDrop™ 8000 spectrophotometer was used. Prior to use the pedestals were cleaned and calibrated using NFW. To measure the sample, 1 µL was placed on the pedestal, the appropriate nucleic acid setting was selected for RNA or DNA. The 280 and 260 ratios were expected to be 2.0, deviation from 2.0 indicated contamination with impurities such as protein and phenols. The upper limit of detection was 4000 ng/µL, if the concentration exceeded this, an aliquot was diluted in NFW and re-analysed.

2.6.3. cDNA synthesis

Depending on the quantity and quality of the RNA available determines which method of reverse transcription is used. SuperScript™ IV requires a lower concentration of RNA compared to MMLV, so is ideal for patient samples, whereas MMLV is sufficient for RNA extracted from cell pellets.

2.6.3.1.SuperScript™ IV Reverse Transcriptase

250ng of RNA is required, the volume is determined by the following calculation:

$$\frac{250ng}{concentration\ (ng/\mu l)}$$

The appropriate volume of RNA is mixed with 1 µL of 50 µM Oligo $d(T)_{15}$ and 10mM dNTP, NFW was used to make the volume up to 13 µL. This mixture was incubated at 65°C for 5 minutes, then immediately transferred to ice for 1 minute. To the RNA mixture 4 µL of SSIV buffer, 1 µL SSIV reverse transcriptase, 1 µL RNase inhibitor and 1 µL 100mM DTT was added. The reaction mixture was incubated at 53°C for 10 minutes then 80°C for 10 minutes. The cDNA samples were stored at -20°C until use.

2.6.3.2.MMLV reverse transcriptase

2000ng of RNA is required, the volume is determined by the following calculation:

$$\frac{2000ng}{concentration\ (ng/\mu l)}$$

The appropriate volume of RNA is mixed with 1 µL of 50 µM Oligo $d(T)_{15}$, NFW was used to make the volume up to 10 µL. This mixture was incubated at 70°C for 5 minutes, then immediately transferred to ice for 1 minute.

To the RNA mixture 5 µL of RT buffer, 1 µL MMLV reverse transcriptase, 0.7 µL RNase inhibitor 1 µL of 10mM dNTP, and 7.3 µL of NFW was added. The reaction mixture was incubated at 40°C for 1 hour then 95°C for 5 minutes. The cDNA samples were stored at -20°C until use.

2.6.4. Primer design

Primers for genes of interest available from Sigma were used where possible, the pairs with the fewest off-target interactions were used. For genes with no available pre-made primers, the following requirements were met; the sequence should end with 1-2 G/C pairs, 18-25 base pairs long, a GC content of 60-80%, melting temperature of 61.5°C and a concentration of 0.05µM. Primer specificity was determined using NCBI primer blast search, returning corresponding amplicons between 100 and 150 base pairs, ensuring high efficiency. Reference sequence for each of the primers used can be found in the relevant reagents and equipment table

The purchased primers were centrifuged to collect all particles at the bottom of the tube, this was resuspended in the specified volume of nuclease free water (NFW) and incubated on ice to fully dissolve for 10 minutes. The stock was vortexed and diluted 1:10 with NFW to make the working dilution.

2.6.5. Primer efficiency

Extracted cDNA from a cell line was diluted in a 1:2 and 1:5 dilution series. SYBR green PCR master-mix was used, the total volume of this was half of the reaction volume, the other half of the reaction volume was made up of 0.5µL of both forwards and reverse primers, NFW to the volume required and 1 µL of sample. Each sample in the dilution series was performed in triplicate, all reagents were vortexed prior to use and kept on ice. Primers for the gene of interest and the stable "house-keeping" gene are tested in the same way, preferably both primer pairs will achieve efficiency at the same temperature. The standard temperature profile used began with a 95°C hold for 1 minute, followed by a cycle of 95°C for 20 seconds, a variable temperature depending on the primer for 15 seconds and 72°C for 20 seconds. To obtain a melt curve the temperature ramped from the lowest annealing temperature through to 95°C.

2.6.6. Quantitative PCR

After the correct temperature for the primer has been established using primer efficiency testing, a similar method is employed to identify the relative gene expression of a sample. The volumes of reagents were consistent with those in the primer efficiency testing; however, samples were not diluted. Both the gene of interest and the stable "house-keeping" gene were run for each sample. The temperature profile used began with a 95°C hold for 1 minute, followed by a cycle of 95°C for 20 seconds, 58°C for 15 seconds and 72°C for 20 seconds. To obtain a melt curve the temperature ramped from 58°C through to 95°C.

2.7. Plasmid transfection in bacteria

2.7.1. Bacterial culture

2.7.1.1.Ampicillin

The bacterial strain used is sensitive to ampicillin, to select the bacteria which were successfully transfected ampicillin selection was used. Ampicillin salt was dissolved in double

distilled (DD) H_2O, to obtain a concentration of 100mg/mL, this was filter sterilised using 0.2-micron filter. Aliquots were frozen at -20°C.

2.7.1.2. Lauria Broth Agar

Lauria Broth (LB) Agar plates were made according to the manufacturer's instructions. LB agar was dissolved in DD H_2O, this was autoclaved to sterilise. When the molten agar had cooled to below 40°C 100mg/mL of ampicillin was added. 15ml of agar was used per plate

2.7.1.3. Lauria Broth

Lauria Broth (LB) was made according to the manufacturer's instructions. LB was dissolved in DD H_2O, this was autoclaved to sterilise. When the broth had cooled 100mg/mL of ampicillin was added.

2.7.1.4. Bacterial stocks

A single colony from the plasmid transfected *E.coli* LB agar plates was taken and grown overnight in LB containing 100mg/mL of ampicillin. A glycerol solution of 30% was made using DD H_2O which was filter sterilised using 0.2-micron filter. From the overnight cultures, 250µL was mixed with 250µL of 30% glycerol, the final concentration of glycerol was 15%, these were snap frozen on dry ice and stored at -80°C.

2.7.2. Bacterial transformation

Stbl3 strain E.coli competent cells were used, which can stabilise supercoiled DNA, making them ideal for transfection with the shRNA plasmid. This strain of E.coli is sensitive to ampicillin, therefore the shRNA plasmid used gives the transfected cells resistance. Chemically competent cells were thawed on ice and 2µl of plasmid DNA was added to the vial, sterile water was used as a control to ensure the bacteria were not resistant to ampicillin, and there was no contamination. Reference sequence for each of the plasmids used can be found in the relevant reagents and equipment table. After a 30-minute incubation on ice, the aliquots were heated to 42°C for 45 seconds before returning to ice for 2 minutes. S.O.C. medium was added to the cells and incubated at 37°C for 1 hour. 100µL was plated on LB agar with ampicillin, and incubated overnight at 37°C. From the experimental plates a single colony is selected and cultured in 25mL of LB broth with ampicillin and incubated overnight at 37°C.

2.7.3. Plasmid extraction

The Plasmid Midi Kit from Qiagen was used, following the manufactures instructions. Bacteria containing the transfected plasmid were cultured overnight in LB with ampicillin as describes above. The bacterial culture was centrifuged at 5000 rpm for 15 minutes at 4°C, the supernatant was discarded, and the pellet was resuspended in 4 mL of chilled Buffer P1. To the bacterial suspension, 4 mL of buffer P2 was added and vigorously mixed. After a 5-minute incubation at room temperature, 4 mL of chilled Buffer P3 was added, mixed vigorously. The bacterial lysate was poured into the QIAfilter Cartridge and incubated at room temperature for 10 minutes. The QIAGEN-tip 100 was prepared by adding 4 ml of Buffer QBT which was allowed to flow through the tip. The plunger was inserted into the QIAfilter Cartridge to expel the lysate into the QIAGEN-tip 100. Using gravity flow the solution entered the resin and the flow-through was discarded. The resin and sample were washed twice by adding 10 mL buffer QC to the QIAGEN-tip 100. Buffer QF was heated to 65°C, 5 mL was added to the QIAGEN-tip 100 and the flow-through was collected in a sterile falcon tube. To the flow-through, 3.5 mL isopropanol was added, this was mixed and centrifuged at 5000 rpm for 30 minutes at 4°C. The supernatant was discarded, and the pellet was resuspended in 1mL of 70% ethanol, which was centrifuged at 13,000rpm for 10 minutes at 4˚C. The supernatant was discarded, any remaining liquid was allowed to evaporate from the pellet at room temperature. The dry pellet was resuspended in 50µL of TE buffer, before quantification on the nanodrop.

2.8. Viroid construction

2.8.1. Lipofectamine transfection

HEK293 cells are ideal for viroid construction as they contain the gene for adenovirus. HEK239 cells were seeded in t25 flasks and allowed to adhere and grow to 70% confluence. To the flask 8µg of plasmid of interest, 6µg of packaging plasmid and 2µg of envelop plasmid were added to the flasks with 17µl lipofectamine and 32µl p3000 reagent, in FCS free optiMEM medium. The flasks were incubated at 37°C with 5% CO2 for 24 hours.

2.8.2. Virus harvest

Twenty-four hours after HEK239 cells were treated with lipofectamine p3000, the medium was changed to remove the reagents used in transformation. Forty-eight and seventy-two hours after cells were treated the medium was harvested, these fractions should contain the

constructed virus. The harvested medium was centrifuged at 300xg for 5 minutes, and filtered using a 0.2-micron filter, removing any remaining cells preventing contamination of cell lines. The fractions were frozen at -80°C until required.

2.9. Viral Transduction

2.9.1. Transduction of adherent cell lines

Cells were seeded and allowed to adhere and grow to be 70-80% confluent. Medium was removed and cells were treated with Polybrene diluted in medium, at a concentration of 10µg/µL. After a short incubation, the viral stock was added to the test wells and medium was added to the control wells.

After 24 hours, spent medium was removed and replaced with fresh medium containing Puromycin at a concentration determined by a drug titration. Puromycin will kill cells not containing the transfected plasmid within a few days whereas transfected cells will continue to grow in culture. When control cells are dead due to the Puromycin treatment, the transfected cells are grown until confluent.

2.9.2. Transduction of suspension cell lines

Due to suspension cells experiencing poor transfection efficiency or sensitive to polybrene toxicity, an alternative method of transfection was used. Cells in suspension were centrifuged at 300xg for 5 minutes, spent medium was discarded and cells were resuspended in 1mL of fresh medium containing polybrene at a concentration of 10µg/µL. After a short incubation with the polybrene medium, the viral stock was added to the test cells and medium was added to the control cells. The falcon tubes containing the cell suspensions were centrifuged at 800xg for 30 minutes at 32°C. After the centrifugation the supernatant was removed, and cells resuspended in 3mL of fresh medium, to be seeded in a 6-well plate.

After 24 hours Puromycin was added to the wells at a concentration specific to the cell line as determined by a drug titration. The Puromycin will kill cells not containing the transfected plasmid within a few days, whereas those that were successfully transfected will continue to grow in culture. When control cells are dead due to the Puromycin treatment, the transfected cells are grown until confluent. These cell lines are stably transfected and can continue to be cultured in the appropriate medium as described previously with the addition of Puromycin at a concentration determined by the drug titration. Stocks can be frozen and thawed according to the methods previously mentioned.

2.9.3. Testing viral transfection

It is important to test for the presence of the virus in the harvested fractions from the HEK239 cells. MDA-MB-468 cells were used in the method outlined for adherent cells above. The concentration of Puromycin used was 1µg/mL. After 2 days when control cells were dead due to the Puromycin treatment, the transfected cells were grown until confluent and there was a sufficient quantity for RNA extraction and cDNA is synthesis, which were performed according to the previously described methods.

2.9.4. Knock-down efficiency

Once the cells have been transfected and are numerous enough to harvest, the RNA is extracted, and cDNA is synthesised according to the previously described methods. For both the gene of interest (GOI) and a stable housekeeper gene qPCR was performed on all samples: cells treated with the control plasmid or the experimental plasmids. The average cycle threshold (CT) value for replicates was calculated for the housekeeper and GOI. The change in the CT values (ΔCT) from GOI and housekeeper was calculated, and the change of the change in CT values (ΔΔCT) were calculated using the following equations:

$$\Delta CT = Housekeeper\ CT\ - GOI\ CT$$

$$\Delta\Delta CT = \Delta CT\ Experimental\ Plasmid\ - \Delta CT\ Control\ Plasmid$$

Using information provided by ThermoFisher[165] on percentage Knockdown of gene expression, the approximate knockdown efficiency can be determined using the calculated ΔΔCT value where Δ represents the change.

2.10. XTT cell proliferation and viability assay

Following gene Knock-down, cell lines were seeded at $1x10^6$/mL and treated with the predetermined concentration of Daunorubicin and TGFβ, wells containing no cells were also plated with the corresponding drug titrations. The plates were incubated for 24 hours.

After the incubation, the plates were mixed using a 1mL pipette. 100µL of each cell suspension and cell-free medium control was transferred to a 96-well plate. A stock of XTT working reagent was made to a concentration of 20µL/mL of electron coupling reagent in the XTT Labelling reagent, which was then vortexed. 50 µL of the XTT working reagent was added to the wells of the 96-well plate, which were mixed by pipetting and then incubated

for 4 hours. After the incubation the plates were mixed by tapping before they were read at 470 nm with a reference wavelength of 650 nm.

2.11. NanoString

Cell line samples were harvested and prepared according to the RNA extraction protocol above. The RNA was measured with the nanodrop in triplicate to obtain an accurate concentration. 150ng of RNA was pipetted into the strip tubes and water was added to equalise the volume to 5µL. The reporter probe and capture probes were added to the strip tubes according to the manufacturer's instructions. After mixing by tapping the tubes, they were centrifuged to collect the mixture in the bottom of the tubes which were then incubated at 65°C for 24 hours. After the incubation the strip tube was centrifuged briefly to collect the mixture in the bottom of the tube and loaded into the prep-station, the reagent plates were centrifuged at 2000xg for 2 minutes and loaded into the prep-station with the required consumables, as per the manufacturer's instructions. Once the prep station had finished the chip containing the samples was sealed and loaded onto the nCounter. The data from the nCounter was analysed using the nSolver software.

2.12. Reagents and equipment

2.12.1. Cell Culture Media and supplements

Reagent	Supplier
EMDM	SLS (Lonza)
Foetal Calf Serum (FCS)	GE Healthcare Hyclone
IMDM	SLS (Lonza)
L-Glutamine	SLS (Lonza)
OptiMEM media	Gibco
RPMI-1640	SLS (Lonza)

2.12.2. Cell Culture Reagents

Reagent	Supplier
Cytarabine	Sigma Aldrich
Dimethyl sulfoxide (DMSO)	Santa Cruz Biotechnology
Dulbecco's phosphate buffered saline (DPBS)	SLS (Lonza)
Puromycin	GIBCO
Solution-18	Chemometric
Transforming Growth Factor β1 (TGFβ)	Peprotech
Trypan Blue solution 0.4%	Sigma-Aldrich
Trypsin and EDTA	SLS (Lonza)

2.12.3. Bacterial Culture Reagents

Reagent	Supplier
Ampicillin sodium salt	Sigma-Aldrich
Glycerol	Sigma-Aldrich
Luria broth (millers)	Sigma-Aldrich
Luria broth (millers) Agar	Sigma-Aldrich

2.12.4. Kits

Reagent	Supplier
Cell Proliferation Kit II (XTT)	Roche
Competent cell sampler	Invitrogen
HiSpeed Plasmid Midi Kit	Qiagen
RNeasy Mini Kit	Qiagen

2.12.5. Chemical Reagents

Reagent	Supplier
Deoxyribonucleotide triphosphate (dNTP)	Promega
Double distilled water (ddH2O)	Barnstead
Ethanol absolute Electran® molecular biology	VWR Chemicals
Isopropanol	fisher Chemical
lipofectamine P3000	Thermo Scientific
Liquid nitrogen	BOC
MMLV Reverse Transcriptase	Promega
MMLV RT 5x Buffer	Promega
Nuclease-free water	Ambion
Oligo(dT)15 Primer	Promega
Polybrene	Sigma-Aldrich
Presept tablets	Johnson and Johnson
RNaseZAP	Ambion
RNasin	Promega
SSIV buffer	Invitrogen
SuperScript™ IV Reverse Transcriptase	Invitrogen
SYBR® Green	BioRad
β-mercaptoethanol	Sigma-Aldrich
Protein assay dye reagent concentrate	BioRad

2.12.6. Plastics, Glassware and Sharps

Reagent	Supplier
A8-slides	Chemometric
Bijou tubes (7 mL)	Starlab
Cell culture flasks (T25, T75, T175)	Sarstedt
Cell scraper	Sarstedt
Clear flat bottom 6-well plate, sterile	Sarstedt
Clear flat bottom 96-well plate	Starlab
Cryogenic vials (1.0 mL)	Starlab
Falcon tubes (15 mL, 50 mL)	Sarstedt
Filter tips (10ul, 20ul, 100ul, 200ul, 1000ul)	Starlab
Glass bottles	Duran
Micro tubes (0.5 mL, 1.5 mL, 2.0 mL)	Sarstedt
Plate spreaders	Sarstedt
Rotor-Gene Strip Tubes & Caps	Starlab
Serological pipettes (5 mL, 10 mL, 25 mL)	Sarstedt
Syringe filter 0.2µm	Sartorius
Syringe filter 0.4µm	Sartorius
Syringes (20 mL)	Medicina

2.12.7. Equipment

Reagent	Supplier
Autoclave	Rodwel
Automated Cell Analyser NucleoCounter® NC-250™	Chemometric

Axio Observer.Z1 microscope	ZEISS
Benchtop vortex mixer	Scientific Industries
Centrifuge 5804R	Eppendorf
Centrifuge 5810R	Eppendorf
Class II Safety Cabinet	Walker
Digital Sight DS-Fi1 camera	Nikon
Freezer -20°C	LEC Medical
Freezer -80°C	Panasonic 2100
Fridge 4°C	LEC Medical
Haemocytometer	Weber
Incubator 37°C	GenLab
Incubator 37°C 5% CO2	Scientific Laboratory Supplies
Centrifuge, Heraeus Megafuge 16R	Thermo Scientific
Micropipettes (2 µl, 10 µl, 100 µl, 200 µl, 1000 µl)	Gilson/Starlab
Minispin benchtop centrifuge	Eppendorf
Multichannel pipette (300 µl)	Eppendorf
Nanodrop ND-8000 spectrophotometer	Thermo Scientific
Nanopure Diamond water reservoir	Barnstead
Nikon Eclipse Ts100 Light Microscope	Olympus
Orbital Incubator 37°C	Stuart
PCR workstation cabinet	Grant-Bio
Real-time qPCR thermal cycler	Qiagen
Thermoblock	Biometra

Waterbath	Clifton
Weighing Scale	Fisher Scientific
Microscope, EVOS M5000	invitrogen
iMark microplate Reader	Bio-rad

2.12.8. Software

Programme	Supplier
GraphPad Prism v7	GraphPad Software Inc.
Rotor-GeneQ Series Software v2.3.1	Qiagen
SPSS statistics 26	IBM Corp
R x64 3.6.0	The R Foundation for Statistical Computing
NucleoView™ software	Chemometric

2.12.9. shRNA Plasmid DNA

TRC number	Sequence
TRCN0000073652 MISSION shRNA	CCGGGCAGCCCTAATGAATACAGAACTCGAGTTCTGTATTCATTAGGGCTGCTTTTTG
TRCN0000073650 MISSION shRNA	CCGGCCCGGAGGAAATGTGACTATTCTCGAGAATAGTCACATTTCCTCCGGGTTTTTG
control Plasmid	addGene pLKO.1, Puromycin, empty vector
packaging plasmid	addGene psPAX2, PL-12260
envelop plasmid	addGene pMD2.G, PL-12259

2.12.10. Primers for qPCR

Gene	Sequence	Supplier
CD109 F	AGTATACATAGGGAAGCCAG	sigma-Aldrich
CD109 R	ATCCAGTATTCAGAAAGTCC	sigma-Aldrich
YWHAZ F	ACCGTTACTTGGCTGAGGTTGC	sigma-Aldrich
YWHAZ R	CCCAGTCTGATAGGATGTGTTGG	sigma-Aldrich

Chapter 3 - Biomarker discovery

3.1. Introduction

3.1.1. The need for clinically relevant biomarkers

There are a number of biomarkers already available in a clinical setting, these are broken down into seven different categories[120] shown in table 3.1. Clinical biomarkers are a biological molecule found in body fluids or tissues that are indicative of a disease state or condition[166]. Biomarkers can be proteins, peptides or nucleic acid based[167], although for a clinical setting, molecules should be easily obtained and analysed. Key molecules that are a focused for biomarker discovery are abnormal RNA expression, DNA mutations and abnormal protein expression, which are easily obtained molecules that can be processed from patient samples.

Type	Application	Example
Diagnostic	Detection of early disease and subtype	Prostate cancer: Prostate specific antigen
Prognostic	likelihood of an outcome or clinical event	Prostate cancer: Gleason score
Predictive	Prediction of effective treatment response	Brest cancer: HER2, ER, PR
Response	Identification of a response to therapy	Solid cancers: Tumour volume
Risk	Identifying those who may develop the disease	Cervical cancer: HPV
Monitoring	Assessing disease status	Solid cancers: Tumour volume and metastasis
Safety	Indicate adverse effects from therapy	All cancers: Neutropenia

TABLE 3.1. TYPES OF BIOMARKERS. SEVEN MAIN TYPES OF BIOMARKERS ARE AVAILABLE[120]. ALTHOUGH BIOMARKERS ARE AVAILABLE FOR MANY DIFFERENT DISEASES, THE EXAMPLES GIVEN ARE CANCER SPECIFIC.

Biomarkers are a vital tool for patient care, which are used to inform treatment options and achieve the best possible outcome, the more informed a clinician is the better care a patient will receive. An example of a current biomarker is t(15;17)[73] which is characteristic of APL, a very treatable form of AML[81], this is a large chromosomal translocation and a DNA base

marker. A patient with APL receives all-trans-retinoic acid in combination with Arsenic trioxide or chemotherapy if required[81], as opposed to chemotherapy as standard for AML. The different treatment of this subcategory of patients provides better care and overall better outcomes. As part of diagnosis, patient biopsies are taken from the affected tissues to determine if a patient has cancer and which type and stage it is[168]. If a patient is suspected of having breast cancer the biopsies are stained using an antibody cocktail for HER2, oestrogen receptor and progesterone receptor, all of which have targeted therapeutics available[169–174]. In a pathology report the presence or absence of these markers will be confirmed, if a patient doesn't have these markers it is described as triple negative breast cancer which has fewer treatment options and poorer prognosis including reduced overall survival and reduced relapse free survival[175–177].

The issue currently facing the biomarker sector is that, although many are proposed each year, very few are approved for use, in part to the fact they are not transparent in their discovery so are not necessarily trusted by the wider scientific community[178]. The need to communicate the methods use to a lay audience can be challenging when heavily computational methods are used, which can be a barrier to their approval[178,179]. Another barrier to approval is the lack of high-performance and high accuracy panels, which is especially important for markers that may impact patient care and the medication available.

In AML there are biomarkers available which indicate which subcategory of AML a patient has and how aggressive the cancer can be. Using the ELN[73] guidelines, these categories of AML are based on the large genetic mutations and translocations causing the cancer, an example of this t(15;17)[13] is an indicator for APL which has a separate treatment strategy to AML involving all-trans-retinoic acid[79] in conjunction with arsenic trioxide or chemotherapy. Another example is flt3-ITD[99], where tyrosine-kinase-inhibitors are included in the therapy to increase the chance of successful treatment[99,180,181]. Within AML a large proportion of patient's relapse, almost 50% which leads to further adverse outcomes including death. With such a large proportion of patients facing relapse it is important to identify these patients early so treatment can begin and a more positive outcomes can be achieved. As such the aim of this chapter is to develop a biomarker discovery workflow, based on the prediction of relapse in acute myeloid leukaemia. A panel of genes with high performance in the prediction of relapse will be developed, which will also be easily understood by a wider scientific audience.

3.1.2. Current methods of biomarker discovery

There are many different methods of biomarker discovery, many of these require specialist skills or knowledge to utilise properly. Here we discuss the different methods of biomarker discovery that already exist, and strategies for improvement. There are many machine learning approaches some are transparent in the logic used and easily understood, referred to as "white-box" methods[179], where some are heavily programmed and the results are not easily explained, referred to as "black-box" methods[179]. There has been a push from the scientific community to move away from black-box methods to increase the transparency when using the results in a clinical setting. Although both methods are valid approaches to biomarker discover white-box methods have the added benefit of transparency and reproducibility whereas black box methods can give slightly different results each time. Although there's a notion that heavily computational and complex methods are somehow better[178,179] this is not strictly true, and depends heavily on the data available, there is also overlap between the results of both.

In addition, these methods can be supervised, unsupervised or semi-supervised. Supervised machine learning relies on data which has both input data and an output classifying data, for example, RNA expression as input data and relapse status of patients as classifying output data. Unsupervised methods rely only on input data for the algorithm to cluster or associate the different samples. Semi-supervised methods rely on large amounts of input data, some with labelling data, it can be used to classify and label the samples without this information. For this body of work the focus will be on supervised methods which will be used to classify patients based on clinical data.

Artificial neural networks (ANN) are an example of a black-box method, as it is not transparent how the outcome is achieved as it is determined by the algorithm. In these instances, the algorithm "learns" from the input information and provides the output based on the algorithm detail. For biomarker discovery the input required is a dataset of the desired molecule for example, an RNA sequence dataset, and information used to label the samples or segregate the samples, such as relapse status or a specific gene of interest.

The use of "decision trees" are an example of a "white-box" method. They are based on sequential questions to target and define the differences between groups, which are transparent and logical. Once the algorithm is set up, it can be used repeatedly on different datasets and it is a very transparent process, however categorical data is required as an input which doesn't describe much of the data obtained in patient. This is another example of a

supervised method as the algorithm requires input from the user, in the form of the questions asked and the cut off when continuous data is used, the results obtained have the potential to be biased based on the human input. Another white-box method of biomarker discovery is using statistics to determine the association of a marker to the outcome, the calculations are published, and the method is fully reproducible. One way in which statistics can be used for biomarker discovery is looking how the gene expression is associated to the outcome of a patient through the means of t-test or binary logistic regression, although many other methods are available.

Although each method has advantages and disadvantages, each one can be used in different settings and may be more applicable in certain circumstances. The overall aim of this chapter is to develop a robust workflow of biomarker discovery that can be applied to patient outcome, which is understandable and logical, to be applied in a clinical setting. As such the workflows designed will be assessed through the development of a biomarker panel to predict relapse in AML patients. The best overall panel will be determined by the usefulness of the score created and could be used for future clinical biomarker discovery.

3.1.3. Dataset for biomarker discovery

To discover biomarkers, we require a large body of relevant biological data from a particular disease but are restricted by what pre-published bodies of work are available. To create a patient dataset large numbers of patients, money, and time are required, as well as informed consent, ethical approval, and a team of experts to extract, process and analyse patient samples. The creation of a patient dataset is not feasible during a PhD so we must look to already published, publicly available datasets, which allow us to utilise larger quantities of patient data making biomarkers discovery faster, easier, cheaper, and more reproducible. Data set repositories such as Array express[132], and gene expression omnibus[131] provide a platform to search datasets that have been published. Other sources, such as cBio-portal[182,183] provide an interface to quickly manipulate the data and determine some preliminary answers based on the datasets available, which is an excellent starting point to see the information held in the dataset and the possible applications before downloading the dataset for more thorough analysis.

There are several biological components that can be used for biomarker discovery including DNA, RNA, and protein. Although all these molecules can be excellent biomarkers, we are limited by the data which is available to us. When using publicly available datasets, DNA sequences are often restricted access and requires advanced permissions and ethical

approval from the human tissue authority (HTA), due to the fact a patient may be identifiable by their DNA sequence. Although DNA sequence datasets can be difficult to access, DNA microarray data can be easier to obtain as this technology does not identify the specific sequence, but the mutations in the context of specific genes. The advantage of using DNA sequencing is that it provides large amounts of data, but it can be expensive and requires experts to analyse fully. Array technology provides a lower quality data as it is predetermined which genetic markers are detected, therefor it is less informative, but it is cheaper and easier to analyse. Key information obtained from DNA analysis is predominantly mutation information present in a sample, whether that is translocations, duplications, fusion genes, internal tandem repeats, insertions, deletions or point mutations, all can directly impact cell functions.

There are very few protein datasets as these require mass spectrometry for large scale analysis, or a panel of antibodies to obtain expression for specific proteins, which reduces the quantity of data available. Typically, the tests carried out in a clinic would be antibody base, looking as specific targets known to determine patient care options, such as HER2 for targeted breast cancer treatment. Mass-spectrometry allows for large amounts of data to be obtained but relies on the researcher to prepare the sample appropriately and can be dependent on the machine used, although if mass-spectrometry data is available this provides a wealth of knowledge about the disease state and what functions and processes are happening within a sample.

Publicly available RNA datasets are readily obtainable, historic datasets often use microarray data, where newer datasets use RNA sequencing, or more recently NanoString. Patients can't be identified through RNA expression, which is ideal for ethical approval although it is still a requirement for the initial study and to obtain and process patient samples. RNA transcripts are associated with a cells function, cell cycle stage and disease state, this is ideal for showing dysregulation in a disease. Although microarray data is readily available it can be difficult to process as there are multiple probes for the same gene, whereas RNA sequencing aligns the fragments of RNA to determine a frequency of RNA expression for a particular gene.

The availability of patient information associated with a dataset can make finding a relevant biomarker difficult, due to the lack of annotations or poorly annotated datasets. If a specific disease state or outcome is being examined, the analysis requires some annotation, such as survival, relapse, or Gleason score to allow for grouping of patients and identification of more relevant markers. It may also be relevant to have disease free samples, this allow for disease

specific biomarkers to quickly determine if a person has a disease or if they are at risk from a disease, just as how HPV is used to separate patients at high and low risk of cervical cancer, where those who are negative are excluded from further testing.

Although DNA mutations are important and can severely impact certain diseases, small mutations can be irrelevant due to the redundancy in the genetic code, for example, silent point mutations can substitute a base changing the codon sequence but does not change the corresponding amino acid. In addition, mutations can occur in introns and untranslated regions which are thought to have no impact on transcription and translation, although recent research has suggested these small mutations can have an impact on oncogenes[184]. As small DNA mutations are not always translated into RNA and protein, they may not impact the disease state; large mutations can show where a disease has begun, for example t(15;17) is a common mutation which is present in PML (pre myelocytic leukaemia) and is a key biomarker for its diagnosis. Other large mutations in the DNA may inform treatment options such as flt-3 ITR which confers poor prognosis and require a tyrosine-kinase inhibitor as part of treatment, or the t(15;17) mutation which is treated with retinoic acid. There is large person to person variation in DNA sequencing which may confound the results obtained through NGS and can interrupt the sequence binding to the probes in Array technology, but again, large mutations are a valid source for determining disease state and treatment options.

looking at the datasets available quickly determines which molecule can be used to develop a biomarker panel; The overwhelming majority of the datasets are RNA based which shows the disease state within a sample, and allows for validation within a separate datasets, compared to protein datasets, which are infrequent, show the disease state of a sample but due to the infrequency cannot be validated easily, or DNA which is difficult to obtain but would be informative of key mutations driving the disease.

3.2. Methods

3.2.1. Obtaining Datasets

Dataset repositories, Array express and gene expression omnibus, were searched to Identify useful datasets, within certain criteria. The molecule chosen for the basis of the biomarker panel was RNA due to the readily available data in comparison to DNA and protein. The datasets chosen should have over 40 patients to allow for separation of groups within the dataset, overall survival and relapse free survival should be available, and of course the dataset should comprise of human samples. In addition to these criteria, the initial search

phrase was chosen carefully to encompass all the different titles that could be used when identifying AML, although most search functions were excellent in identifying alternative names itself. Several phrases were used to get the breath of datasets, including “AML”, “Acute Myeloid leukaemia”, “Acute Myelogenous Leukaemia” and the alternative American spelling of “Leukemia” were used in turn.

Overall, three datasets were identified with sufficient data to use for discovery and validation of a relapse signature score. The fist dataset, which was used for biomarker discovery, was the HOVON[148] dataset which is an Affymetrix array dataset. The CEL files for the HOVON[148] datasets were downloaded from array express[149] and were merged using the “ReadAffy” function from the R package *affy*[150]. The dataset was quality checked using the *ArrayQualityMetrics* package to identify outliers, looking at the Relative Log Expression (RLE) and the Normalized Unscaled Standard Error (NUSE). Once the outliers were removed from the dataset, it was RMA normalised using the *affy*[150] package. The “collapse rows” function from the R package *WGCNA*[151] was used to select one representative Probe-ID for each gene, the settings used select the probe-ID with the fewest missing values, and secondarily the Probe-ID with the largest mean value to resolve any tied probes.

The TCGA[83] dataset was downloaded using the R package *TCGAWorkflowData*[153], all cases were downloaded from the repository as FPKM data. TARGET-AML[154] was downloaded from the TARGET site in the RPKM data format. The datasets were quality checked using the *arrayQualityMetrics*[151] package, looking at the distance between arrays, any outliers were removed.

To ensure the biomarkers discovered were specific to relapse, certain filters were applied to all the cohorts. Only diagnostic samples were used, all other time points were removed from the dataset. only event free survival and relapse patients were used, patients who experienced all other events were removed, an additional check was applied to the EFS patients, to remove patients who had been annotated as event free but subsequently also described as dead, presumably from other causes but could not be a certainty with the information provided. Other filters were also applied to improve the quality of the score developed. All patients with FAB-M3 and t(15;17) translocation were removed, due to the difference in treatment, subsequently If the FAB or ELN classification were missing the sample was removed.

3.2.2. Bioinformatics techniques

3.2.2.1. Artificial neural networks

Binary ANN was used by classifying cases based on their relapse status, event free survival (0) or relapse (1), the output was sorted by average test error. The genes identified either from the binary ANN or another statistical test were identified and the expression values were used as the input for the continuous ANN, the output was sorted by average test error.

3.2.2.2. Welch's T-test

The "t.test" function in R was used to calculated P-values, this test was used to determine if the mean gene expression of relapse patients significantly differed from event free survival patients. A $P<0.05$ is considered statistically significant, indicating the means of the population are significantly different

3.2.2.3. Receiver operator characteristic curves

The R package *pROC*[160] was used to perform the calculations for the ROC curves, this was implemented to determine if a specific gene or genes can discriminate between Event free survival and relapse patients. A $P<0.05$ was considered significant. The area under the curve (AUC) is equal to the probability of patients being correctly identified.

3.2.2.4. Binary logistic regression

The calculations were computed using SPSS, the binary event used was the relapse status, where 1 represents relapse and 0 indicates event free survival, the independent variable was the gene expression value or values. A $P<0.05$ was considered significant. The Beta correlation coefficient describes the trend of the relationship, these values were used to calculate the overall predictive score.

3.2.2.5. Cox regression

The calculations were computed using SPSS, the binary event used was relapse status, where 1 represents relapse and 0 indicates event free survival, the independent variable was the gene expression value or values and the time variable was the event free survival time. A P-value of $P<0.05$ was considered statistically significant.

3.2.2.6. Kaplan–Meier survival curve

Kaplan–Meier survival analysis curves, illustrate the survival of a population based on an event and time to the event, in this instance the event was relapse (1), the censoring event was event free survival (0), and the time used was event free survival time. It was used to compare the difference of event free survival time in populations of high or low relapse prognostic index. A Mann-Whitney u test was calculated alongside the Kaplan-Meier survival curve, which is detailed previously.

3.2.3. Type one error correction

The method implemented in this investigation to reduce type one error, was the Benjamini & Hochberg[163] method otherwise known as false discovery rate (FDR). To obtain corrected p-values the r function "*p.adjust*" was used. A P<0.05 after FDR was considered significant.

3.2.4. Discovery workflows

To develop a robust method of biomarker discovery, several workflows were trialled, which combine statistical and machine learning approaches (figure 3.1). These workflows. After the initial variable workflows, shown in colour in figure 3.1, the gene lists obtained were used in independent multivariate binary logistic regression, to identify genes that work synergistically to predict patient outcome.

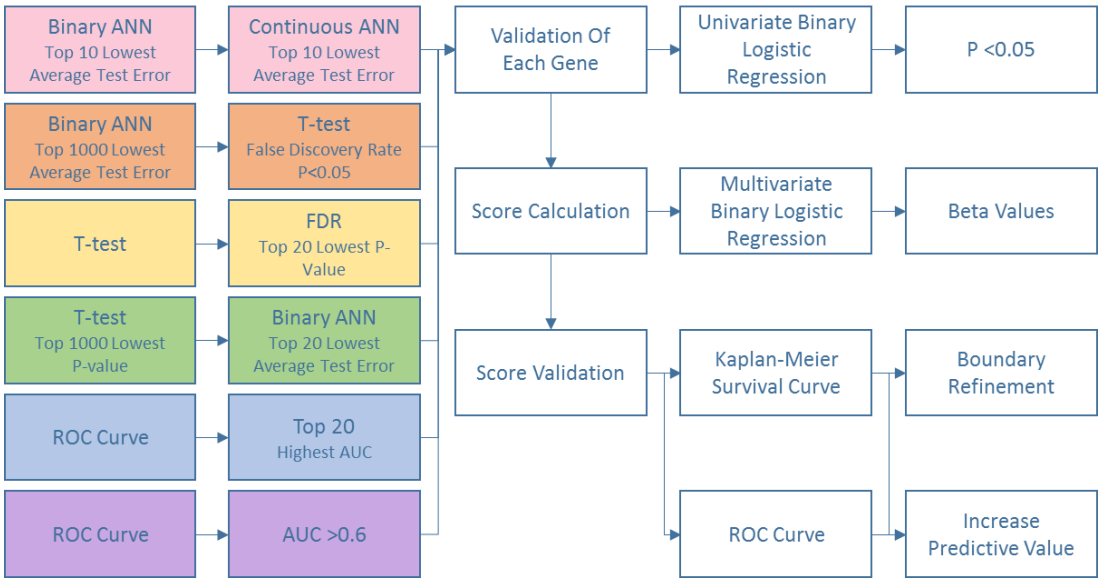

FIGURE 3.1. BIOMARKER DISCOVERY WORKFLOWS. A COMBINATION OF STATISTICS AND MACHINE LEARNING USED IN THE INITIAL BIOMARKER DISCOVERY WORKFLOW LEADS INTO A UNIFORM SCORE CALCULATION AND VALIDATION PROCEDURE, ENSURING DIRECTLY COMPARABLE SCORES.

The gene values used to calculate the overall score were obtained from the finalised gene list of each workflow, the β values were obtained from the multivariate binary logistic regression of the finalised gene list for each workflow. The number of elements within the score calculation can be increased or decreased according to the number of genes obtained for each method. To calculate the score the following equation was used[164], where Σ indicates the sum of, and β values were obtained though binary logistic regression:

$$\sum(gene \times \beta\ value\ of\ the\ gene)$$

3.2.4.1.ANN

The ANN workflow is indicated in pink in figure 3.1. Using the ANN algorithm explained previously, relapse status was used as the initial binary input where 1 indicates relapse and 0 indicates event free survival. The top 10 genes by lowest average test error, from the initial binary ANN, were each used as continuous inputs for the next ANN analysis. The top 10 genes by lowest average test error from each of the continuous ANN were taken resulting in 110 genes in total, not accounting for duplicates in the results.

3.2.4.2.ANN-first

The ANN-first workflow is indicated in orange in figure 3.1. Using the ANN algorithm explained previously, relapse status was used as the initial binary input where 1 indicates relapse and 0 indicates event free survival. The top 1000 genes by lowest average test error were taken, and a T-test was performed, where patients were grouped by their relapse status. After a false discovery P-value correction was used, up to 20 genes with the lowest P-value were selected.

3.2.4.3.T-test

The T-test workflow is indicated in yellow in figure 3.1. A T-test was performed where patients were grouped by their relapse status, event free survival versus relapse. After a false discovery P-value correction was used, the top 20 genes by lowest P-value were selected.

3.2.4.4.ANN-Second

The ANN-second workflow is indicated in green in figure 3.1. A T-test was performed where patients were grouped by their relapse status, event free survival versus relapse. The top 1000 genes by lowest P-value were taken as a truncated dataset. Using the ANN algorithm

explained previously, relapse status was used as a binary input where 1 indicates relapse and 0 indicates event free survival. The top 20 genes by lowest average test error were selected.

3.2.4.5. ROC

The ROC workflow is indicated in blue in figure 3.1. The R package *pROC*[160] was used to determine the P-value and the AUC for each gene in the dataset, to identify genes which can discriminate between Event free survival and relapse patients. The genes were sorted by AUC and the top 20 highest AUC were selected.

3.2.4.6. BLR

The binary logistic regression workflow is indicated in purple in figure 3.1. The R package *pROC*[160] was used to determine the P-value and the AUC for each gene in the dataset, to identify genes which can discriminate between Event free survival and relapse patients. The genes were sorted by AUC, all the genes with an AUC >0.6 were selected. A binary logistic regression was performed on these genes, which were then sorted by beta-value. The genes with the 10 highest and 10 lowest beta-values were selected.

3.2.5. Assigning weight to individual markers

Biomarkers can function individually or synergistically to achieve an appropriate level of confidence in its predictive ability. Individual biomarkers can predict the outcome independently, with an appropriate level of confidence and statistical significance. Synergistic biomarkers are typically a panel of individual biomarkers markers, working together to improve the level of confidence in the prediction and therefore the overall predictive value.

Individual biomarkers can have a positive or negative correlation to an outcome, it is important to recognise these differences to ensure the individual markers aren't being miss-represented and detract from the overall predictive value of the panel, and as such the individual biomarkers can be assigned weights when creating a score. When assigning a weight to a biomarker, the direction of the correlation is the first consideration. A positive weight indicates a positive correlation between the expression and outcome, therefore an increase in the marker would increase the likelihood of the outcome. The opposite is true of a negative weight, it indicates a negative correlation between the expression and outcome, therefore a decrease in the marker would increase the likelihood of the outcome. Although each marker in a panel work independently to predict the outcome, it is important to note

not all these markers do so with the same level of confidence. By assigning a weight to the constituting markers of a panel the relative contribution of the individual markers to the score is accounted for, making the score more reliable and accurate. For this project the weights were determined by using the beta values obtained from the binary logistic regression or cox regression, as outlined in the Nottingham prognostic index[164,185,186].

3.3. Results

3.3.1. Comparing Binary logistic regression and cox regression

Although the principals of Cox Regression (CR) and Binary Logistic Regression (BLR) are the same, it is important to establish a standard procedure for use in the workflow. The workflow should be applicable to different scenarios therefor it is important to determine which type of regression to use. Following the t-test workflow, the top 20 most significant genes were used in a univariate CR and BLR. Figure 3.2 shows the differences in the β value obtained from both CR and BLR: It is evident that the β values for BLR are consistently higher in the positive range, and consistently lower in the negative range, than those of CR, however the standard error was also consistently higher in BLR. The β values of BLR being more extreme than those of CR can influence and impact the score, as they are used directly in the calculation. The genes that were still significant in the univariate analysis were then used in a multivariate CR and BLR, removing the least significant each time, until all genes were significant. Overall, the CR score consists of four genes: HAL, SOCS2-AS1, TES and NYNRIN. The BLR score consists of five genes: HAL, TES, SOCS2-AS1, NYNRIN and CD109. The β values obtained in the multivariate analysis and the gene expression values were used to calculate the score.

The comparison of the prognostic scores for both BLR and CR are shown in figure 3.3. The Kaplan-Meier survival graph (figure 3.3A) shows a median split based on the score calculated in each method. There is a significant difference between the high and low group for the scores calculated using both BLR and CR, there is a slight difference between the splits with BLR separating the groups to a slightly greater extent. When comparing the predictive value of the scores over a 5-year period, there is minimal difference between BLR and CR however BLR preformed marginally better as demonstrated by the greater split between groups. When looking at figure 3.3B, the AUC for the ROC curve was extremely similar therefor the predictive value of the scores are similar. Although the AUC are similar the predictive value of the BLR method is better than CR. When looking at figure 3.3C, the distribution of the

score for relapse and EFS patients, the inter quartile range is larger for BLR when compared to CR, although both scores significantly differentiate between relapse and event free survival patients. there is a significant difference between the two low scoring groups owing to the lower median of the BLR group in comparison to the CR group, similarly there is a significant difference between the high score groups but not to the same extent as the low groups, in part due to the larger interquartile range of the BLR group. Overall BLR performed better that CR in each instance, and although CR takes time into account, this was not a necessary variable for prediction of relapse, as the aim of the investigation is to predict relapse at any time in the patient's future. In addition, the use of BLR allows for the method to be consistently used even with variables that do not have an associated time, such as mutation status, as such will be the method used to carry forward onto further workflows.

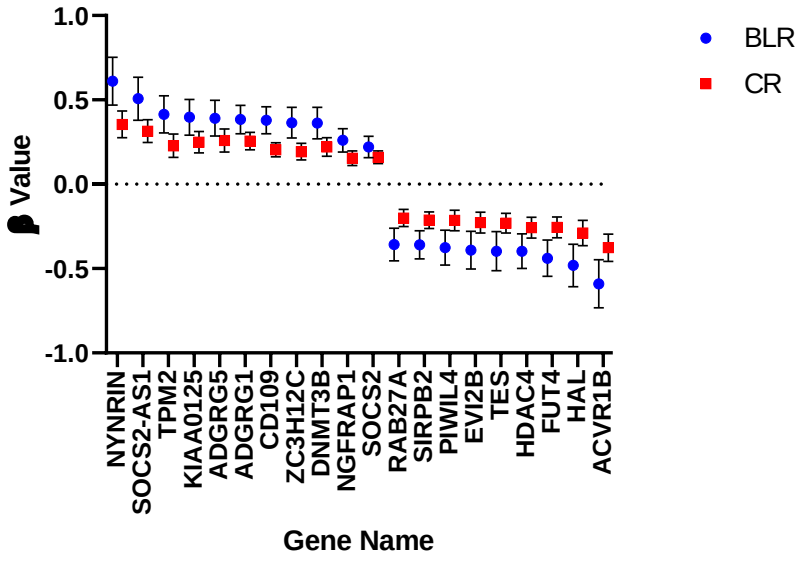

FIGURE 3.2. COMPARISON OF B VALUES FROM BINARY LOGISTIC REGRESSION AND COX REGRESSION. THE B VALUES OBTAINED FROM BINARY LOGISTIC REGRESSION (BLR) AND COX REGRESSION (CR) ALONG WITH THE STANDARD ERROR OF THE MEAN, AS CALCULATED IN SPSS. EACH OF TOP 20 GENES MOST SIGNIFICANTLY ASSOCIATED WITH RELAPSE AS DETERMINED BY T-TEST IN THE HOVON DATASET ARE PRESENTED HERE.

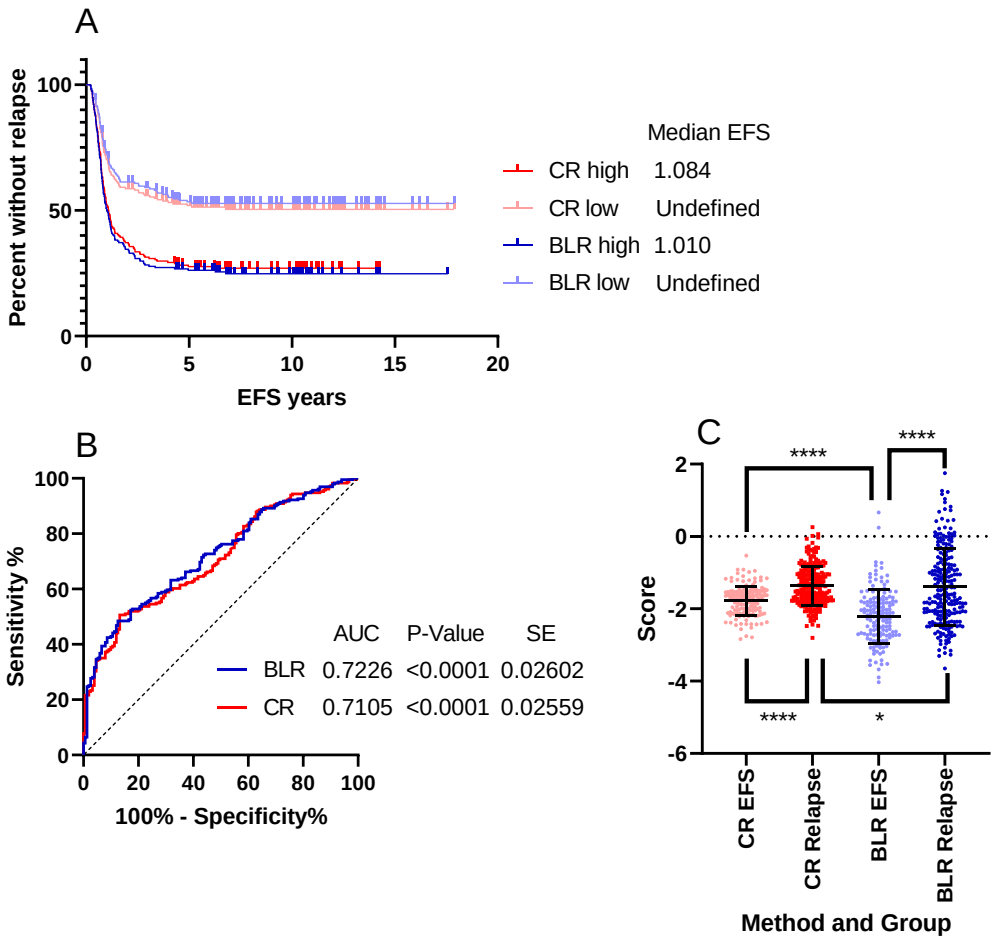

FIGURE 3.3. COMPARING SCORES CREATED USING CR AND BLR. SCORES PREDICTING RELAPSE WERE CALCULATED USING GENE EXPRESSION AND B VALUES FROM EITHER COX REGRESSION (CR) OR BINARY LOGISTIC REGRESSION (BLR), THE OVERALL PREDICTIVE VALUE IS SHOWN HERE IN THE HOVON DATASET. A) KAPAN-MEIER SURVIVAL CURVE SHOWING THE MEDIAN SPLIT OF THE SCORE IN EACH METHOD, THE MEDIAN EVENT FREE SURVIVAL (EFS) IS SHOWN. DIFFERENCES BETWEEN THE HIGH AND LOW GROUP OF EACH METHOD WERE SIGNIFICANT WITH A P-VALUE OF <0.0001, HOWEVER THE DIFFERENCES BETWEEN THE LOW GROUPS WAS NOT SIGNIFICANT AS WERE THE HIGH GROUP. B) ROC CURVE SHOWING THE PREDICTIVE POWER OF THE SCORE, THE AREA UNDER THE CURVE (AUC), P-VALUE AND STANDARD ERROR (SE) ARE SHOWN. C) SCATTER PLOT COMPARING THE DISTRIBUTION OF THE SCORES.

3.3.2. Biomarker discovery Work-flow comparison

The RMA normalised HOVON AML dataset was used as the discovery cohort for the workflows outline in figure 3.1. By comparing the different black box and white box methods and using them in combination a predictive score was developed. To determine the best method of biomarker discovery from those outlined in figure 3.1, the results were validated in both TCGA and TARGET AML datasets.

3.3.2.1.ANN Alone

ANN looks at the associations of the input variable to all other variable in the dataset. By using the EFS versus Relapse as the initial input and then expanding this with the top ten genes by average test error a network of up to 110 genes are identified. In figure 3.4 the network of genes is shown, there are 84 unique genes in the network, which create two main hubs of association. One hub contains four genes which are directly linked in the analysis: ADGRG1, CD109, ZC3H12C, KIAA0125 and there are links between their associated genes, additionally NYNRIN is linked to this hub by two common associated genes. The second hub of association contains three genes: HAL, SIRPB2 and ACVR1B, which are linked directly and indirectly by their associated genes. There are two genes which are not linked to another hub in the network, FUT4 and SOCS2-AS1, which are not linked directly or indirectly via there associated genes. The hubs of association have a likelihood of working synergistically in a predictive panel of genes, however this network alone is not detailed enough to form a predictive score.

As a follow up to the ANN, and to refine the gene list to those that will work synergistically in a biomarker panel, a BLR is used. The univariate BLR validates the genes which work as predictive biomarkers while identifying those which do not, therefor only genes which are significantly associated with relapse were used in the multivariate BLR. There were 52 genes which were identified as significantly associated with relapse through the univariate BLR (supplementary table 1), these genes were carried forward into further analysis. To refine the 52 genes and identify those which work synergistically to predict relapse in AML, a multivariate BLR was implemented. The least significant gene was removed from the analysis each time, until all genes were significant together, the results of which are shown in table 3.2. From here clinical annotations were added to the seven remaining significant genes, of which none significantly contributed to the panel (supplementary tables 2-5) therefore the seven genes in table 3.2 were the basis of this score.

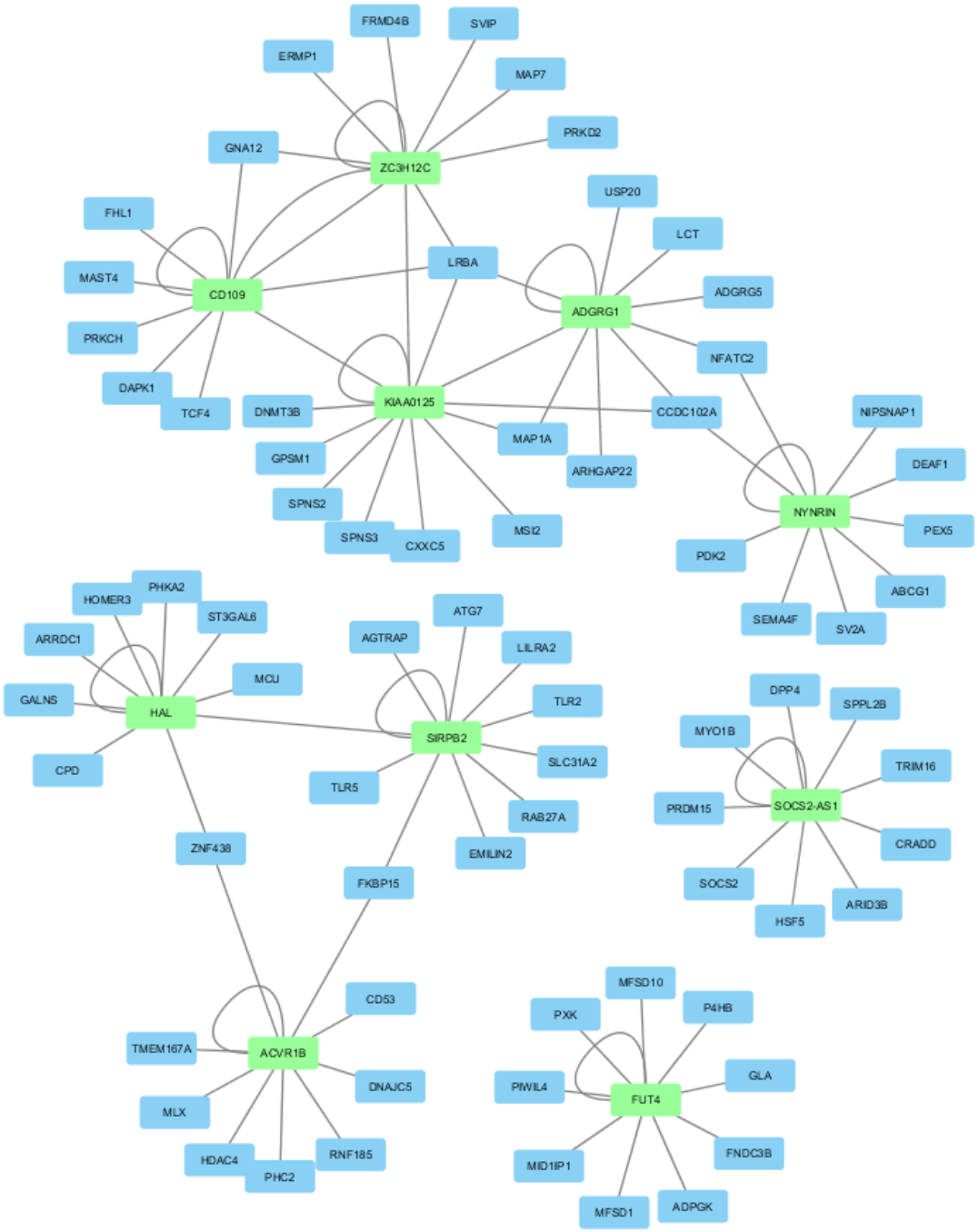

FIGURE 3.4. NETWORK OF GENES ASSOCIATED WITH RELAPSE. USING CYTOSCAPE, THE NETWORK OF ASSOCIATIONS DISCOVERED USING THE ANN WORKFLOW IN THE HOVON DATASET IS VISUALISED. THERE ARE TWO MAIN HUBS OF ASSOCIATIONS, WHICH THE GENES ARE LINKED DIRECTLY AND INDIRECTLY THROUGH THE ANN ANALYSIS.

The score was calculated using the beta values from the multivariate BLR in table 3.2. The score was calculated for each patient in the HOVON dataset, which was then split by the median score value. The dataset was split into the EFS and relapse patients to compare the average score values, indicated in figure 3.5A. There was a significant difference between

the average score of the EFS and relapse patients, with the median score of the relapse patients being significantly higher than that of the EFS patients. To assess the scores predictive ability a ROC curve was used shown in figure 3.5B, indicating an AUC of 0.7376 which is equal to a 73.76% accuracy. Figure 3.5C shows a Kaplan-Meier survival plot based on the median split of the dataset by the calculated score; in which the high score group had an EFS of less than one year, meaning 50% of the group had relapsed before one year after they achieved remission. The low score group had an undefined median EFS, indicating the group never reached 50% relapse, although some people in the low score group did relapse, far fewer than the high score group.

Gene name	B	S.E.	Wald	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
ACVR1B	-0.773	0.237	10.647	0.001	0.462	0.290	0.734
DNMT3B	0.545	0.171	10.178	0.001	1.725	1.234	2.411
SOCS2-AS1	0.557	0.175	10.082	0.001	1.746	1.238	2.462
CD109	0.308	0.098	9.831	0.002	1.360	1.122	1.649
ADPGK	-0.948	0.338	7.860	0.005	0.388	0.200	0.752
SLC31A2	0.399	0.182	4.798	0.028	1.491	1.043	2.131
MSI2	-0.265	0.127	4.363	0.037	0.767	0.598	0.984

TABLE 3.2. MULTIVARIATE BINARY LOGISTIC REGRESSION FROM THE ANN WORKFLOW. THE GENES SIGNIFICANTLY ASSOCIATED WITH RELAPSE FROM THE UNIVARIATE BLR WERE REFINED USING THE MULTIVARIATE BLR TO ONLY THOSE WORKING SYNERGISTICALLY IN THE HOVON DATASET.

To validate this score, the same genes were selected from the TCGA and TARGET AML datasets, the score was calculated for the EFS and relapse patients, using the beta values as before. Figure 3.6A shows the dot plot diagrams comparing the distribution of the score in the datasets, neither dataset showed a significant difference between the relapse or EFS patients, in contrast to the HOVON discovery dataset. Figure 3.6B shows the ROC curve of the score in both datasets, which have a similar AUC of around 0.58, indicating 58% accuracy of this score, which is low in comparison to the HOVON dataset. Figure 3.6C shows the Kaplan-Meier survival plot, the datasets were split by the median value of the score, there is no significant difference between the high and low groups in each dataset. Due to the inconsistent results between datasets, this is not an ideal panel of markers to use to predict relapse in AML. The large variation in the AUC when comparing the discovery cohort to the validation cohorts means the score is only accurate in the discovery cohort.

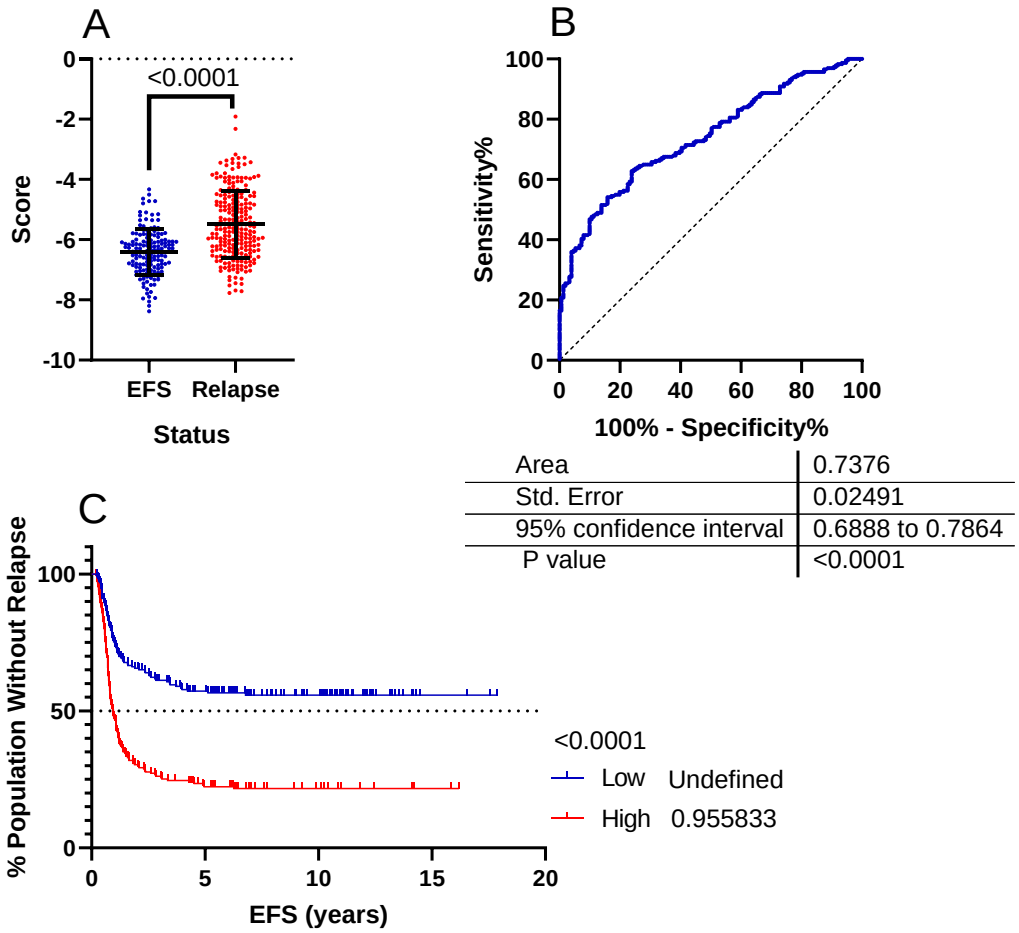

FIGURE 3.5. PREDICTIVE ABILITY OF THE ANN ALONE WORKFLOW IN THE HOVON DATASET. FIGURE A SHOWS THE SCORE CALCULATED FOR EACH PATIENT WHICH WAS SPLIT BY THEIR RELAPSE STATUS, THE MEDIAN VALUE FOR RELAPSE PATIENTS IS SIGNIFICANTLY HIGHER THAN EFS PATIENTS. FIGURE B SHOWS THE ROC CURVE WITH AN AREA OF 0.7376, INDICATING A 73.76% ACCURACY. FIGURE C SHOWS THE PATIENTS SPLIT BY THE MEDIAN VALUE OF THE SCORE, THE GROUPS ARE SIGNIFICANTLY DIFFERENT, AND THE MEDIAN EFS FOR THE LOW SCORE GROUP WAS UNDEFINED WHEREAS THOSE WITH A HIGH SCORE HAD A MEDIAN EFS OF UNDER A YEAR.

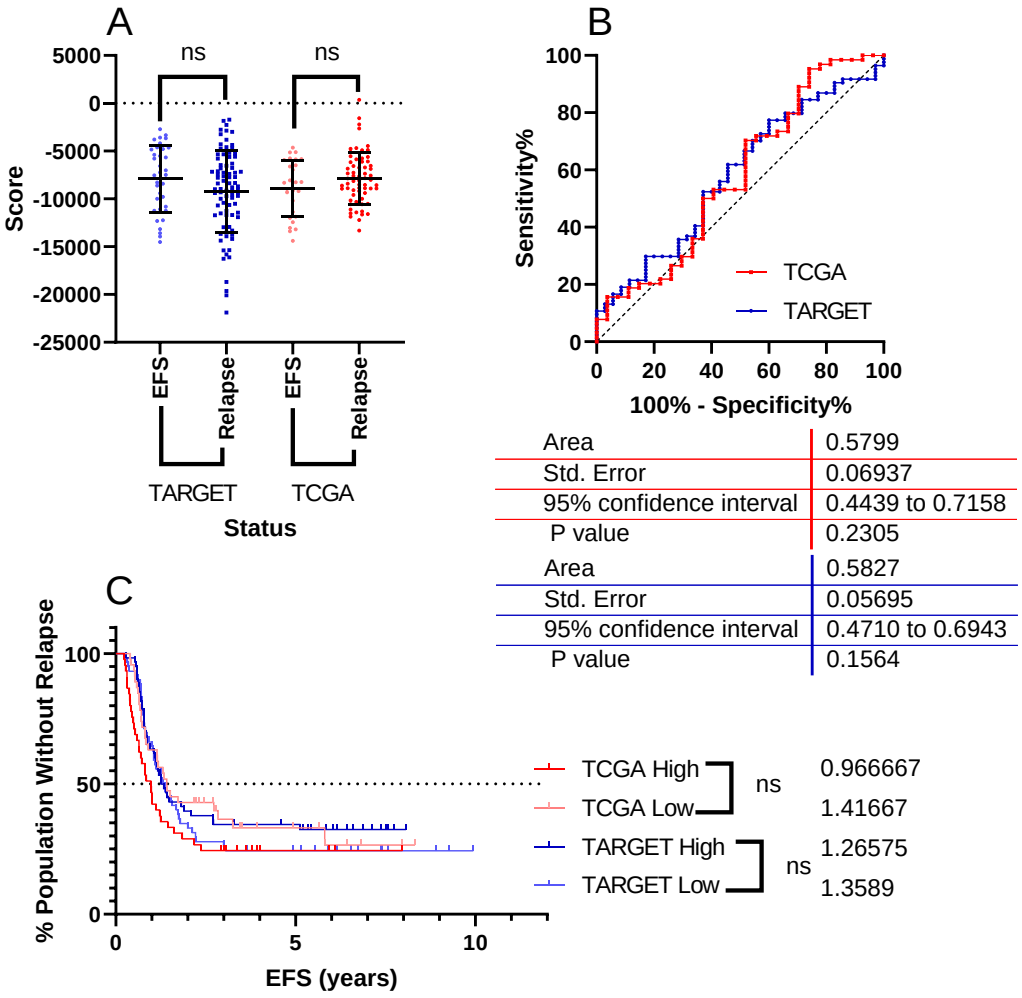

FIGURE 3.6. VALIDATION OF THE ANN ALONE WORKFLOW IN TCGA AND TARGET AML. THE SCORE WAS CALCULATED FOR THE RELAPSE AND EFS PATIENTS IN EACH DATASET. FIGURE A SHOWS THE DISTRIBUTION OF THE SCORE IN THE DATASETS COMPARING THE EFS AND RELAPSE PATIENTS, NEITHER DATASET SHOWED A SIGNIFICANT DIFFERENCE BETWEEN THE GROUPS. FIGURE B SHOWS THE ROC CURVE FOR THE SCORE IN THE DATASETS, BOTH DATASETS HAS A SIMILAR AUC OF 0.57, INDICATING A 57% ACCURACY. FIGURE C SHOWS THE KAPLAN-MEIER SURVIVAL PLOT, SPLIT BY THE MEDIAN VALUE OF THE SCORE FOR EACH DATASET, THERE IS NO SIGNIFICANT DIFFERENCE BETWEEN THE HIGH AND LOW GROUPS IN EITHER DATASET.

3.3.2.2.ANN First

The ANN algorithm was used to identify the top 1000 genes with the lowest average test error. The genes identified were narrowed down using a t-test comparing EFS and relapse patients and the P-values were corrected with the p.adjust R script, to correct for the false discovery rate (supplementary table 6). The top 20 most significant genes were validated using a univariate BLR (supplementary table 7), of these only the significant genes were used in the multivariate BLR, which was repeated until all the genes were significantly contributing to the panel shown in table 3.3. Once the genes of the panel had been selected, clinical annotations were added to the multivariate BLR (Table 3.4) to determine which, if any, contributed to the panel of genes. From table 3.3 to 3.4 the gene CD109 was removed but the clinical annotation of cytogenetic risk group was included in the overall score calculation.

	B	S.E.	Wald	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
HAL	-0.535	0.146	13.450	0.000	0.585	0.440	0.779
TES	-0.386	0.139	7.681	0.006	0.680	0.518	0.893
SOCS2-AS1	0.385	0.178	4.672	0.031	1.469	1.037	2.083
NYNRIN	0.407	0.196	4.291	0.038	1.502	1.022	2.208
CD109	0.202	0.099	4.163	0.041	1.224	1.008	1.487

TABLE 3.3. MULTIVARIATE BLR OF THE ANN FIRST METHOD IN THE HOVON DATASET. THE GENES WHICH WERE SIGNIFICANTLY ASSOCIATED WITH RELAPSE AFTER THE UNIVARIATE BLR WERE TAKEN FORWARD TO THIS MULTIVARIATE BLR, THESE ARE THE REMAINING GENES THAT SIGNIFICANTLY CONTRIBUTE TO THE PREDICTIVE PANEL.

	B	S.E.	Wald	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
HAL	-0.519	0.147	12.411	0.000	0.595	0.446	0.794
NYNRIN	0.480	0.188	6.484	0.011	1.616	1.117	2.338
SOCS2-AS1	0.439	0.173	6.428	0.011	1.551	1.105	2.179
TES	-0.327	0.136	5.793	0.016	0.721	0.553	0.941
Cytogenetic risk group	0.452	0.200	5.125	0.024	1.571	1.063	2.324

TABLE 3.4. MULTIVARIATE BLR INCLUDING CLINICAL ANNOTATIONS OF THE ANN FIRST METHOD. CLINICAL ANNOTATIONS WERE INCLUDED IN THE MULTIVARIATE BLR TO DETERMINE IF ANY SIGNIFICANTLY CONTRIBUTED TO THE PANEL IN THE HOVON DATASET. THIS IS THE COMBINATION OF GENES AND CLINICAL ANNOTATIONS WHICH WORK SYNERGISTICALLY AS A BIOMARKER PANEL.

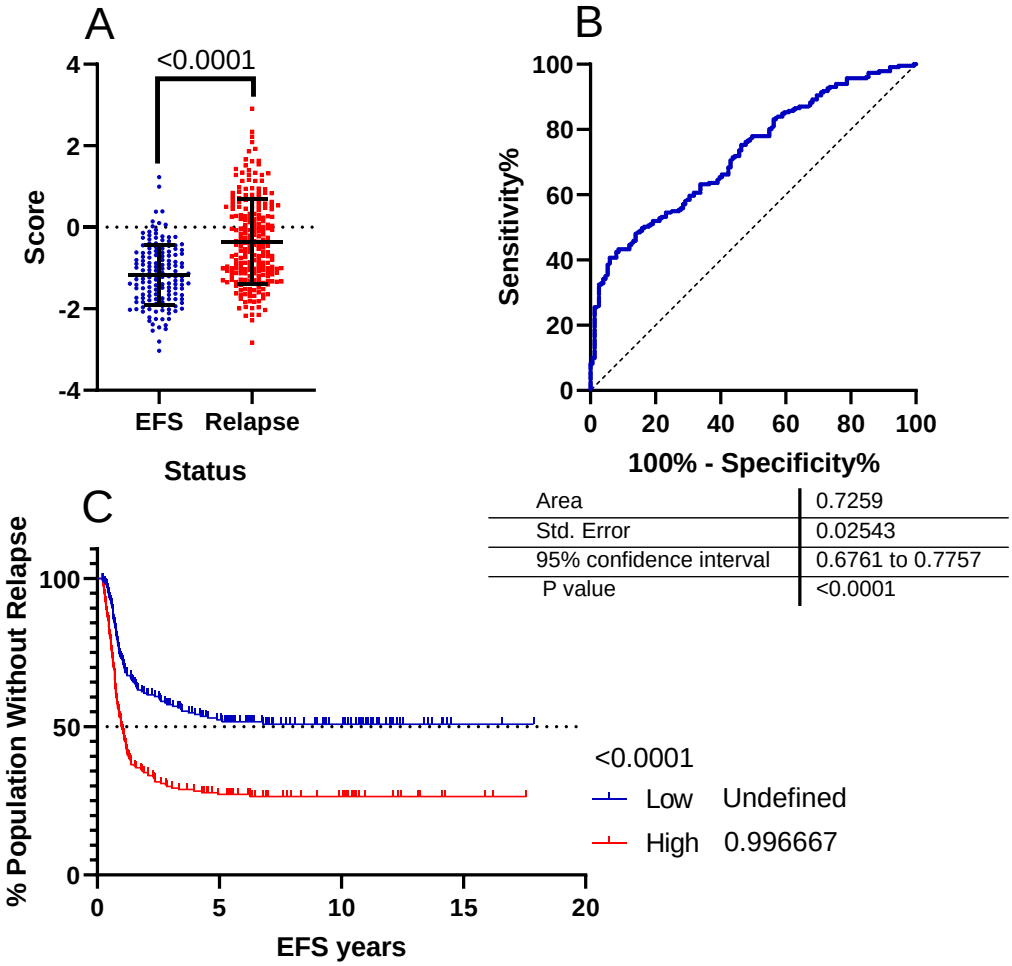

FIGURE 3.7. PREDICTIVE ABILITY OF THE ANN FIRST WORKFLOW IN THE HOVON DATASET. FIGURE A SHOWS THE SCORE CALCULATED FOR EACH PATIENT IN THE HOVON DATASET, WHICH WERE SPLIT BY THEIR RELAPSE STATUS, THE MEDIAN VALUE FOR RELAPSE PATIENTS IS SIGNIFICANTLY HIGHER THAN EFS PATIENTS. FIGURE B SHOWS THE ROC CURVE WITH AN AREA OF 0.7259, INDICATING A 72.59% ACCURACY FOR THIS SCORE. FIGURE C SHOWS THE PATIENTS SPLIT BY THE MEDIAN VALUE OF THE SCORE, THE TWO GROUPS ARE SIGNIFICANTLY DIFFERENT, AND THE MEDIAN EFS FOR THE LOW SCORE GROUP WAS UNDEFINED WHEREAS THOSE WITH A HIGH SCORE HAD A MEDIAN EFS OF UNDER A YEAR.

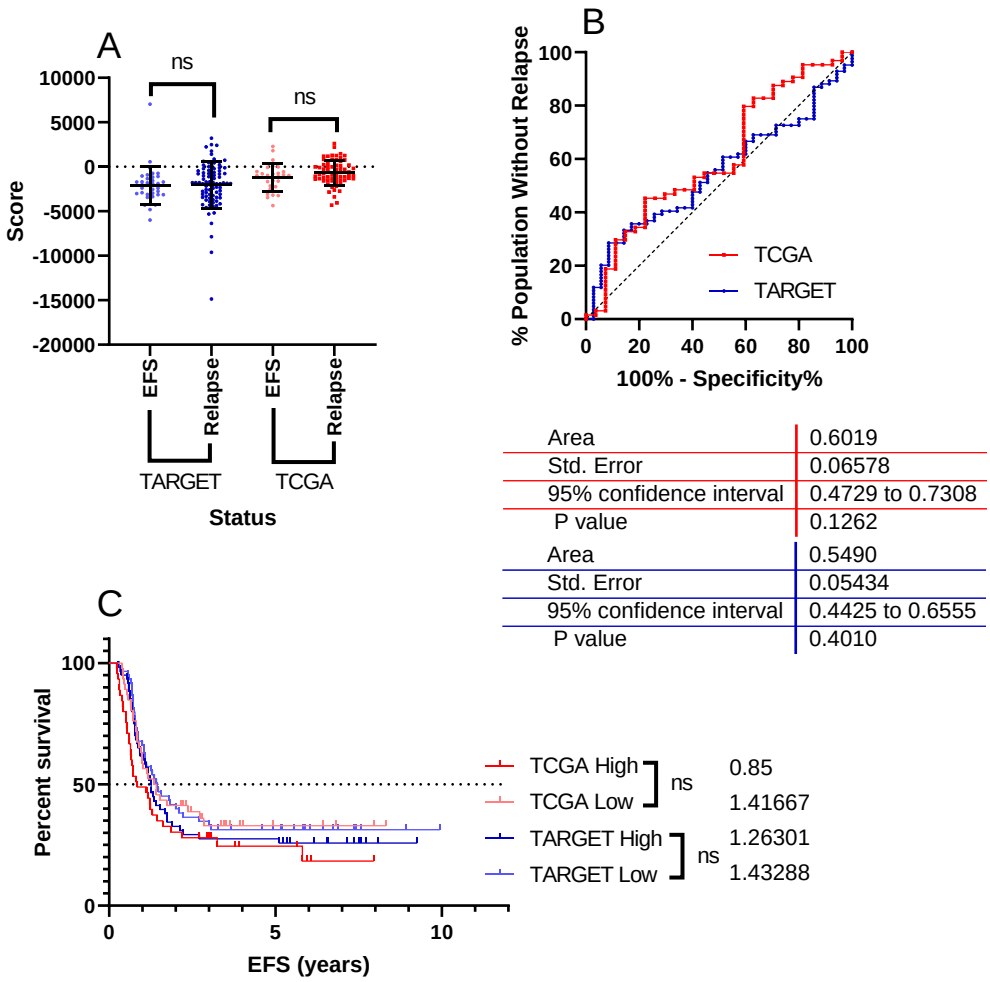

FIGURE 3.8. VALIDATION OF THE ANN FIRST WORKFLOW IN TCGA AND TARGET AML. THE SCORE WAS CALCULATED FOR THE RELAPSE AND EFS PATIENTS IN EACH DATASET. FIGURE A SHOWS THE DISTRIBUTION OF THE SCORE IN THE DATASETS, COMPARING EFS AND RELAPSE PATIENTS, NEITHER DATASET SHOWED A SIGNIFICANT DIFFERENCE BETWEEN GROUPS. FIGURE B SHOWS THE ROC CURVE FOR THE SCORE IN THE DATASETS, TCGA HAD AN AUC OF 0.6019 AND TARGET OF 0.5490, INDICATING AN ACCURACY OF 60.19% AND 54.90% RESPECTIVELY. FIGURE C SHOWS THE KAPLAN-MEIER SURVIVAL PLOT SPLIT BY THE MEDIAN SCORE IN EACH DATASET, THERE IS NO SIGNIFICANT DIFFERENCE BETWEEN THE HIGH AND LOW GROUPS IN EITHER DATASET.

For each patient in the HOVON dataset the score was calculated, in figure 3.7A the distribution of the score in relapse and EFS patients is compared. There is a significant difference between the two groups, where the average score of a relapse patient is much higher than an EFS patient, however there is a large overlap in the distribution of the score between the groups. Comparing the results in figure 3.7A to figure 3.8A the range of the score is very different owing to the different technologies used in the datasets. Although the score was significantly different between relapse and EFS patients in figure 3.7A the same is not said for figure 3.8A where there was no significant difference in either dataset. Although the discovery cohort showed significance the score is not valid when applied to other validation datasets.

In figure 3.7B the AUC is 0.7259 indicating a 72.6% accuracy of the score, therefor the score is valid for the prediction of relapse within the dataset. The AUC for the validation datasets, shown in figure 3.8B, were 0.6019 for TCGA and 0.5490 for TARGET indicating a 60% and 55% accuracy of the score in the validation datasets. As the AUC for the validation datasets were lower than that of the discovery dataset the panel of genes used in this score are not ideal for the prediction of relapse. The results from figure 3.8B coupled with figure 3.8A show the score is not preforming well in the validation datasets.

In figure 3.7C there is an excellent separation of the high and low score groups, with the low group never reaching 50% relapse and the high group reaching this milestone in just under a year. The highly separated and distinct groups indicate the score can separate the patients well in this dataset. In contrast, figure 3.8C shows highly overlapping groups with no significant difference. Although the median EFS differs by 0.18 of a year for TARGET and 0.56 of a year for TCGA, this is of little value given the overlapping of the groups. Although the score was valid in the HOVON discovery dataset, the score does not translate to the TARGET and TCGA validation datasets as there is no significant difference between the high and low scoring groups in figure 3.8C or the EFS and relapse patients in figure 3.8A.

3.3.2.3.T-test

A t-test was used to identify the genes most significantly associated with relapse, the P-values were corrected using the *p.adjust* script in R. The intention was to take the top 20 most significant genes, however as displayed in table 3.5, after the P-value was adjusted for the FDR only 5 genes remained significant. The remaining significant genes were validated using a univariate BLR, shown in table 3.6, all were significantly associated to relapse in AML. The genes were then refined using a multivariate BLR, removing the least significant gene

each time until all genes were significantly contributing to the score, the remaining genes were CD109, NYNRIN and FUT4, as displayed in table 3.7. Once the combination of genes had been determined, the clinical annotations were included in the multivariate BLR, only cytogenetic risk group significantly contributed to the score as shown in table 3.8.

Once the basis of the panel had been determined through the multivariate BLR the, the score was calculated using the beta values displayed in table 3.8. When comparing the score of the EFS to relapse patients in the HOVON dataset, the groups were significantly different. Although there is still a large overlap between the two outcomes the score is much higher in patients who relapse compared to EFS patients, as shown in figure 3.9A. The difference is not evident when looking at figure 3.10A where there is no significant difference between the EFS and release patients in either of the validation datasets.

In the HOVON dataset the AUC was 0.7044 indicating a 70.44% accuracy of the score, as displayed in figure 3.9B, which is in contrast with figure 3.10B, where the AUC was 0.5712 for TCGA and 0.5418 for TARGET. Although in the discovery dataset the predictive ability of the score was high this is not translated to the discovery datasets. This lack of validity to the score is also apparent in figure 3.10C, where there is no significant difference between the high and low groups for either dataset. The discovery dataset did show a significant difference between the high and low groups in figure 3.9C, as this is not translated across to the validation datasets this score would be unreliable as a predictor of relapse beyond the discovery dataset.

Gene	T-test	BH
CD109	3.06E-07	0.006517
ADGRG1	1.22E-06	0.012992
NYNRIN	3.4E-06	0.024138
SIRPB2	4.97E-06	0.026463
FUT4	8.33E-06	0.035482
ACVR1B	1.56E-05	0.055375
SOCS2-AS1	1.85E-05	0.056288
ZC3H12C	3.53E-05	0.093977
RAB27A	6.94E-05	0.143084
HDAC4	7.2E-05	0.143084
DNMT3B	7.39E-05	0.143084
HAL	9.14E-05	0.16222
PIWIL4	0.000118	0.18649
NGFRAP1	0.000124	0.18649
TPM2	0.000131	0.18649
TES	0.000163	0.204777
KIAA0125	0.000174	0.204777
ADGRG5	0.000182	0.204777
EVI2B	0.000183	0.204777
SOCS2	0.000211	0.224258

TABLE 3.5. TOP 20 MOST SIGNIFICANT GENES FROM THE T-TEST WORFLOW. A T-TEST WAS PERFORMED ON ALL THE GENES IN THE HOVON DATASET, THE TOP 20 ARE DISPLAYED WITH THE BENJAMINI HOCHBERG FLASE DISCOVERY RATE CORRECTION.

	B	S.E.	Wald	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
CD109	0.379	0.080	22.349	0.000	1.461	1.248	1.709
ADGRG1	0.384	0.084	20.833	0.000	1.468	1.245	1.731
NYNRIN	0.611	0.142	18.489	0.000	1.842	1.394	2.433
SIRPB2	-0.359	0.084	18.351	0.000	0.698	0.593	0.823
FUT4	-0.438	0.108	16.518	0.000	0.645	0.522	0.797

TABLE 3.6. UNIVARIATE BLR OF THE T-TEST WORKFLOW. ONCE THE GENES HAD BEEN DISCOVERED USING A T-TEST AND THE FALSE DISCOVERY RATE WAS CORRECTED, THE REMAINING GENES WERE ANALYSED TO ENSURE THEY WERE SIGNIFICANTLY ASSOCIATED WITH RELAPSE IN THE HOVON DATASET.

	B	S.E.	Wald	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
CD109	0.228	0.093	6.039	0.014	1.256	1.047	1.506
NYNRIN	0.541	0.177	9.407	0.002	1.718	1.216	2.429
FUT4	-0.488	0.116	17.666	0.000	0.614	0.489	0.771

TABLE 3.7. MULTIVARIATE BLR OF THE T-TEST WORKFLOW. THE GENES WHICH WERE SIGNIFICANTLY ASSOCIATED WITH RELAPSE AFTER THE UNIVARIATE BLR WERE TAKEN FORWARD TO THIS MULTIVARIATE BLR, THESE ARE THE REMAINING GENES THAT SIGNIFICANTLY CONTRIBUTE TO THE PREDICTIVE PANEL IN THE HOVON DATASET.

	B	S.E.	Wald	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
CD109	0.190	0.094	4.022	0.045	1.209	1.004	1.455
NYNRIN	0.510	0.178	8.248	0.004	1.666	1.176	2.359
FUT4	-0.416	0.120	11.973	0.001	0.659	0.521	0.835
Cytogenetic Risk Group	0.446	0.198	5.057	0.025	1.562	1.059	2.303

TABLE 3.8. MULTIVARIATE BLR OF THE T-TEST WORKFLOW INCLUDING CLINICAL ANNOTATIONS. CLINICAL ANNOTATIONS WERE INCLUDED IN THE MULTIVARIATE BLR TO DETERMINE IF ANY SIGNIFICANTLY CONTRIBUTED TO THE PANEL. THIS IS THE COMBINATION OF GENES AND CLINICAL ANNOTATIONS WHICH SIGNIFICANTLY CONTRIBUTE TO THE PANEL IN THE HOVON DATASET.

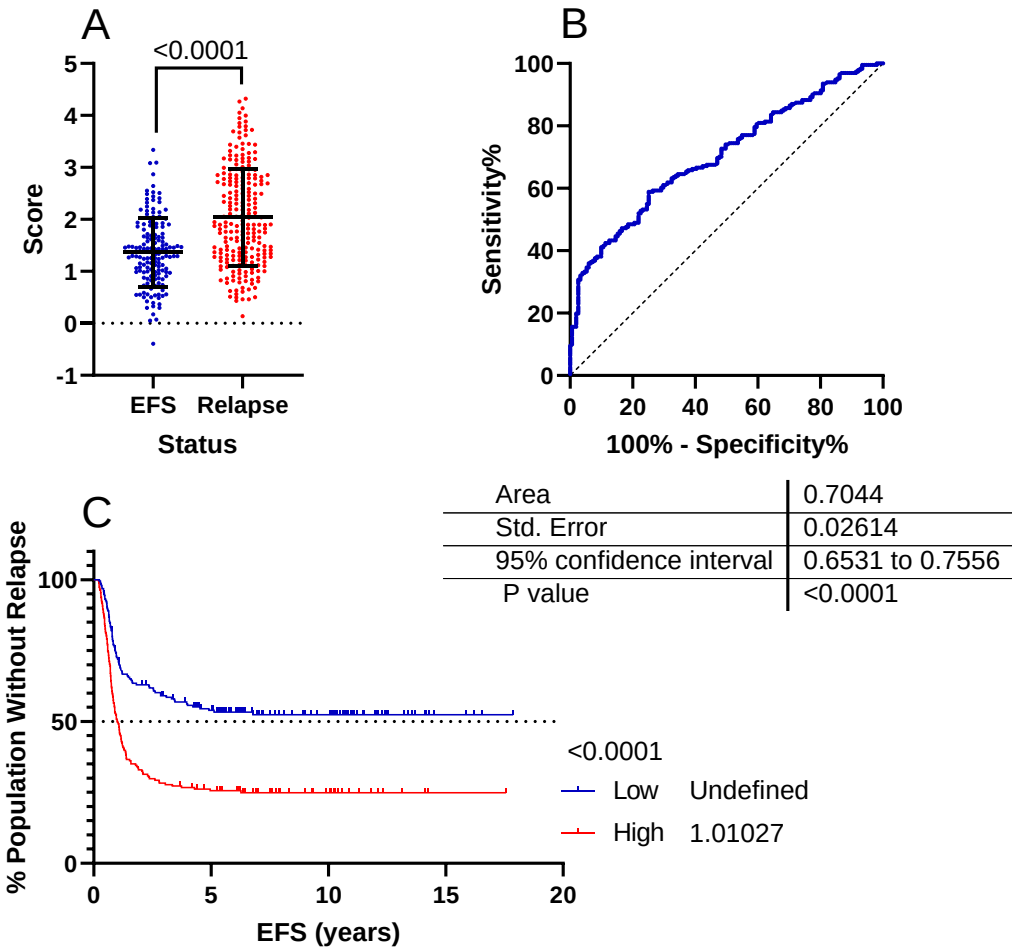

FIGURE 3.9. PREDICTIVE ABILITY OF THE T-TEST WORKFLOW IN THE HOVON DATASET. FIGURE A SHOWS THE DISTRIBUTION OF THE SCORE CALCULATED FOR EACH PATIENT IN THE HOVON DATASET, WHO WERE SPLIT BY THEIR RELAPSE STATUS, THE MEDIAN VALUE FOR RELAPSE PATIENTS IS SIGNIFICANTLY HIGHER THAN EFS PATIENTS. FIGURE B SHOWS THE ROC CURVE WITH AN AUC OF 0.7044, INDICATING A 70.44% ACCURACY. FIGURE C SHOWS THE MEDIAN SPLIT OF THE CALCULATED SCORE, THE GROUPS ARE SIGNIFICANTLY DIFFERENT, THE MEDIAN EFS FOR THE LOW SCORE GROUP WAS UNDEFINED WHEREAS THOSE WITH A HIGH SCORE HAD A MEDIAN EFS OF JUST OVER A YEAR.

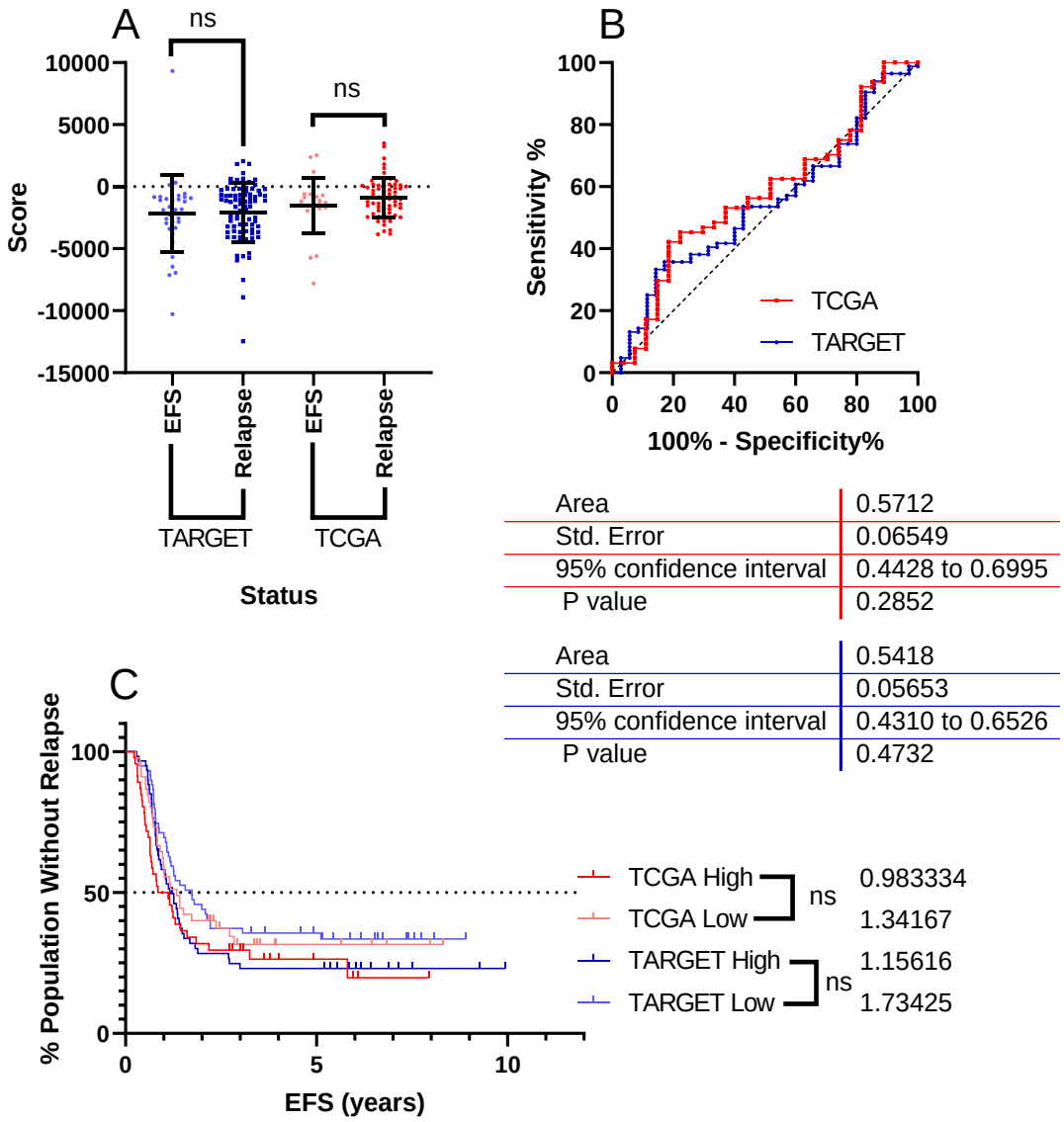

FIGURE 3.10. VALIDATION OF THE T-TEST WORKFLOW IN TCGA AND TARGET AML. THE SCORE WAS CALCULATED FOR EACH PATIENT IN BOTH DATASET. FIGURE A SHOWS THE DISTRIBUTION OF THE SCORE IN RELAPSE AND EFS PATIENTS, NEITHER DATASET SHOWED A SIGNIFICANT DIFFERENCE BETWEEN THE GROUPS. FIGURE B SHOWS THE ROC CURVE FOR THE SCORE IN THE DATASETS, THE AUC FOR TCGA WAS 0.572 AND FOR TARGET 0.5418, INDICATING AN ACCURACY OF 57.72% AND 54.18% RESPECTIVELY. FIGURE C SHOWS THE KAPLAN-MEIER SURVIVAL PLOT SPLIT BY THE MEDIAN VALUE OF THE CALCULATED SCORE, THERE IS NO SIGNIFICANT DIFFERENCE BETWEEN THE HIGH AND LOW GROUPS IN EITHER DATASET.

3.3.2.4.ANN Second

For this workflow a t-test comparing EFS and relapse patients was performed, and the top 1000 most significant genes were identified, these were narrowed down using the binary ANN algorithm to select the top 20 genes with the lowest average test error. To validate these genes a univariate BLR was used (supplementary table 8), all of which were significant so were used in the multivariate BLR. The least significant gene was removed each time until all genes significantly contributed to the score (table 3.9). Once all genes were significantly contributing to the panel the clinical annotations were included, of which only the cytogenetic risk group contributed to the panel but removed HAL as a contributor, (table 3.10). The final panel of genes and clinical attributes in table 3.10 were used to calculate the score.

	B	S.E.	Wald	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
SOCS2-AS1	0.551	0.163	11.372	0.001	1.734	1.259	2.389
CD109	0.273	0.093	8.612	0.003	1.314	1.095	1.576
ACVR1B	-0.498	0.191	6.782	0.009	0.608	0.418	0.884
HAL	-0.357	0.174	4.226	0.040	0.699	0.497	0.983

TABLE 3.9. MULTIVARIATE BLR FOR THE ANN SECOND WORKFLOW IN THE HOVON DATASET. THE GENES WHICH WERE SIGNIFICANTLY ASSOCIATED WITH RELAPSE AFTER THE UNIVARIATE BLR WERE TAKEN FORWARD TO THIS MULTIVARIATE BLR, THESE ARE THE REMAINING GENES THAT SIGNIFICANTLY CONTRIBUTE TO THE PREDICTIVE PANEL.

	B	S.E.	Wald	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
ACVR1B	-0.660	0.156	17.941	0.000	0.517	0.381	0.701
SOCS2-AS1	0.447	0.156	8.265	0.004	1.564	1.153	2.121
CD109	0.252	0.094	7.221	0.007	1.287	1.071	1.547
Cytogenetic Risk Group	0.484	0.199	5.940	0.015	1.623	1.099	2.396

TABLE 3.10. MULTIVARIATE BLR FOR THE ANN SECOND WORKFLOW INCLUDING CLINICAL ANNOTATIONS. CLINICAL ANNOTATIONS WERE INCLUDED IN THE MULTIVARIATE BLR TO DETERMINE IF ANY SIGNIFICANTLY CONTRIBUTED TO THE PANEL. THIS IS THE COMBINATION OF GENES AND CLINICAL ANNOTATIONS WHICH SIGNIFICANTLY CONTRIBUTE TO THE PANEL IN THE HOVON DATASET.

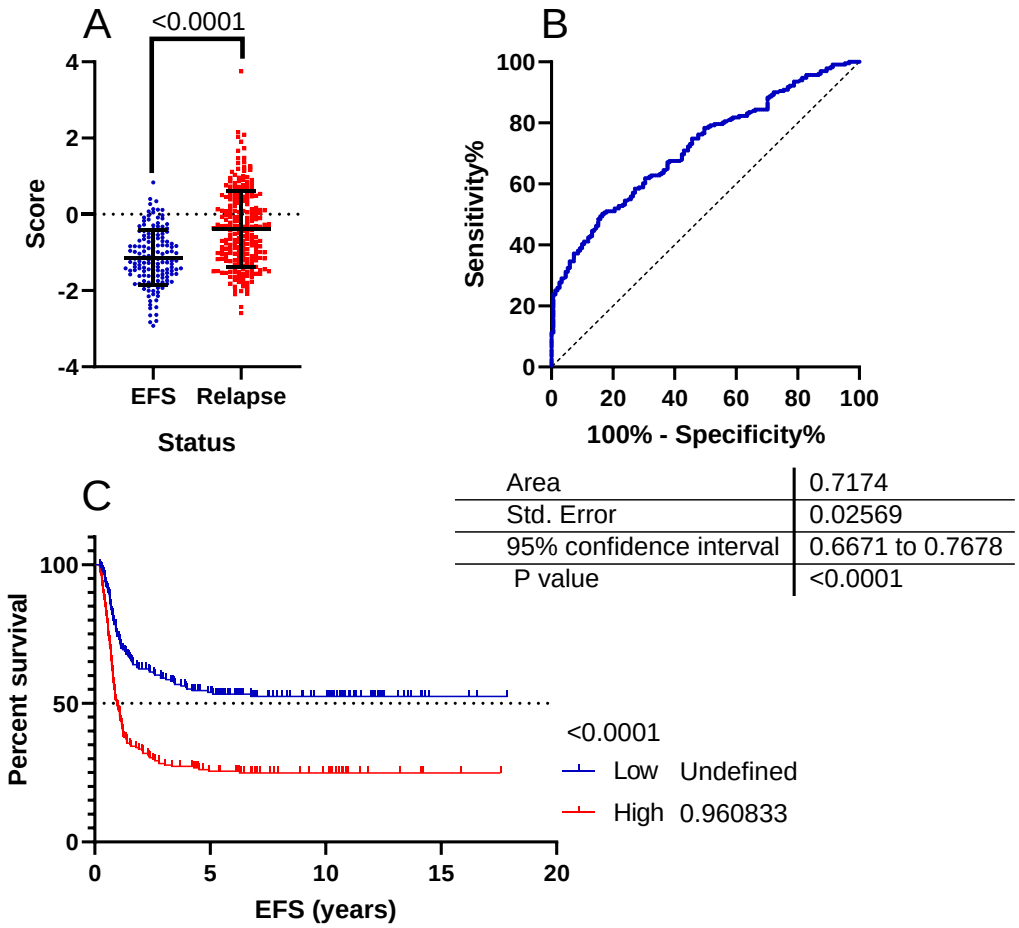

FIGURE 3.11. PREDICTIVE ABILITY OF THE ANN SECOND WORKFLOW IN THE HOVON DATASET. FIGURE A SHOWS THE DISTRIBUTION OF THE SCORE CALCULATED FOR EACH PATIENT, WHICH WAS SPLIT BY THEIR RELAPSE STATUS, THE MEDIAN VALUE FOR RELAPSE PATIENTS IS SIGNIFICANTLY HIGHER THAN EFS PATIENTS. FIGURE B SHOWS THE ROC CURVE WITH AN AUC OF 0.7174, INDICATING A 71.74% ACCURACY. FIGURE C SHOWS THE KAPLAN-MEIER SURVIVAL PLOT SPLIT BY THE MEDIAN VALUE OF THE CALCULATED SCORE, THE GROUPS ARE SIGNIFICANTLY DIFFERENT, THE MEDIAN EFS FOR THE LOW SCORE GROUP WAS UNDEFINED WHEREAS THOSE WITH A HIGH SCORE HAD A MEDIAN EFS OF UNDER A YEAR.

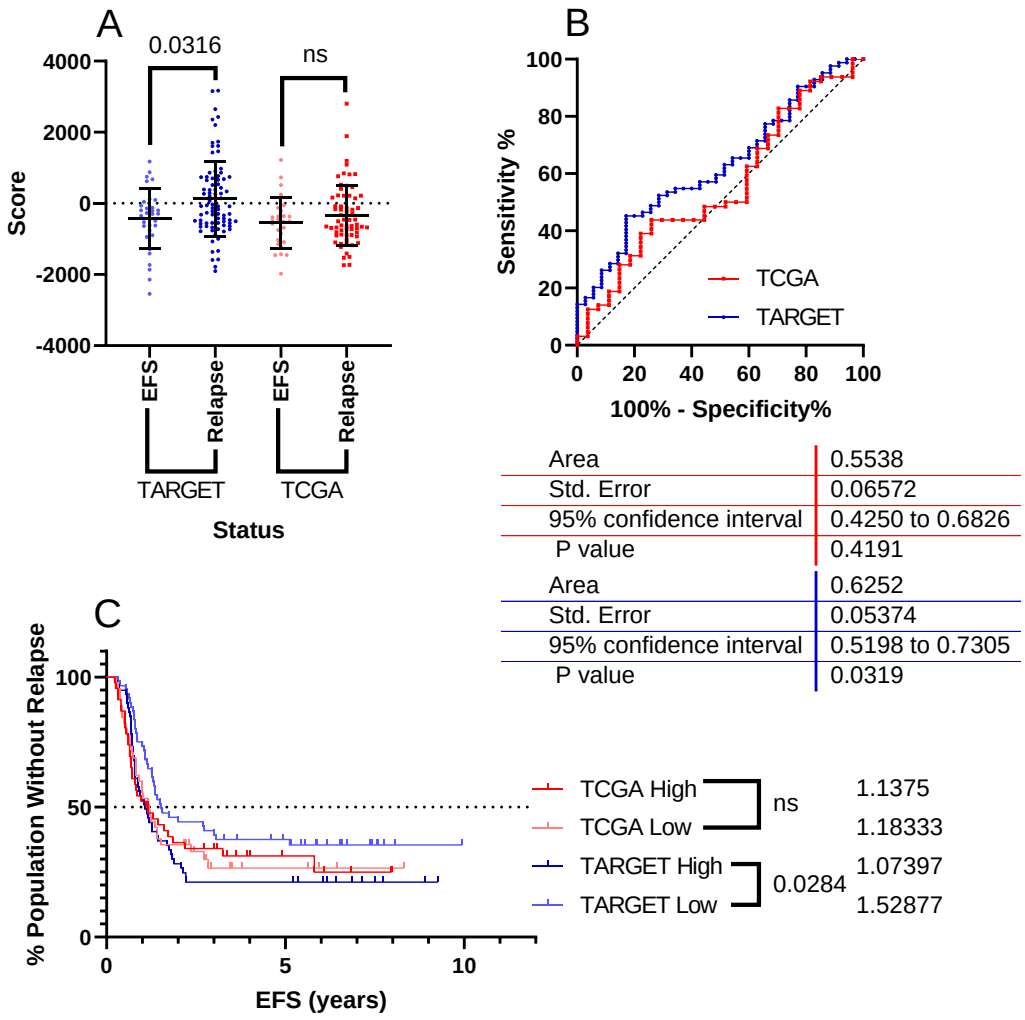

FIGURE 3.12. VALIDATION OF THE ANN SECOND WORKFLOW IN TCGA AND TARGET AML. THE SCORE WAS CALCULATED FOR THE RELAPSE AND EFS PATIENTS IN EACH DATASET. FIGURE A SHOWS THE DISTRIBUTION OF THE SCORE IN THE DATASETS SPLIT BY THE PATIENTS RELAPSE STATUS, TARGET AML SHOWED A SIGNIFICANT DIFFERENCE BETWEEN THE RELAPSE AND EFS GROUP WHEREAS TCGA WAS NOT SIGNIFICANTLY DIFFERENT. FIGURE B SHOWS THE ROC CURVE FOR THE SCORE IN THE DATASETS, THE AUC FOR TCGA WAS 0.5538 INDICATING AN ACCURACY OF 55.38%, WHEREAS TARGET HAD AN AUC OF 0.6252 INDICATING AN ACCURACY OF 62.52%. FIGURE C SHOWS THE KAPLAN-MEIER SURVIVAL PLOT SPLIT BY THE MEDIAN VALUE OF THE SCORE IN EACH DATASET, THERE IS NO SIGNIFICANT DIFFERENCE BETWEEN THE HIGH AND LOW GROUP IN TCGA, BUT THERE IS A SIGNIFICANT DIFFERENCE BETWEEN THE HIGH AND LOW GROUP IN TARGET AML.

In the HOVON dataset there is a significant difference between the EFS and Relapse patients, although the distribution of the scores does overlap largely in the centre of the range of scores (figure 3.11A). In contrast the score was only significantly different between relapse and EFS patients in the TARGET dataset, but not in the TCGA dataset (figure 3.12A), again with a large overlap of the scores between the groups. Looking at figure 3.11B, the HOVON dataset had an AUC of 0.7174, indicating a predictive value of 71.74%, in comparison the AUC for TCGA was 0.5538 and for TARGET 0.6252 (figure 3.12B). The AUC for the validation datasets is expected to be lower than that of the discovery dataset, due to the inherent differences between them. As the score is not significantly different between the EFS and relapse patients in the TCGA dataset the AUC is reflective of this.

In figure 3.11C there is a large separation between the high and low scoring groups for the Kaplan-Meier survival plot, which is significantly different. The low scoring group has an undefined median EFS, indicating the group did not reach 50% relapse; in contrast the median EFS for the high scoring group was 0.96, therefore the group reached 50% relapse in under one year. In figure 3.12C there is a significant difference between the high and low groups of the TARGET dataset, while both groups reached 50% relapse, the high scoring group reached this milestone sooner than the low scoring group by 0.45 of a year. The TCGA dataset showed no significant difference between the high and low scoring groups in figure 3.12C, where the median EFS was almost identical; the low scoring group had an EFS of 1.14 of a year and the high scoring group 1.18 of a year.

3.3.2.5. ROC

Statistics associated with ROC curves, including AUC, sensitivity and specificity were calculated for every gene in the HOVON dataset used the pROC package in r. The top 20 genes with the highest AUC were selected (supplementary table 9) and validated using a univariate BLR (supplementary table 10). All the top 20 genes were significant according to the univariate BLR therefor were used in the multivariate BLR, where the least significant gene was removed each time until all the genes were significantly contributing to the panel (table 3.10). Once the genes that contribute to the panel had been determined, clinical annotations were included in the multivariate BLR, only cytogenetic risk group significantly contributed to the panel (table 3.11).

Once the genes and clinical annotations significantly contributing to the panel had been determined, the score was calculated in the HOVON dataset. Although there was an overlap between the relapse and EFS patients, there was a significant difference between the score

in EFS compare to relapse patients (figure 3.13A), with the relapse patients having a much higher score on average compared to EFS patients. When looking at the score distribution in the validation datasets (figure 3.14A), there is more overlap between the relapse and EFS patients compared to the discovery dataset. The score is significantly different between the relapse and EFS patients in the TARGET dataset but not in the TCGA dataset.

	B	S.E.	Wald	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
HAL	-0.542	0.144	14.181	0.000	0.582	0.439	0.771
TMEM243	-0.455	0.134	11.542	0.001	0.634	0.488	0.825
CD109	0.231	0.098	5.543	0.019	1.260	1.039	1.527
NYNRIN	0.426	0.193	4.859	0.028	1.531	1.048	2.236
SOCS2-AS1	0.379	0.178	4.543	0.033	1.461	1.031	2.071

TABLE 3.11. MULTIVARIATE BLR FOR THE ROC WORKFLOW IN THE HOVON DATASET. THE GENES WHICH WERE SIGNIFICANTLY ASSOCIATED WITH RELAPSE AFTER THE UNIVARIATE BLR WERE TAKEN FORWARD TO THIS MULTIVARIATE BLR, THESE ARE THE REMAINING GENES THAT SIGNIFICANTLY CONTRIBUTE TO THE PREDICTIVE PANEL.

	B	S.E.	Wald	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
CD109	0.197	0.100	3.902	0.048	1.218	1.002	1.482
HAL	-0.484	0.147	10.899	0.001	0.616	0.462	0.821
NYNRIN	0.400	0.195	4.206	0.040	1.492	1.018	2.186
SOCS2-AS1	0.366	0.179	4.171	0.041	1.442	1.015	2.050
TMEM243	-0.429	0.135	10.048	0.002	0.651	0.500	0.849
Cytogenetic risk group	0.399	0.202	3.912	0.048	1.491	1.004	2.214

TABLE 3.12. MULTIVARIATE BLR INCLUDING THE CLINICAL ANNOTATIONS FOR THE ROC WORKFLOW. CLINICAL ANNOTATIONS WERE INCLUDED IN THE MULTIVARIATE BLR TO DETERMINE IF ANY SIGNIFICANTLY CONTRIBUTED TO THE PANEL. THIS IS THE COMBINATION OF GENES AND CLINICAL ANNOTATIONS WHICH SIGNIFICANTLY CONTRIBUTE TO THE PANEL IN THE HOVON DATASET.

When looking at figure 3.13B there is an AUC of 0.7347, indicating a predictive accuracy of 73.47%, a high performing panel in the discovery dataset. When comparing the predictive ability in the validation datasets, the AUC is similar for both datasets TCGA at 0.63 and TARGET at 0.65, however this is only significant in the TARGET dataset, which reflects the results shown in figure 3.14B. Figure 3.13C shows excellent separation between the high

and low score groups, where the median EFS is undefined for the low scoring group and 0.96 of a year for the high score group. The difference between the high and low groups in the HOVON dataset is significant, however only the TCGA dataset showed a significant difference in figure 3.14C. The result of TCGA showing significant differences in the group separation and TARGET showing no significant difference is in contrast to the result in figure 3.14A and B.

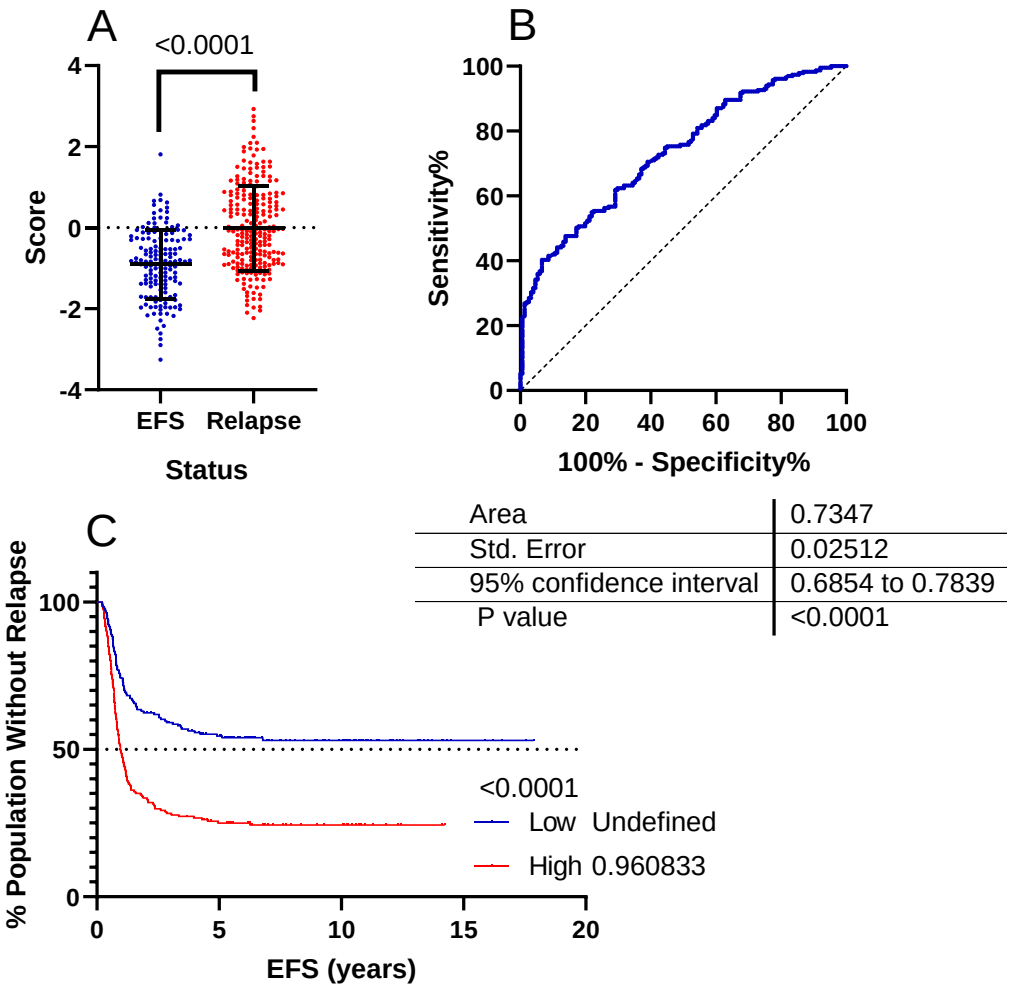

FIGURE 3.13. PREDICTIVE ABILITY OF THE ROC WORKFLOW IN THE HOVON DATASET. FIGURE A SHOWS THE DISTRIBUTION OF THE SCORE CALCULATED FOR EACH PATIENT WHICH WAS SPLIT BY THEIR RELAPSE STATUS, THE MEDIAN VALUE FOR RELAPSE PATIENTS IS SIGNIFICANTLY HIGHER THAN EFS PATIENTS. FIGURE B SHOWS THE ROC CURVE WITH AN AUC OF 07347, INDICATING A 73.47% ACCURACY. FIGURE C SHOWS THE KAPLAN-MEIER SURVIVAL PLOT SPLIT BY THE MEDIAN VALUE OF THE SCORE, THE GROUPS ARE SIGNIFICANTLY DIFFERENT, THE MEDIAN EFS FOR THE LOW SCORE GROUP WAS UNDEFINED WHEREAS THOSE WITH A HIGH SCORE HAD A MEDIAN EFS OF UNDER A YEAR.

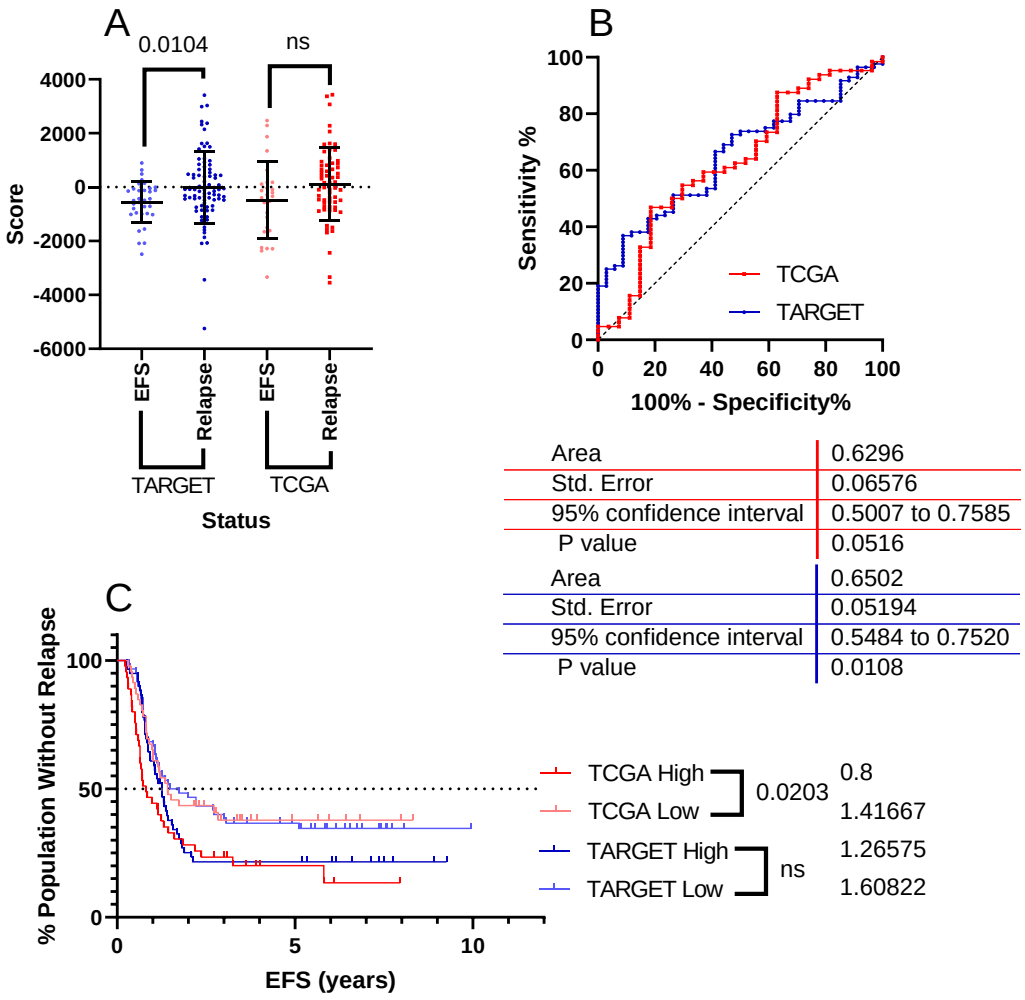

FIGURE 3.14. VALIDATION OF THE ROC WORKFLOW IN TCGA AND TARGET AML. THE SCORE WAS CALCULATED FOR THE RELAPSE AND EFS PATIENTS IN EACH DATASET. FIGURE A SHOWS THE DISTRIBUTION OF THE SCORE, SPLIT BY THE PATIENTS RELAPSE STATUS, IN EACH DATASET. THERE WAS A SIGNIFICANT DIFFERENCE BETWEEN THE EFS AND RELAPSE PATIENTS IN TARGET AML, WHERE THE MEDIAN SCORE WAS SIGNIFICANTLY HIGHER IN RELAPSE PATIENTS. THERE WAS NO SIGNIFICANT DIFFERENCE IN THE TCGA DATASET. FIGURE B SHOWS THE ROC CURVE FOR THE SCORE IN THE DATASETS, BOTH DATASETS HAS A SIMILAR AUC AROUND 0.63, INDICATING A 63% ACCURACY. FIGURE C SHOWS THE KAPLAN-MEIER SURVIVAL PLOT, WHERE THE PATIENTS WERE SPLIT BY THE MEDIAN SCORE OF THE DATASET, THERE IS NO SIGNIFICANT DIFFERENCE BETWEEN THE HIGH AND LOW GROUPS IN THE TARGET DATASET WHERE THERE IS A SIGNIFICANT DIFFERENCE IN TCGA.

3.3.2.6. BLR

Using the pROC package in r, statistics associated with ROC curves, including AUC, sensitivity and specificity were calculated for every gene in the HOVON dataset. There were 41 genes with an AUC above 0.6, these were selected and a BLR was performed (supplementary table 11), all except one gene was significant in the BLR. From the univariate BLR, the genes with the top 10 highest beta-values and the top 10 lowest beta-values were selected for the multivariate BLR, in each round the least significant gene was removed until all genes significantly contributed to the panel of genes (table 3.13). Once the multivariate BLR had determined the remaining 9 genes in the panel, the clinical annotations were included in the multivariate analysis, none of which significantly contributed to the panel.

	B	S.E.	Wald	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
PYROXD1	-0.768	0.203	14.330	0.000	0.464	0.312	0.691
FEM1C	-0.692	0.198	12.170	0.000	0.501	0.339	0.738
NYNRIN	0.639	0.219	8.528	0.003	1.895	1.234	2.911
MXRA7	0.410	0.159	6.686	0.010	1.507	1.104	2.056
AGTPBP1	0.444	0.180	6.070	0.014	1.559	1.095	2.219
LAPTM4B	0.161	0.067	5.794	0.016	1.175	1.030	1.340
MPO	-0.160	0.066	5.768	0.016	0.853	0.748	0.971
HAL	-0.373	0.175	4.555	0.033	0.689	0.489	0.970
SOCS2-AS1	0.368	0.184	4.016	0.045	1.445	1.008	2.072

TABLE 3.13. MULTIVARIATE BLR FOR THE BLR WORKFLOW IN THE HOVON DATASET. THE GENES WHICH WERE SIGNIFICANTLY ASSOCIATED WITH RELAPSE AFTER THE UNIVARIATE BLR WERE TAKEN FORWARD TO THIS MULTIVARIATE BLR, THESE ARE THE REMAINING GENES THAT SIGNIFICANTLY CONTRIBUTE TO THE PREDICTIVE PANEL.

The beta values from table 3.13 were used to calculate the score in the HOVON dataset, the distribution of the score between the EFS and relapse patients can be seen in figure 3.15A. The scores of the relapse patients are significantly higher compared to the EFS patients, however there is still an overlap between the two groups (figure 3.15A). In figure 3.16A there is only a significant difference between the EFS and relapse patients in the TCGA dataset and not in the TARGET dataset.

When looking at figure 3.15B, the AUC of this score is very high indicating an improved accuracy in differentiating between relapse and EFS patients compared to other methods discussed previously. In figure 3.16B the AUC for TCGA was 0.6759 quite high for the

validation datasets, especially compared to the AUC of the TARGET dataset of 0.5054, although the AUC is distinctly different between the validation datasets, this is indicating the differences between the two datasets overall.

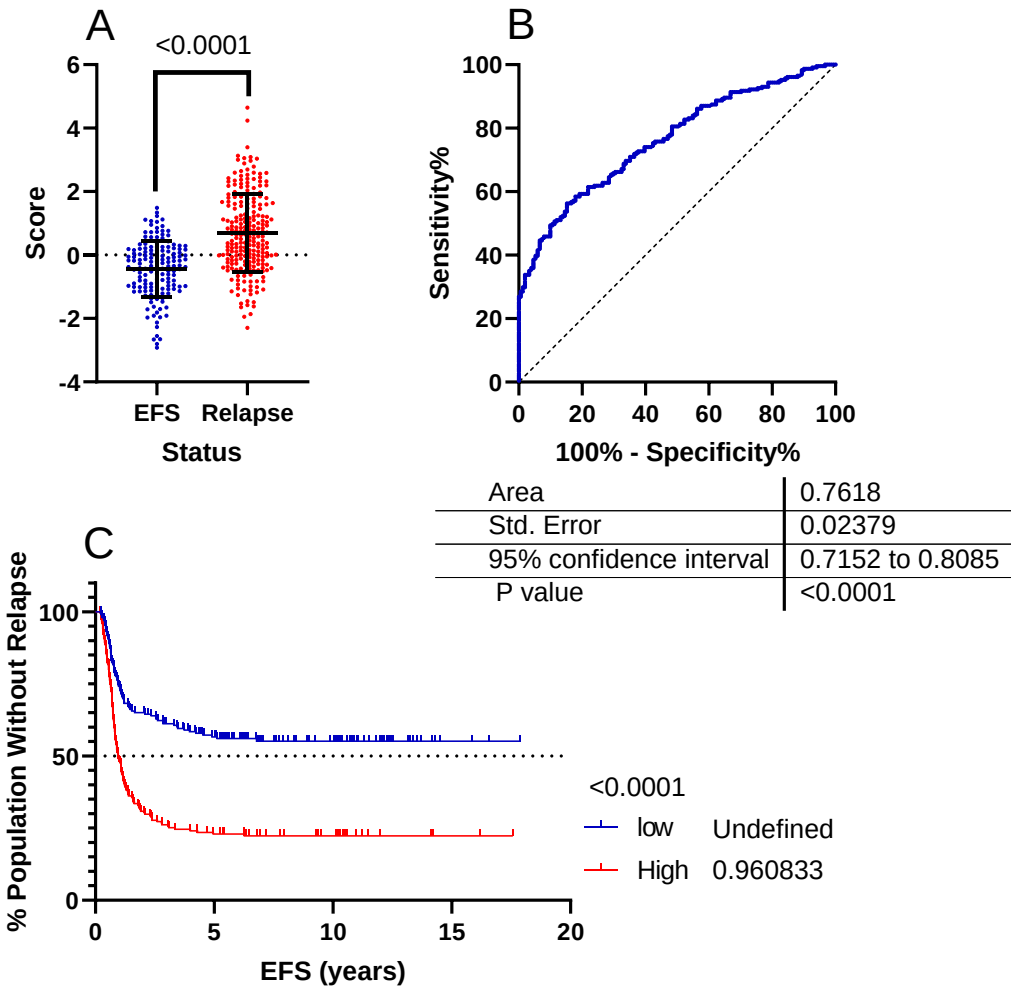

FIGURE 3.15. PREDICTIVE ABILITY OF THE BLR WORKFLOW IN THE HOVON DATASET. FIGURE A SHOWS THE DISTRIBUTION OF THE SCORE CALCULATED FOR EACH PATIENT WHICH WAS SPLIT BY THEIR RELAPSE STATUS, THE MEDIAN VALUE FOR RELAPSE PATIENTS IS SIGNIFICANTLY HIGHER THAN EFS PATIENTS. FIGURE B SHOWS THE ROC CURVE WITH AN AREA OF 0.7618, INDICATING A 76.18% ACCURACY. FIGURE C SHOWS THE MEDIAN SPLIT OF THE CALCULATED SCORE, THE GROUPS ARE SIGNIFICANTLY DIFFERENT WITH THE MEDIAN EFS FOR THE LOW SCORE GROUP WAS UNDEFINED WHEREAS THOSE WITH A HIGH SCORE HAD A MEDIAN EFS OF UNDER A YEAR.

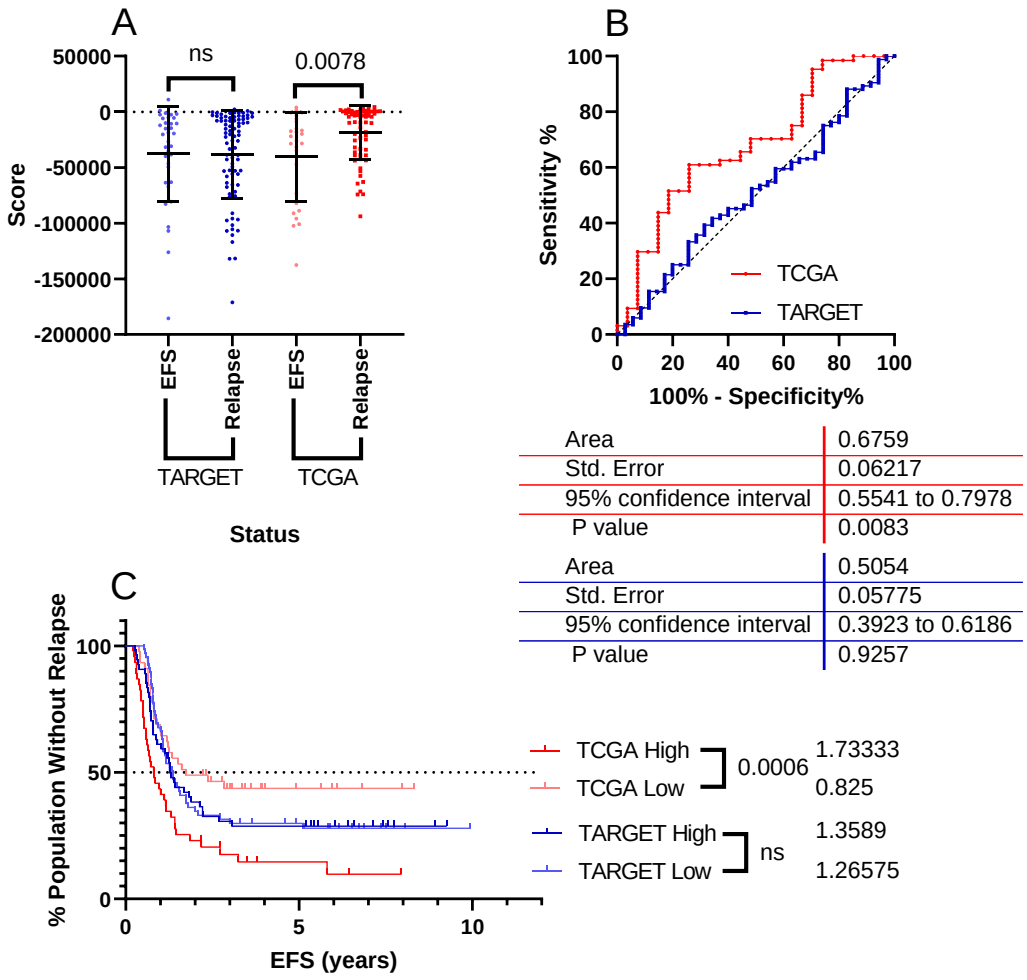

FIGURE 3.16. VALIDATION OF THE BLR WORKFLOW IN TCGA AND TARGET AML. THE SCORE WAS CALCULATED FOR THE RELAPSE AND EFS PATIENTS IN EACH DATASET. FIGURE A SHOWS THE DISTRIBUTION OF THE SCORE IN THE DATASETS SPLIT BY THE PATIENT RELAPSE STATUS, TCGA SHOWED A SIGNIFICANT DIFFERENCE BETWEEN THE RELAPSE OR EFS GROUP WHEREAS TARGET AML SHOWED NO SIGNIFICANT DIFFERENCE. FIGURE B SHOWS THE ROC CURVE FOR THE SCORE IN THE DATASETS, THE AUC FOR TCGA WAS 0.6759 INDICATING AN ACCURACY OF 67.59% COMPARED TO TARGET WITH AN AUC OF 0.5054, INDICATING AN ACCURACY OF 50.54%. FIGURE C SHOWS THE KAPLAN-MEIER SURVIVAL PLOT SPLIT BY THE MEDIAN SCORE VALUE IN EACH DATASET, THERE WAS NO SIGNIFICANT DIFFERENCE BETWEEN THE HIGH AND LOW GROUPS OF THE TARGET DATASET, BUT THERE WAS A SIGNIFICANT DIFFERENCE BETWEEN THE HIGH AND LOW GROUPS IN TCGA, WHERE THE HIGH GROUP HAD A MEDIAN EFS OF UNDER A YEAR WHERE THE LOW SCORE GROUP HAD AN EFS OF 1.7 YEARS.

In figure 3.15C there is an excellent separation between the high and low scoring groups. With the median free survival for the high group at 0.96, indicating that 50% of the group relapsed before a year and the low group with an undefined median free survival, indicating the group did not reach 50% relapse. In comparison, figure 3.16C, the separation for the high and low groups in TCGA dataset is distinct and significant, with the median EFS for the low score group of 1.7 of a year and for the high group of 0.825 of a year. The large difference between the median EFS indicates the score can separate the group well. In comparison the TARGET dataset shows no significant difference between the high and low scoring group with the median EFS almost the same at 1.36 for the high score group and 1.27 for the low score group.

3.4. Discussion and conclusions

3.4.1. Comparison of prognostic scores

When looking at summary table 3.14, although the ANN alone, ANN first and T-test method all performed well in the discovery dataset they were not significant in either of the validation datasets indicating a poor preforming score. Due to the lack of significance for these workflows they can be disregarded for the discovery of biomarkers in this instance, however, different datasets or a different clinical point for which biomarkers are required may be suitable for these workflows. Although they did not work for this investigation, the statistics and machine learning foundations are still valid and may be more applicable in different situations.

The ANN second, ROC and BLR workflows were significant in the discovery dataset and significant in one of the validation datasets. When looking at the ROC workflow, it significantly distinguished between EFS and relapse patients in the TARGET dataset but not in the TCGA dataset, and the inverse is true when separating high and low scoring patients with a significant separation in the TCGA dataset but not in the TARGET dataset. The contradiction in the results could be corrected with more refinement in the boundary of high and low score, however in this instance, the inconsistency means the results of this workflow are not suitable as a predictive biomarker panel.

The ANN second workflow was significant in the TARGET dataset, when looking at EFS versus relapse and high versus low score, but not in the TCGA dataset. The invers is true of the BLR workflow, which was significant in separating EFS and relapse patients as well as High and low scoring patients in the TCGA dataset rather than the TARGET dataset. Both workflows

have discovered a usable biomarker, and with further refinement they both could be useful predictors of relapse. Overall, the BLR workflow gave the best score as it showed higher significance when separating the EFS and relapse patients in the validation dataset, and the AUC was higher compared to the ROC workflow.

		ANN alone	ANN first	T-test	ANN second	ROC	BLR
HOVON	EFS vs Relapse	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
	ROC	0.7376	0.7259	0.7376	0.7174	0.7347	0.7618
	High vs Low	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
TCGA	EFS vs Relapse	NS	NS	NS	NS	NS	0.0078
	ROC	0.5799	0.6019	0.5712	0.5538	0.6296	0.6759
	High vs Low	NS	NS	NS	NS	0.0203	0.0006
TARGET	EFS vs Relapse	NS	NS	NS	0.0316	0.0104	NS
	ROC	0.5827	0.549	0.5418	0.6252	0.6502	0.5054
	High vs Low	NS	NS	NS	0.0284	NS	NS

TABLE 3.14. SUMMARY OF RESULTS OF DIFFERENT METHODS OF BIOMARKER DISCOVERY. THE RESULTS FROM EACH OF THE VALIDATION GRAPHS IN ALL THREE DATASETS ARE SUMMARISED.

The elements involved in the BLR workflow can all be written into a single streamline script in r, which would ensure its correct application and the reproducibility. As each element is processed through computational methods, human error can be avoided improving the reproducibility and the consistency of the panels being developed in the future. Unlike other workflows, the BLR workflow relies heavily on transparent and trusted statistical methods, which allows for a wider audience to understand the workflow and apply it in a clinical setting.

The score discovered in the BLR workflow was the largest of the workflows with 9 genes as displayed in table 3.15. The common genes that regularly appeared in the workflows are highlighted in table 3.15, with SOCS2-AS1 appearing in 5 finalised panels, where NYNRIN and

CD109 appear in 4 panels. Interestingly, CD109 was also included in the ANN first workflow and HAL was included in the ANN second workflow before clinical annotations were introduced.

ANN Alone	ANN first	T-test	ANN second	ROC	BLR
CD109	SOCS2-AS1	CD109	CD109	CD109	SOCS2-AS1
SOCS2-AS1	HAL	NYNRIN	SOCS2-AS1	SOCS2-AS1	HAL
ACVR1B	NYNRIN	FUT4	ACVR1B	HAL	NYNRIN
DNMT3B	TES			NYNRIN	MXRA7
ADPGK				TMEM243	AGTPBP1
SLC31A2					LAPTM4B
MSI2					MPO
					PYROXD1
					FEM1C
	Cytogenetic risk group	Cytogenetic Risk Group	Cytogenetic Risk Group	Cytogenetic risk group	

TABLE 3.15. SUMMARY TABLE OF THE SCORES. PRIOR TO THE INTRODUCTION OF CLINICAL ANNOTATIONS CD109 WAS INCLUDED IN THE ANN FIRST METHOD AND HAL WAS INCLUDED IN THE ANN SECOND METHOD

As noted previously there are common genes which appear in several panels, the first is CD109 which has the highest AUC (0.64) from every gene in the HOVON dataset (supplementary table 9), indicating that individually this gene could distinguish between relapse and EFS patients with 64% accuracy. CD109 protein is a glycosyl-phosphatidylinositol linked glycoprotein that localizes to the surface of cells, which binds to and negatively regulates signalling by transforming growth factor beta[144], the increased expression of CD109 has been linked to negative outcomes across different cancers[139,140,187,188]. Although this gene did not appear in the score it is clearly an integral part of cancer progression.

As for the genes in the score, SOCS2-AS1 is the only RNA gene, meaning it is not translated into protein. This gene is a long non-coding RNA (lncRNA), on the antisense strand of SOCS2,

which interestingly was not predictive in relapse. SOCS2-AS1 has already been associated with apoptosis inhibition and cellular growth in prostate cancer[189], although has also been linked to inhibition of endometrial cancer[190] and colorectal cancer[191]. It is thought to work as a regulator by inhibiting miR-1264 thus promoting the expression of SOCS2[191], in addition it has been found to bind to and promote the degradation of Aurora kinase A[190], a potent cell cycle regulator.

Histidine ammonia-lyase (HAL) is a protein coding gene, it is a cytosolic enzyme associated with the degradation of L-histidine to trans-urocanic acid, it has no current association to cancer but may increase the sensitivity of cancer cells to certain chemotherapy drugs[192], in addition, histidine as a dietary supplement showed promise in mouse models[193] as a way to improve clinical outcome.

High expressions of NYNRIN are a linked to poor overall survival in AML[194]. Although decreased expression has been shown to increase invasion by melanoma cells and high expression increases overall survival[195]. Mutations in NYNRIN have been associated with a predisposition to Wilms tumours[196] and relapsed ALL[197]. Although this gene and the associated protein are not widely studied.

The MXRA7 gene has not been widely studied but it codes for a transmembrane matrix remodelling protein, the gene has been linked to several different cancers using a datamining approach[198], and an increased levels of the protein is associated with minimal residual disease in AML[199].

The gene AGTPBP1 codes for a protein that catalyses the deglutamylation of post-translational modified proteins and removes polyglutamate chains from the carboxy-terminus of proteins, this is seen particularly in tubulin. A decreased expression was found in lung cancer compared to the surrounding normal tissue, which has been linked to increased motility and proliferation in cellular studies[200].

The gene LAPTM4B codes for a multifunctional protein, which is crucial to lysosomal function including: EGFR degradation, ubiquitination inhibition, recruitment of Leucine transporters to the lysosome, activation of ATPase protein pump and negatively regulates TGFβ1 production in regulatory T cells. Mutations in the gene have been associated with cancer susceptibility[201] and poor prognosis due to increased proliferation and apoptosis resistance. An increase in the gene has been associated with autophagy[202] in cancer cells, this provides

them with a resistance to metabolic stress, allowing increased proliferation. Additionally, an increased expression of this gene has been associated with chemotherapeutic resistance[203].

The gene MPO encodes the protein myeloperoxidase, a lysosomal enzyme involved in the antimicrobial response in neutrophils. The degranulation of neutrophils releases MPO into the extracellular space to combat pathogens. The reactive oxygen species released cause inflammation and could cause DNA damage in near cells. The increased presence of neutrophils and MPO has been associated with colorectal cancer[204], lung cancer[205] and breast cancer[206,207], mutations that reduce MPO production decrease the risk of breast cancer. This is thought to be because of the reduced oxidative stress and inflammation.

Upregulation of the gene PYROXD1 has been found in colorectal cancer[208] compare to surrounding normal tissue, although the gene is not widely studied in cancer. The protein functions as a pyridine nucleotide-disulphide reductase, with C-terminal nitrile reductase domain, and is involved in the oxidative stress response.

The downregulation of FEM1C in colorectal cancer has been associated with poor prognosis[209]. Although the processes of FEM1C are largely unknown it is through to play a role in post-transcriptional modification of RNA[210] and the post-translational modification of proteins[211]. Each one of the genes discovered in this score show a relationship with cancer and the cell cycle, these all have the potential to be markers for diagnostics and targets for therapeutics.

3.4.2. Comparison of the datasets

Each dataset in this investigation has its own unique qualities, due to this there will be variation between the score performance within each dataset. For example, TARGET AML is a childhood dataset, whereas HOVON and TCGA are both adult datasets, there are inherent differences between the treatment and survivability of the AML in children compared to adults. These differences are difficult to control and the impact of which is not entirely obvious in the results but is likely to have an impact. In addition, TCGA and TARGET are both RNA sequence datasets, whereas HOVON is an Affymetrix dataset, as such the range of results in addition to the expression of certain genes will differ widely in the Affymetrix data compared to the sequence data. These differences are apparent and are reflected in the results, the range for the scores is typically in the 1000's for the sequence data and in the 10's for the Affymetrix data.

In the workflows discussed the differences between the datasets were not controlled, resulting in the large differences in ranges, the lack of significance for certain workflows and the lower separation of patients in the validation datasets. By making the ranges and distribution as equal as possible, the results will likely be more comparable and have a wider application.

To correct for the differences between the datasets an additional normalisation step is required, this takes all the datasets and attempts to equalise the range of the data. This would make the score more comparable between the platforms and potentially applicable across platforms. There are several ways to achieve cross-platform normalisation, although not all are comparable in their efficiency. Min-max normalisation is used post discover as a penultimate step before score calculation. After a final gene list is obtained, the gene expression values are used along with the minimum and maximum value of the specific gene, within the dataset. The gene expression values are scaled to be between 0 and 1, by using the following equation, where the minimum represents the minimum gene expression of the gene within the whole dataset, and similarly with the maximum:

$$\frac{gene\ value - minimum}{maximum - minimum}$$

Although this method would ensure the range is the same between datasets, it wouldn't normalise the distribution of the data. The data in the HOVON dataset uses the whole range with a central median value, which is not seen the TCGA or TARGET dataset where the results are skewed due to the baseline value of zero.

A different method of normalisation is using a housekeeper gene to calculate fold-change. Potential housekeeper would have to be consistently expressed across all datasets with little variance between relapse and EFS patients, the commonly used housekeepers from Qiagen[212] are an ideal starting point for this. Finding housekeepers that are consistent between conditions and between datasets can be difficult, in addition the fold change can be skewed easily if the gene is reduced compared to the housekeeper, although this can be corrected using a log2 fold change. Although again the range of the gene expression can have a large impact where the RNA sequence datasets are into the 1000s with gene expression the Affymetrix datasets are in the 10s, this large difference will be reflected in the fold change.

Another method of normalisation between different datasets is a workflow described by[213] which can be adapted as shown in figure 3.19, this workflow merges datasets from different technologies and normalises the datasets universally, making the technologies directly

comparable. This normalisation can be achieved using the "*NormalizeBetweenArrays*" function from the *limma*[214] package in R, selecting the a-quantile normalisation setting. The benefit to this method is the results from discovery dataset can be directly compared to the validation datasets, however this also means the results should be deconvoluted after to provide a broader perspective in the original dataset.

All three methods have advantages and disadvantages, although, the method that would provide the best overall normalisation, is the a-quantile normalisation between arrays. If the results were directly comparable between datasets it would be easier to find genes that were characteristic of relapse and remove those that weren't. These will all be considerations for future biomarker discover and can be used as an improvement on the workflows described here.

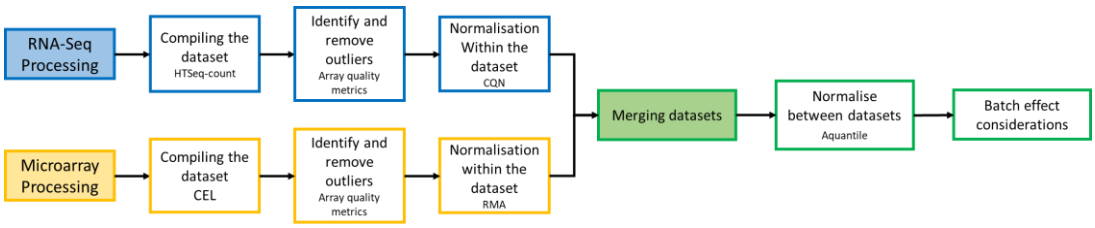

FIGURE 3.17. NORMALISE BETWEEN ARRAYS WORKFLOW. EACH DATASET IS INDEPENDENTLY COMPILED, AND QUALITY CHECKED BEFORE NORMALISATION WITHIN THE DATASET. THE DATASETS ARE MERGED AND NORMALISED TOGETHER USING A-QUANTILE NORMALISATION.

Chapter 4 - Impact of CD109 in AML

4.1. Introduction

As discussed in chapter three the relapse gene signature was discovered using a pipeline of different bioinformatics techniques. The aim of the prognostic score was to predict which patients were likely to relapse, in a clinical setting this would allow for closer monitoring and potentially use of alternative treatment strategies. During the analysis, CD109 was identified as the gene with the highest area under the curve (AUC) (supplementary table 9), indicating it could singly predict relapse in AML. The gene has independently been associated with relapse and poor prognosis in different cancers[139,140,164,187,215,216], all showing CD109 has an important role in cancer progression and survival. In breast cancer CD109 has been associated with poor overall survival and carcinogenisis[139,140,164,187,215,216]. Interestingly CD109 protein expression has been found in basal-like breast carcinomas, but not in non-basal-like carcinomas[139]. In addition, the expression of CD109 was associated with reduced fat invasion in the basal-like carcinoma subset[139]. In the epithelium of the oral cavity, increased expression of CD109 has been found in squamous cell carcinoma and precancerous legions, but normal epithelium did not highly express CD109, indicating its presence is associated with carcinogenesis and progression of pre-cancerous legions[140].

In AML it has been established that CD109 is associated with poor overall survival[217] and when combined with CALCRL and LSP1 forms a prognostic panel allowing for the prediction of patient outcome[164], additionally the gene been associated with resistance to induction chemotherapy[218]. Which is a key part in the clearance of AML and the ability to effectively treat patients.

The mechanisms of CD109 transcript and the associated protein are somewhat unknown; however, the protein is a co-recepter[144,219] for TGFBR1, which promotes the proteolytic degradation of TGFBR1[144,145,147] reducing the downstream signalling. There is also some evidence CD109 sequesters[147] TGFβ1 before it reaches the TGFBR1, again reducing the signalling potential of this pathway. Interestingly, a TGFβ1 neutralising antibody was shown to enhance apoptosis in AML by cytarabine[220], which is contrary to CD109, in that a reduction in the TGFBR1 protein reduces the treatment sensitivity of AML in patients. As CD109 was consistently found to be highly associated with relapse in AML, and its association to other cancers, this gene was investigated further in the context of treatment response and survival.

It is expected that a reduction in CD109 expression will increase cellular response to TGFβ1 and cytarabine induced cell death, which will be used to treat shRNA Knockdown cell lines.

4.2. Methods and materials

4.2.1. Tissue culture maintenance

Cells were grown at 37°C in a humidified atmosphere with 95% air and 5 % CO_2. The medium used was specific to the cell line, outlined in table 4.1. All centrifugation was performed at room temperature, 300xG for 5 minutes. Cells were counted using the "nucleocounter" with solution-18, as per the manufacturer's instructions.

Adherent cells were passaged when 80-90% confluent. Medium was discarded, and the flask was washed with PBS. A solution of 0.05 % trypsin and 0.02 % EDTA was added to the flask to cover surface. Flasks were incubated until cells had completely detached and were suspended individually. The cell suspension was transferred to a falcon tube, and complete growth medium was added before centrifugation. The supernatant was discarded, and cell pellet was resuspended in an appropriate volume of complete growth medium before counting and seeding at the required density.

Suspension cells were passaged as required by transferring the suspension to a falcon tube, which was counted before centrifugation. The supernatant was discarded, and cell pellet was resuspended in an appropriate volume of complete growth medium for the required seeding density.

Cell line	Growth medium requirements
HEK293	DMEM + 10% FCS + 1% L-Glutamine
MDA-MB-468	DMEM + 10% FCS + 1% L-Glutamine
THP-1	RPMI-1640 +10% FCS + 1% L-Glutamine
MolM-13	RPMI-1640 +10% FCS + 1% L-Glutamine
Kasumi-1	RPMI-1640 + 20% FCS + 1% L-Glutamine

TABLE 4.1. GROWTH MEDIUM REQUIREMENTS OF THE CELL LINES USED. THE SPECIFIC MEDIUM REQUIREMENTS AND ARE DEPENDENT ON CELL LINE, WHICH ARE OUTLINED IN THIS TABLE.

4.2.2. Quantitative PCR

Cells were harvested at a density of $1_{x10}{}^{6}$cells/ml, a 1mL sample was centrifuged, and the supernatant was removed. The cell pellet was resuspended in PBS to wash and was centrifuged, the supernatant was removed. The RNA was extracted using an RNeasy kit from Qiagen and quantified using the nanodrop. From the RNA, cDNA was synthesised with oligo-dt_{15} using up to 2000ng of RNA and Promega MMLV reverse transcriptase.

The Qiagen rotorgene quantitative PCR machine was used to perform RT-qPCR. The reagents required for successful qPCR include: SYBR green PCR master-mix, making up half of the total reaction volume, 0.5µL of both forwards and reverse primers (10 pMole/uL) per reaction, nuclease free water to the volume required and 1 µL of cDNA sample. The temperature profile used began with a 95°C hold for 1 minute, followed by a cycle of 95°C for 20 seconds, 58°C for 15 seconds and 72°C for 20 seconds. To obtain a melt curve the temperature ramped from 58°C through to 95°C.

4.2.3. Viroid construction

The viral particles were constructed in the HEK239 cell line, which contains the gene for adenovirus, using the optimised µg ratio of 8:6:2 for the shRNA plasmid, the packaging plasmid and the envelope plasmid. The cells were seeded in t25 flasks and allowed to adhere and grow to 70% confluence. To the flask 8µg of plasmid of interest, 6µg of packaging plasmid and 2µg of envelop plasmid were added to the flasks with 17µl lipofectamine p3000 and 32µl p3000 reagent, in 5ml of FCS free optiMEM medium. The flasks were incubated at 37°C with 5% CO_2.

Twenty-four hours after HEK239 cells were transfected, the medium was removed, the cells were washed with PBS and fresh growth medium was added. After forty-eight and seventy-two hours of the initial transfection, the medium containing the constructed virus was harvested. The harvested medium was centrifuged at 300xg for 5 minutes to remove large debris and cells, the supernatant was filtered using a 0.2-micron filter, removing any remaining cells and smaller debris, preventing contamination of cell lines. After the first harvest 5mL of fresh HEK293 medium was added for fraction-2 collection. The fractions were frozen at -80°C until required.

4.2.4. Viral transduction and spin-fection

To transduce adherent cells, they were seeded and grown to 70-80% confluent. Medium was removed and cells were treated with Polybrene diluted in medium, at a concentration of 10µg/µL. After a short incubation, the viral stock was added to the test wells and medium was added to the control wells. After 24 hours, spent medium was removed and replaced with fresh medium containing Puromycin at a concentration determined by a drug titration, which will kill the non-transfected cells.

Cells in suspension were centrifuged at 300xg for 5 minutes, spent medium was discarded and cells were resuspended in 1mL of fresh medium containing polybrene at a concentration of 10µg/µL. After a short incubation with the polybrene medium, the viral stock was added to the test cells and medium was added to the control cells. After a short incubation, the cell suspensions were centrifuged at 800xg for 30-45 minutes at 32°C, the supernatant was removed, and cells resuspended in 3mL of fresh medium. Once the cells had recovered, usually 24-48 hours, Puromycin was added to the wells at a concentration determined by a drug titration (1ug/mL), to remove the cells that do not contain the plasmid.

4.2.5. XTT cell proliforation assay

Cell lines were seeded at $1x10^6$/mL, the relevant treatment was added where applicable, whether that be cytarabine, Puromycin, TGFβ1 or no treatment, and control wells containing media with the corresponding treatment were also plated, the plates were incubated for 24 hours. After the incubation, the suspension was mixed and 100µL was transferred to a 96-well plate. A stock of XTT working reagent was made by combining 20µL/mL of electron coupling reagent and 5mL of the XTT Labelling reagent, which was then vortexed. 50 µL of the XTT working reagent was added to the wells of the 96-well plate, which were mixed by pipetting and then incubated for 4 hours. After the incubation the plates were mixed by tapping, any air bubbles were removed before the plates were read at 470 nm with a reference wavelength of 650 nm in a bioRad iMark microplate reader.

4.2.6. NanoString

Cell line samples were harvested, and the RNA was extracted as described above in "Quantitative PCR". The RNA concentration was measured with the Nanodrop 8000 and diluted to a working stock concentration of 70ng/µL. the concentration of the samples were confirmed using the Nanodrop 8000 in triplicate. 150ng of RNA was pipetted into the strip

tubes provided in the kit, water was added to equalise the volume to 5µL before the reporter and capture probes were added. The cancer pathways panel, containing 770 genes, was used in this investigation. The tubes were mixed by tapping then centrifuged to collect the mixture in the bottom of the tubes which were then incubated at 65°C for 24 hours in a thermal cycler with a heated lid. After incubation the strip tubes were centrifuged briefly to collect the mixture in the bottom of the tubes, then loaded into the prep-station with the reagent plates and consumables, as per the manufacturer's instructions. After the prep station had loaded the samples onto the chip, it was sealed and transferred to the nCounter, a High-resolution scan (fov 555) performed to acquire the digital count, and the RLF file "NS_CancerPath_C2535" used to generate the data. The output files (.RCC) were quality controlled using the nSolver software (Version 4.0). Differential expression and further analysis performed within the software.

4.3. Results

4.3.1. Drug titrations

4.3.1.1. Puromycin titration

Puromycin is an amino nucleoside antibiotic, an analogue of the 3' terminal end of aminoacyl-tRNA. Puromycin incorporates itself into a growing polypeptide chain and causes its premature termination, thereby inhibiting protein synthesis. Human cells have some tolerance to Puromycin, although when the threshold is reached it becomes toxic to cells apoptosis is initiated. In cell lines transduced with the PLKO.1 plasmid, Puromycin resistance is conferred on the plasmid, so Puromycin is transported out of the cell before it reaches the threshold. To determine the natural resistance of the cell line a titration was performed, this effectively determines the concentration required to remove non-transduced cells, leaving only cells transduced with the PLKO.1 plasmid.

As shown in figure 4.1, cell lines grow exponentially without the inhibition of Puromycin, however with the introduction of a small concentration of Puromycin the cell growth was halted. Although cell growth was stopped after just 0.5µg/mL of Puromycin it is important to see this in conjunction with cell viability, as the aim is to remove all sensitive cells. Ideally the concentration should kill the sensitive cells within 2-3 days, allowing for transduced cells to grow without contamination from non-transduced cells. In figure 4.2, the cell viability for the titration is displayed, which shows as the concentration of the Puromycin increases the cell viability decreases. Taking figure 4.1 in conjunction with figure 4.2, we see that cell cycle

arrest is induced with a small concentration of Puromycin and there is minimal effect of increasing the concentration on further population decrease (figure 4.1), however looking at the cell viability (figure 4.2) shows that the concentrations not only stop cell division but actively reduce the viability of those cells. To achieve the removal of Puromycin sensitive, non-transduced, cells from a population the cell viability should be reduced as well as the cell population. All measurements were taken using the NucleoCounter, although this method is reliable and consistent, it anecdotally has a tendency to over-estimate the cell population.

The concentration of Puromycin required to kill the cells is different depending on the cell line, although the graphs show that many of the concentrations used would be suitable to clear sensitive cells, it is important to look at the health of the cells and visualise which concentration causes the cells to begin apoptosis in a short time scale. Figure 4.3 shows photographs of the line MolM-13 at x20 magnification, treated with different concentrations of Puromycin. The photographs demonstrate that the cell population increases over time in the condition without Puromycin, in contrast with the conditions treated with Puromycin where there is a reduction in the cell population and a decreased cell viability, shown by the granular appearance of the cells and the reduced cell number. Similarly, in figure 4.4 and figure 4.5, the THP-1 and Kasumi-1 cell lines follow the same trend as in figure 4.3. Looking at the results in combination, each cell line requires a different concentration of Puromycin to be effective, the concentrations chosen for MolM-13, THP-1 and Kasumi-1 were 1µg/mL, 3µg/mL and 1µg/mL respectively. The concentrations were chosen based on the reduced cell viability and evidence of cell clearance within 3 days of treatment.

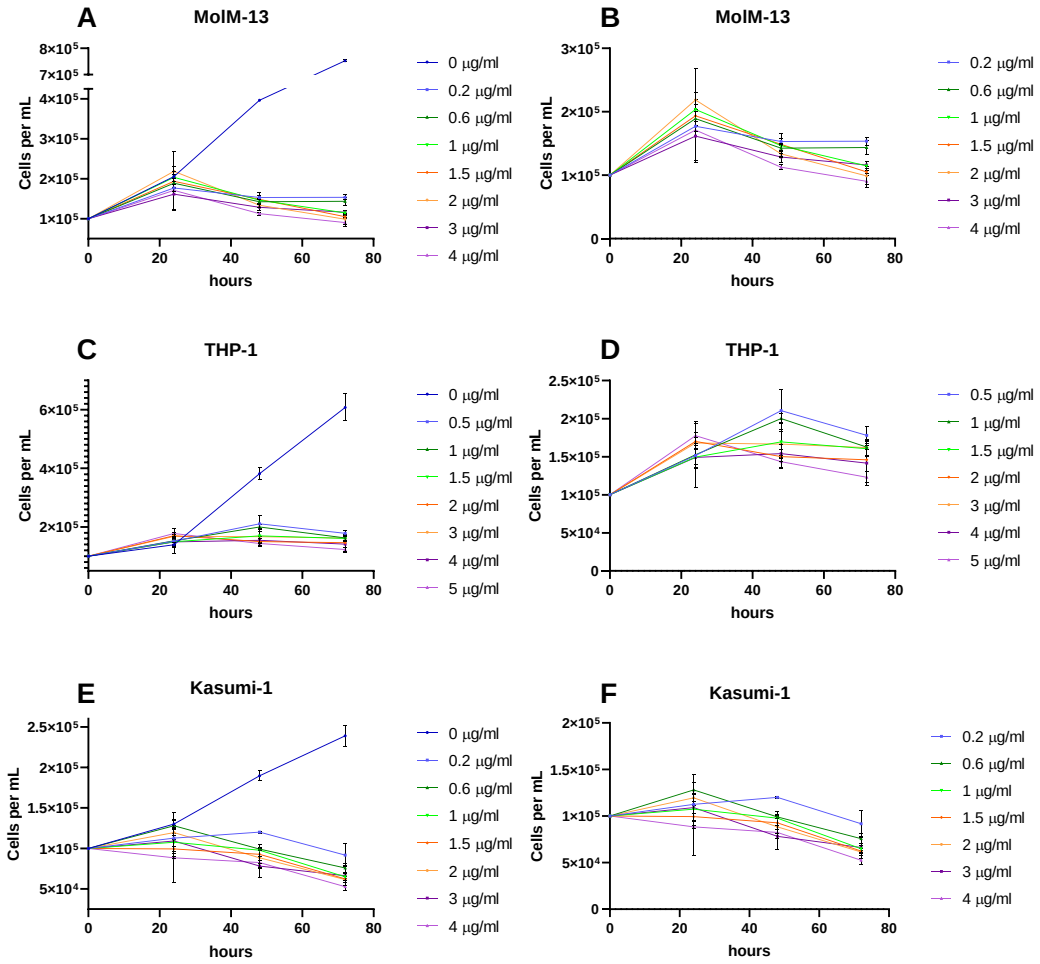

FIGURE 4.1. PUROMYCIN TITRATION AND CELL GROWTH IN AML CELL LINES. THE MINIMUM CONCENTRATION OF PUROMYCIN REQUIRED TO KILL THE CELLS WAS DETERMINED BY COUNTING THE CELLS AT 24-HOUR INTERVALS. FIGURES A, C AND E SHOW THE FULL RANGE OF CONCENTRATIONS USED WITH A COMPARISON TO THE CELL GROWTH WITHOUT PUROMYCIN. FOR CLARITY, FIGURES B, D AND F SHOW THE TITRATION WITHOUT THE UNTREATED RESULT TO ALLOW FOR A MORE DETAILED VIEW OF THE RESULTS. FIGURES A, C AND D SHOW EXPONENTIAL GROWTH OF THE UNTREATED CELL LINES, COMPARED TO ALL CONDITIONS WITH PUROMYCIN. N=3

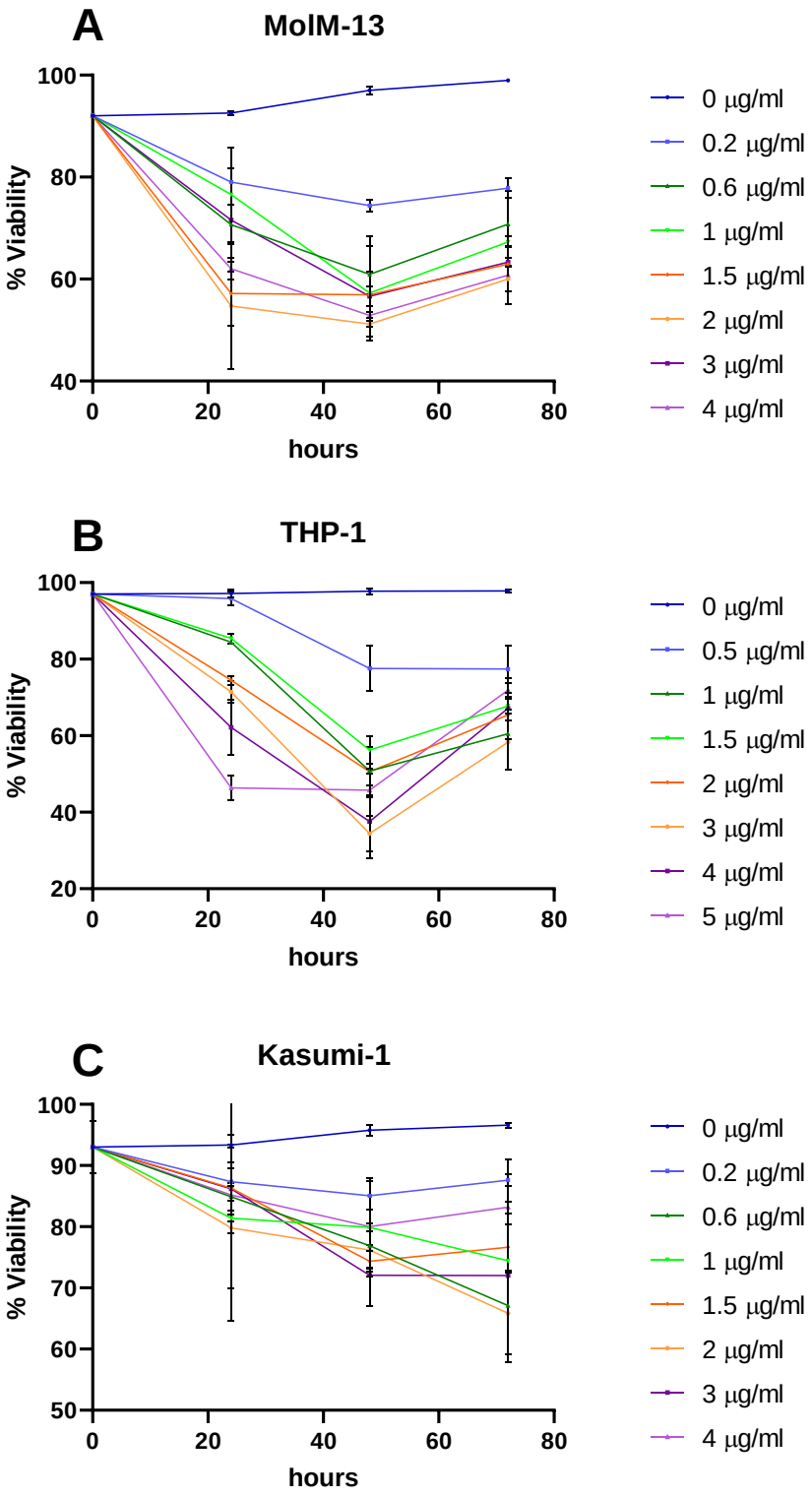

FIGURE 4.2. PUROMYCIN TITRATION AND CELL VIABILITY OF AML CELL LINES. EACH CELL LINE WAS TREATED WITH PUROMYCIN AT DIFFERENT CONCENTRATIONS WITH AN AIM TO REMOVE ALL SENSITIVE CELLS. THE CELL VIABILITY IS CALCULATED AS A THE % OF LIVE CELLS FROM THE WHOLE CELL POPULATION INCLUDING THE DEAD CELLS. N=3

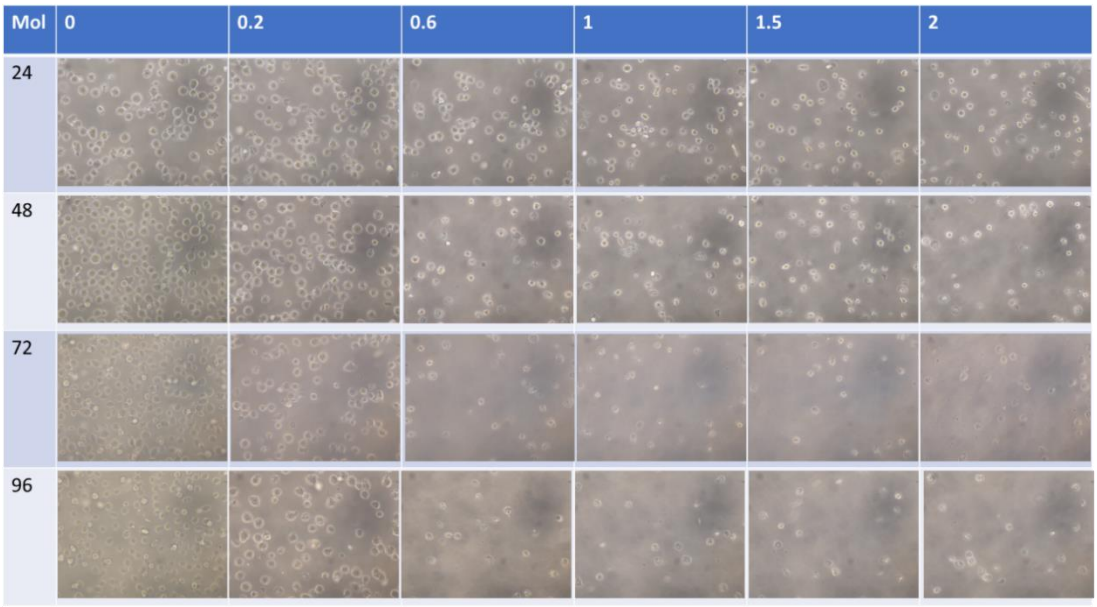

FIGURE 4.3. PHOTOGRAPHS AT X20 MAGNIFICATION FOR MOLM-13 CELLS TREATED WITH PUROMYCIN. TIME IN HOURS IS DISPLAYED ON THE LEFT AND PUROMYCIN CONCENTRATION IN µG/ML IS DISPLAYED AT THE TOP. CELL POPULATIONS IN THOSE TREATED WITH PUROMYCIN DECREASE OVER TIME, WHEREAS THE UNTREATED CELLS GREW EXPONENTIALLY

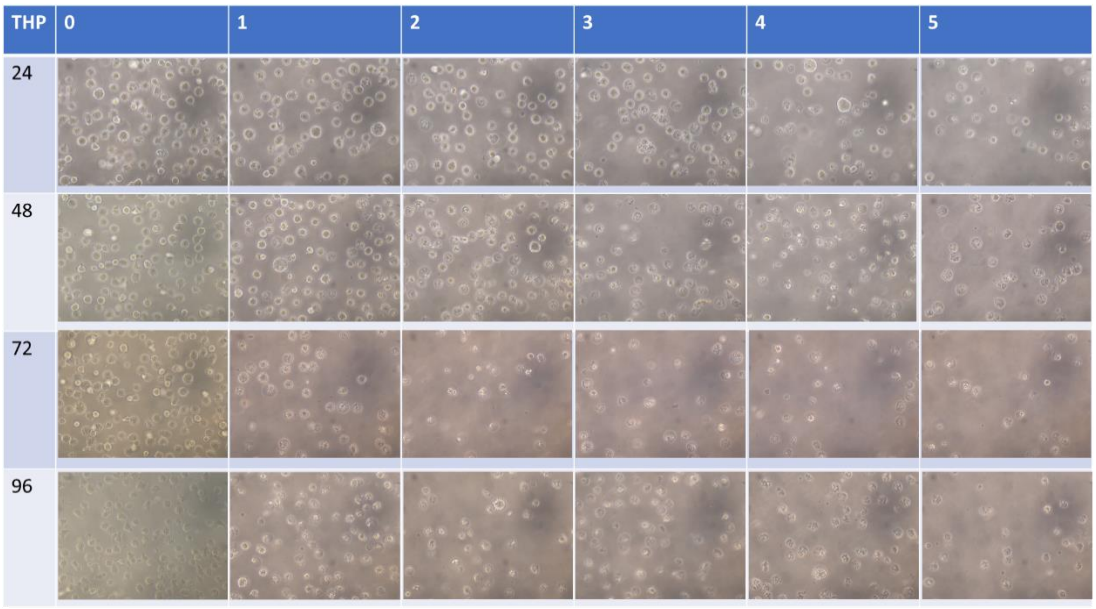

FIGURE 4.4. PHOTOGRAPHS AT X20 MAGNIFICATION FOR THP-1 CELLS TREATED WITH PUROMYCIN. TIME IN HOURS IS DISPLAYED ON THE LEFT AND PUROMYCIN CONCENTRATION IN µG/ML IS DISPLAYED AT THE TOP. CELL POPULATIONS IN THOSE TREATED WITH PUROMYCIN DECREASE OVER TIME, WHEREAS THE UNTREATED CELLS GREW EXPONENTIALLY.

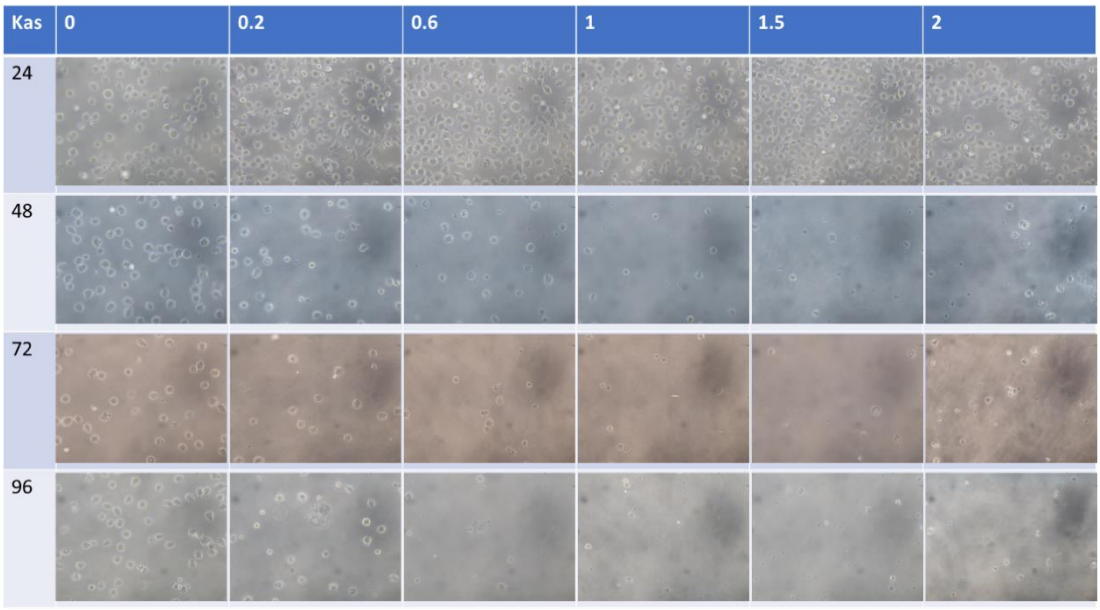

FIGURE 4.5. PHOTOGRAPHS AT X20 MAGNIFICATION FOR KASUMI-1 CELLS TREATED WITH PUROMYCIN. TIME IN HOURS IS DISPLAYED ON THE LEFT AND PUROMYCIN CONCENTRATION IN µG/ML IS DISPLAYED AT THE TOP. CELL POPULATIONS IN THOSE TREATED WITH PUROMYCIN DECREASE OVER TIME, WHEREAS THE UNTREATED CELLS GREW EXPONENTIALLY.

4.3.1.2. Cytarabine titration

Cytarabine is a cytosine analogue which competitively binds to the corresponding purine during S-phase[221]. When incorporated into the DNA structure it sterically hinders the rotation of the molecule within DNA, in addition it inhibits alpha-DNA polymerase and DNA repair through beta-DNA polymerase. These compounding factors mean DNA replication ceases, and cell cycle arrest occurs specifically in S-phase[221]. Cytarabine is used in combination with Daunorubicin to treat AML in a regime known at 7+3 where 7 days of Daunorubicin is given followed by 3 days cytarabine[222–224].

To determine which concentration of cytarabine was effective in each of the AML cell lines a titration was performed. As there was no existing data on appropriate concentrations to use in cell lines, a wide scale titration was performed with ranges from 1nM to 1mM. To determine the effect of cytarabine on the cell lines the cell density was recorded at 24-hour intervals, showing the growth rate of the cell lines at the different concentrations. The initial titration in figure 4.6A and 4.7A was used to target a more selective range for the cell line. The growth rate for cells without cytarabine is exponential, with the addition of 1nM of cytarabine there was very little effect on the growth rate. The condition with 100nM of

cytarabine inhibited the growth rate of the cell lines noticeably, and the conditions with 1µM, 100µM and 1mM induced complete cell cycle arrest. Interestingly, in Figures 4.6B and 4.7B the condition with 1 µM of cytarabine shows cell cycle arrest but the cell viability remains high, this is important to consider for the next titration as the goal is to induce cell cycle arrest with a minimal concentration of cytarabine.

The second titration range was between 250nM and 50 mM as shown in figure 4.6C and 4.7C. The conditions without cytarabine showed exponential growth and every condition with cytarabine shows growth inhibition, generally with the higher the concentration the more inhibition. In figure 4.6D, the concentrations of 10µM and 50mM reduced the cell viability whereas all other concentrations did not, despite the reduced cell growth in figure 4.6C. Similarly, in figure 4.7D, only the 10µM and 50mM treatment reduced cell viability, as such the final titration was between 0.9µM and 8µM to bridge the gap between the first two titrations. Figure 4.6E all conditions show a similar growth curve to the control condition, with only a slight reduction in the cell density, this is reflected in the viability of the cells in figure 4.6F with only the two highest concentrations reducing the viability. As such the concentration used to treat THP-1 cell lines was 10µM. In figure 4.7E all the concentrations reduced the cell growth, however in figure 4.7F, the 0.9µM and 2µM concentrations minimally affected the cell viability compared to the 4µM, 6µM and 8 µM which reduced the cell viability, as such the most appropriate concentration to inhibit the cell cycle for the MolM-13 cell line is 4 µM. As expected, the effective concentrations are slightly different for each cell-line.

The Kasumi-1 cell line was titrated separately and based on the experimental results obtained for both MolM-13 and THP-1 the initial concentrations were between 250nM and 50µM as shown in figure 4.9. As shown in figure 4.9A all the concentrations had an effect on the cell growth rate, however in figure 4.9B the 250nM concentration showed less of an effect on the cell viability compared to the other concentrations used. The second titration was between 150nM and 450nM, shown in figure 4.9C and D, all concentrations inhibited the cell growth rate however as the 150nM concentration did not reduce the cell viability as drastically as the other concentrations used the concentration of 300nM was chosen for further experiments.

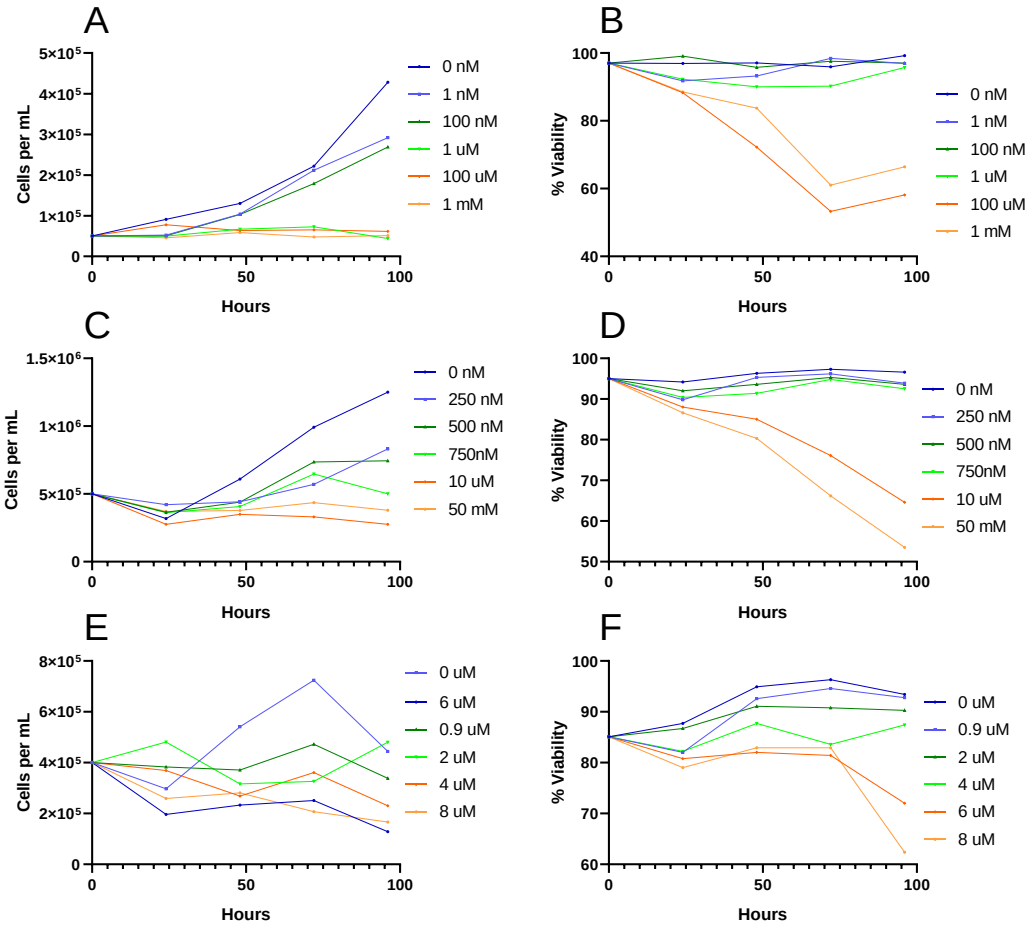

FIGURE 4.6. CYTARABINE TITRATION IN THP-1 CELL LINE. FIGURES A, C AND D SHOW THE CELL DENSITY OF THE THREE TITRATIONS, AND FIGURES B, D, F SHOW THE CELL VIABILITY OF THE CORRESPONDING TITRATION. FIGURES A AND B WERE THE INITIAL LARGE-SCALE TITRATION TO DETERMINE THE RANGE OF THE SUBSEQUENT TITRATIONS. THE CONCENTRATION CHOSEN FOR FUTURE TREATMENTS OF THE THP-1 CELL LINE WAS 10μM. N=1

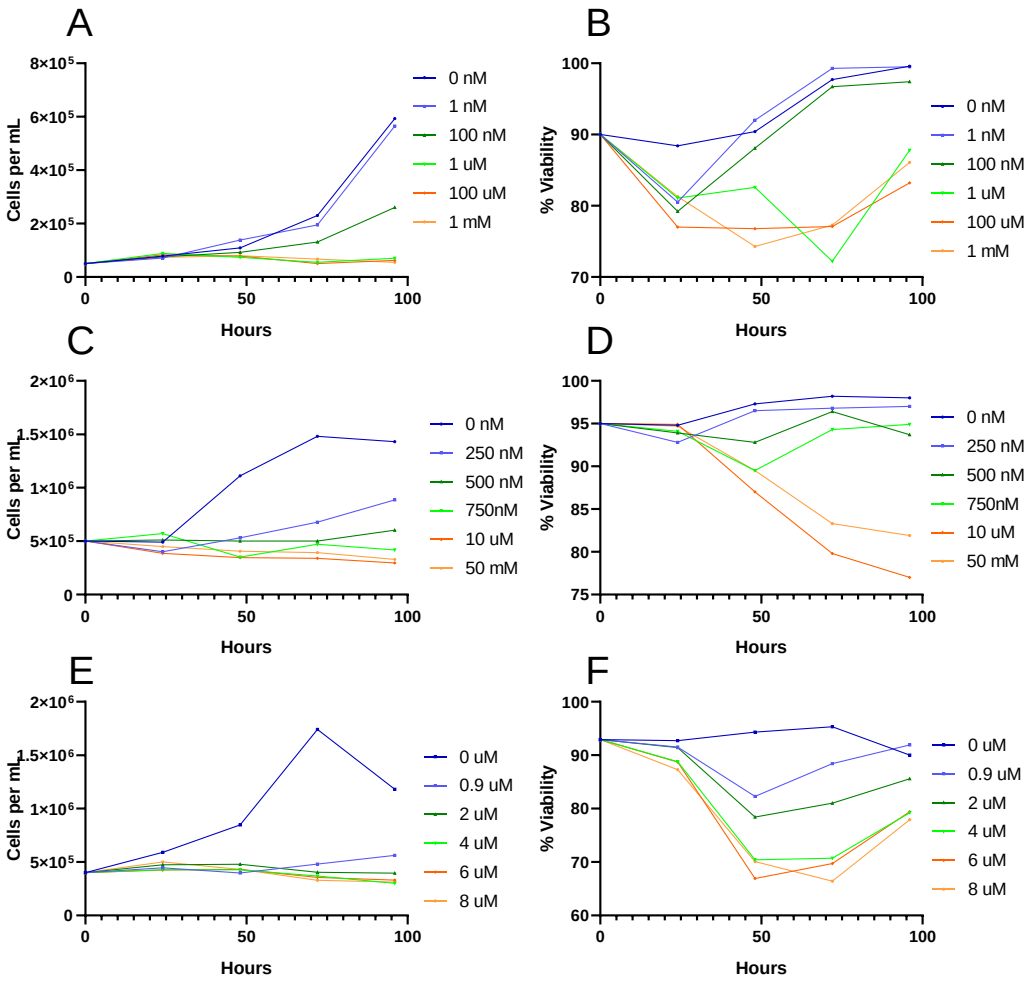

FIGURE 4.7. CYTARABINE TITRATION IN MOLM-13 CELL LINE. FIGURES A, C AND E SHOW THE CELL DENSITY OF THE TITRATIONS, AND FIGURES C, D AND F SHOW THE CORRESPONDING CELL VIABILITY. THE INITIAL TITRATION IN FIGURES A AND B WERE USED TO SELECT AN APPROPRIATE RANGE FOR SUBSEQUENT TITRATIONS. THE SELECTED CONCENTRATION FOR THE MOLM-13 CELL LINE WAS 4μM. N=1

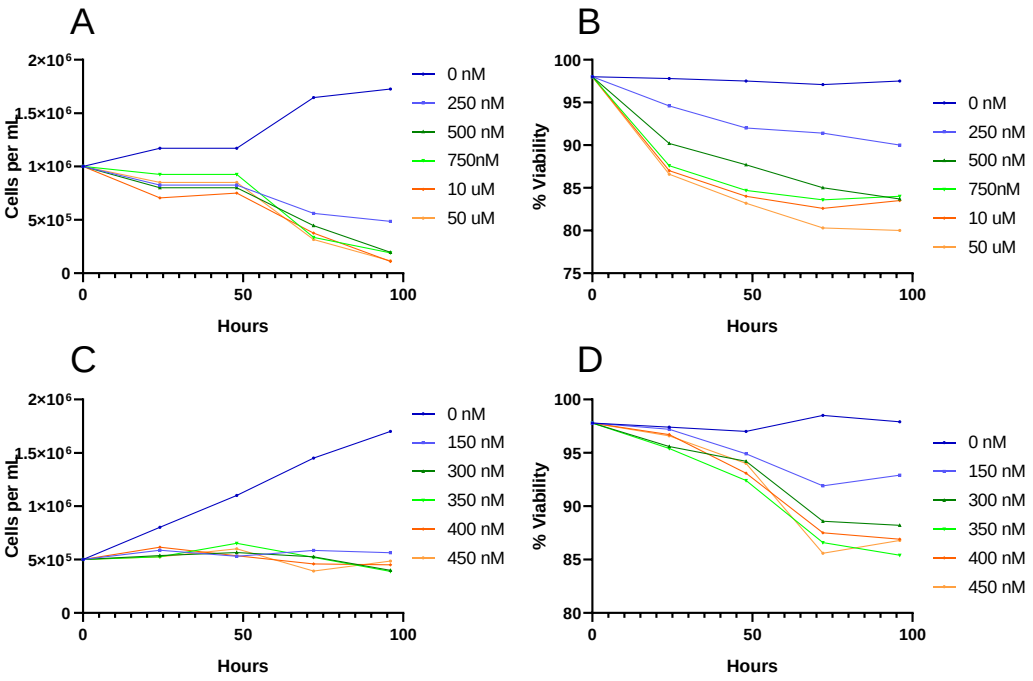

FIGURE 4.8. CYTARABINE TITRATION IN KASUMI-1 CELL LINE. FIGURES A AND C SHOW THE CELL DENSITY OF THE TITRATIONS, AND FIGURES B AND D SHOW THE CORRESPONDING CELL DENSITY. THE INITIAL TITRATION IN FIGURES A AND B WERE USED TO DISCOVER AN APPROPRIATE RANGE FOR THE FOLLOWING TITRATION IN FIGURES C AND D. THE CONCENTRATION OF CYTARABINE USED TO TREAT THIS CELL LINE WAS 300NM. N=1

4.3.1.3. Transforming growth factor β titration

The gene CD109 was found to be highly expressed in AML patients who go on to relapse. The protein CD109 promotes the degradation of the TGFBR1 protein[135,139,144,146,219] and limits its downstream signalling potential. Transforming growth factor β is a cytokine which has many different functions[225–228], including pro-apoptotic signalling pathways and prevention of cell cycle progression, to induction of differentiation and promotion of metastasis in solid cancers.

According to the literature the treatment concentration is normally around[227,228]10ng/mL, which is why the initial titrations from figure 4.9 attempt to reduce this concentration. The literature is based on solid cancers and the impact it has on epithelial to mesenchymal transition, as shown in figure 4.9 these concentrations had no effect of the AML cell lines, even with the increased concentration shown in figure 4.9E and F. Therefore higher concentrations were required in AML cell lines as discussed later in this chapter.

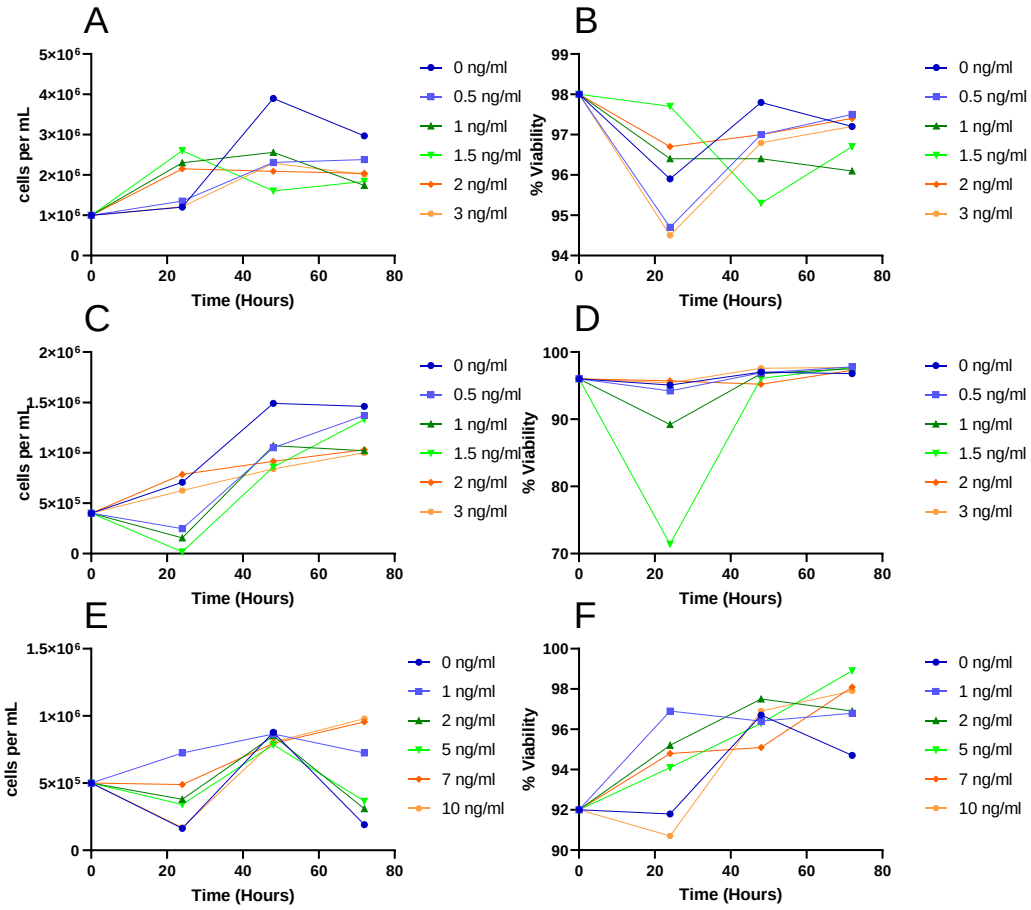

FIGURE 4.9. TGFB1 TREATMENT OF AML CELL LINES. THE AML CELL LINES WERE TREATED WITH VARIOUS CONCENTRATIONS OF TGFB1 TO INDUCE A CHANGE IN GROWTH RATE. FIGURE A, C AND E SHOW THE CELL DENSITY OF EACH OF THE TITRATIONS AND FIGURES B, D AND F SHOW THE CORRESPONDING CELL VIABILITY. FIGURES A AND B SHOW THE MOLM-13 CELL LINE, FIGURES C AND D SHOW THE THP-1 CELL LINE AND FIGURES E AND F SHOW THE KASUMI-1CELL LINE. N=1

4.3.2. PCR primer efficiency calculations

In quantitative PCR (qPCR) amplification efficiency is essential to provide consistent and reliable results. Reaction efficiency is assessed by creating standard curves; both a primer dilution series and a sample dilution series are used. The efficiency testing was performed for all primers used, including the housekeeper and genes of interest. For ease of experimentation it is preferable if both the "housekeeping" primer pair and the gene of interest primer pair achieve efficiency at the same temperature, although this is not a requirement. To assess primer efficiency in the specific PCR conditions, a few key

requirements should be met: the efficiency should be above 0.9 but below 1.1 and the melt should be a single clean peak with no tail.

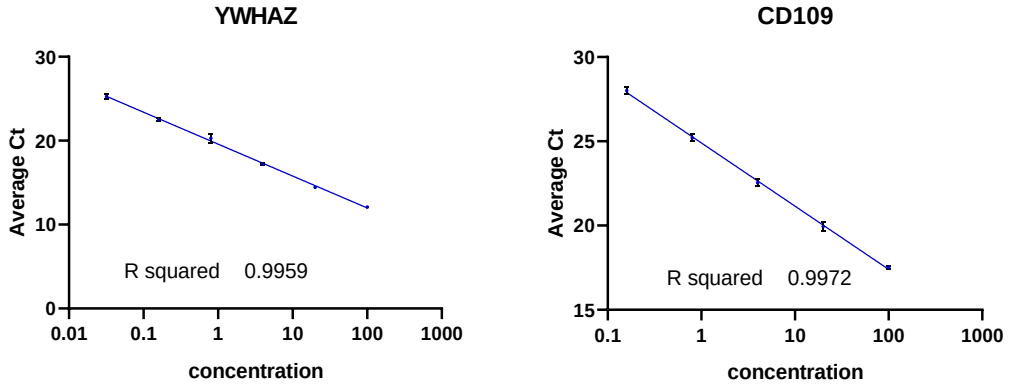

FIGURE 4.10. PRIMER EFFICIENCY TESTING. THE YWHAZ AND CD109 PRIMERS WERE TESTED AT 58°C, A STOCK OF CDNA WAS DILUTED IN A 1IN 5 DILUTION SERIES, THE SAMPLES WERE RUN IN TRIPLICATE TO DETERMINE IF THE PRIMERS WERE EFFICIENT AT THE CHOSEN TEMPERATURE. N=3

The first efficiency testing used is a sample titration; nucleic acid (cDNA) samples were diluted in a 1:2 or a 1:5 dilution series, depending on the native expression of the gene of interest. The dilution series chosen should be reflected in the cycle threshold values obtained, which are used to create the standard curve to assess efficiency in the specific PCR conditions used. Each sample in the dilution series was performed in triplicate to calculate standard deviation of the cycle threshold and exclude any results which were anomalous. Predominantly, when a primer is designed the Tm value selected is around 60°C, the first melt temperature tested was 58°C, and this will allow a stable interaction between the template and the primer for the extension of the DNA sequence prior to the Melt at 60°C. When looking at the primer efficiency the dilutions series should have equally spaced on cT values, which form a straight line giving the r squared value between 0.9 and 1.1, which is equivalent of the primer efficiency.

The first primer pair to be efficiency tested was YWHAZ, which is a stably[229–231] expressed gene in AML making it an ideal "housekeeping" gene. Using a stably expressed gene allows a benchmark to be set for comparing gene expression across various samples, where the concentration of cDNA may vary. The stable expression of the gene allows relative expression to be calculated across different samples, different treatments and different cell lines. Secondly the CD109 primer pair was tested, this is crucial to the understanding of CD109

expression in different cell lines and with different treatments. It will also enable the knockdown efficiency to be calculated. In figure 4.10 a stock of AML cDNA was diluted in a 1 in 5 series, the samples were run at 58°C in triplicate, both the YWHAZ and CD109 primers have an r squared value of 0.99, indicating they are efficient at 58°C.

4.3.3. Viral transduction

There are several methods to knockdown gene expressions in cell lines, including CRISPR-Cas9, siRNA and shRNA. CRISPR-Cas9 is expensive but creates stable cells in which the gene expression is reduced to 0% of the initial expression, as the gene is effectively deleted. When using siRNA the Knockdown is transient and therefore the expression must me checked regularly, this limits the time available to conduct experiments. Using shRNA a stably transfected cell line is established where the Plasmid DNA is inserted into the cell and expressed when the selection condition is applied, although the knockdown efficiency is variable, but there is no time limit to conduct further experiments. As an additional check, the Knockdown efficiency should be tested regularly, especially before beginning experiments.

4.3.3.1. Adherent cells

Following the method outlined above, MDA-MB268 were transduced with the CD109 shRNA knockdown plasmid and the control plasmid. The transduced and a non-transduced control were treated with Puromycin at a concentration of 0.2 µg/mL for a minimum of 72 hours or until the control cells were dead, shown in figure 4.11. The cells which contained the plasmid survived and those without underwent apoptosis, as demonstrated by the control cells in figure 4.11D.

To assess knockdown efficiency of the plasmids used, a sample was harvested for RNA extraction and cDNA synthesis. A qPCR was performed comparing the CD109 gene expression and the expression of the housekeeping gene YWHAZ. Looking at table 4.2 the estimated KD efficiency of each plasmid is shown according to the relative delta-delta-Ct value. In table 4.2, We see plasmid 52 and 50 have very different KD efficiencies, plasmid 50 gave a KD efficiency of around 73% and plasmid 52 gave a KD efficiency of 50%. Although the knockdown efficiency can vary depending on the cell line used, both plasmid 50 and 52 were used to knockdown CD109 expression in the AML cell lines.

	Average Ct YWHAZ	Average Ct CD109	ΔCt	ΔΔCt	Estimated Knockdown efficiency
Plasmid control	18.02	15.81	2.21	NA	NA
Plasmid 50	19.56	15.45	4.11	1.90	73%
Plasmid 52	18.91	15.69	3.22	1.01	50%

TABLE 4.2. KNOCKDOWN EFFICIENCY IN THE MDA-MB268 CELL LINE. THE MDA-MB268 CELL LINE WAS TRANSDUCED WITH PLASMID DNA CONFERRING THE CD109 SHRNA. THE KNOCKDOWN EFFICIENCY WAS CALCULATED BY THE CHANGE IN CT VALUE BETWEEN YWHAZ AND CD109 THEN THE CHANGE OF CT VALUES FOR DIFFERENT PLASMIDS COMPARED TO THE CONTROL PLASMID, THROUGH GENE EXPRESSION ANALYSED USING QPCR. N=3

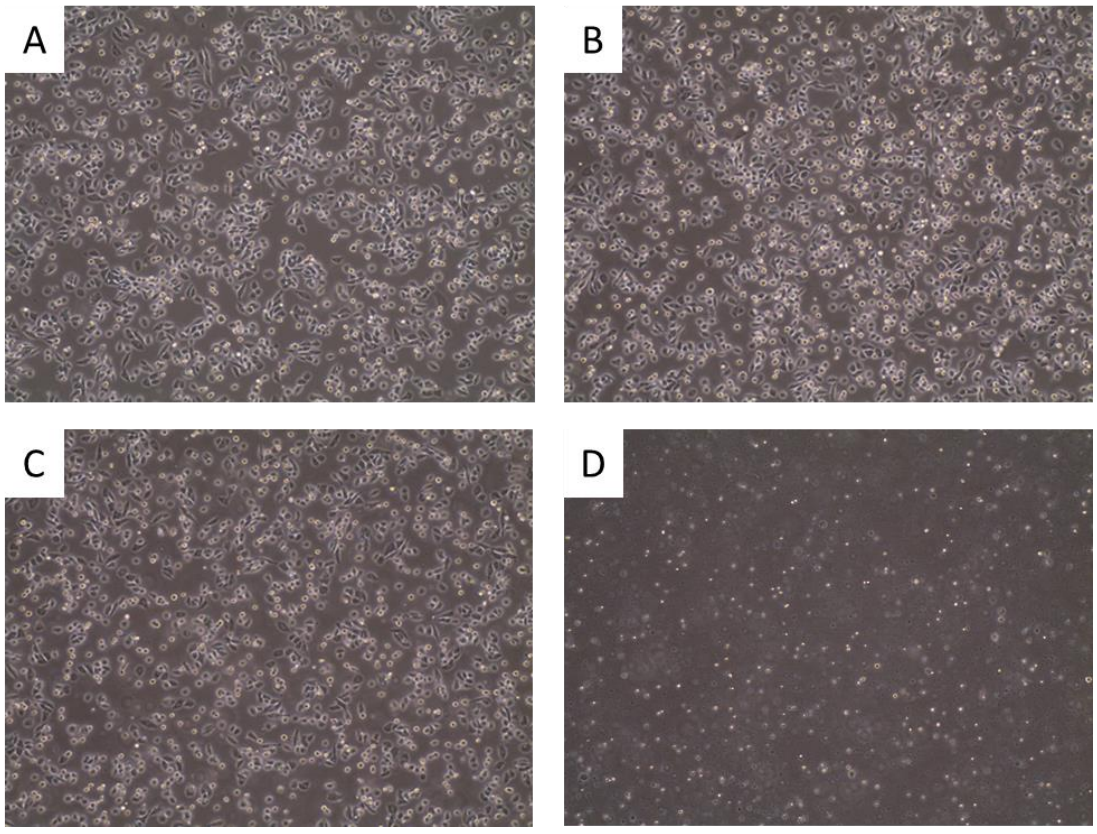

FIGURE 4.11. TRANSDUCED MDA-MB268 CELLS. PHOTOS AT X20 MAGNIFICATION 72 HOURS AFTER THE TRANSDUCED CELLS WERE TREATED WITH 0.2 µG/ML OF PUROMYCIN. FIGURE A SHOWS THE CELL LINE TRANSDUCED WITH PLASMID 50, FIGURE B SHOWS THE CELL LINE TRANSDUCED WITH PLASMID 52, FIGURE C SHOWS THE CELL LINE TRANSDUCED WITH THE CONTROL PLASMID AND FIGURE D SHOWS THE CELL LINE WITHOUT PLASMID TRANSDUCTION.

4.3.3.2.Suspension cells

The method used for adherent cells involved adding the virus to the monolayer and incubating, when this method was applied to suspension cells it was only affective in the THP-1 cell line. In all AML cell lines, except THP-1, the virus failed to enter the cells, did not transcribe the plasmid encoding Puromycin resistance or were destroyed in the process through toxicity. To transduce the remaining AML cell lines a "spin-fection" protocol was applied, which involved a small volume of high-density cell suspension centrifuged at 800xg for 30-45 minutes with the viral particles. The initial attempt with a centrifugation of 30 minutes was successful in the MolM-13 cell line, the second attempt with a centrifugation of 45 minutes was successful in Kasumi-1 cell line. After the cell lines had become established the knockdown efficiency was calculated for these cell lines shown in table 4.3. The Knockdown was not as strong as expected, with only a 20% knockdown in the MolM-13 cell line for both plasmids, and 31% and 45% knockdown in Kasumi-1. The knockdown was strongest in the THP-1 cell line with a 76% and 40% knockdown for plasmid 50 and 52 respectively. From this it is evident that different cell lines respond differently to the plasmid DNA and each plasmid has a different potency of knockdown, showing the importance of using more than one cell line and plasmid for the transfection.

Cell line	plasmid	Average Ct YWHAZ	Average Ct CD109	ΔCt	ΔΔCt	Estimated Knockdown efficiency
THP	control	14.91	18.08	3.17	NA	NA
	50	14.14	19.58	5.44	2.27	76%
	52	14.43	18.34	3.92	0.74	40%
Kasumi-1	Control	15.15	18.92	3.77	NA	NA
	50	14.89	19.50	4.61	0.84	45%
	52	14.78	19.07	4.29	0.52	31%
MolM-13	Control	13.54	18.64	5.10	NA	NA
	50	13.79	19.25	5.45	0.35	20%
	52	13.81	19.28	5.47	0.37	20%

TABLE 4.3. KNOCKDOWN EFFICIENCY OF THE AML CELL LINES. THE AML CELL LINES WERE TRANSDUCED WITH PLASMID DNA CONFERRING THE CD109 SHRNA KNOCKDOWN. THE RELATIVE EXPRESSION OF CD109 COMPARED TO YWHAZ WAS CALCULATED (ΔCT) AND THE CHANGE IN EXPRESSION (ΔΔCT) CALCULATED FROM THERE. THE ESTIMATED RELATIVE KNOCKDOWN OF THE CD109 GENE IS SHOWN.

When cell lines are transduced with plasmid DNA, the cellular mechanisms transcribe the DNA which provides a survival advantage for those cells, in this case the resistance to Puromycin which was in the growth medium, this is a selective pressure to encourage the continued transcription of the plasmid DNA. To increase KD efficiency, the cell lines were grown in a higher concentration of Puromycin to force the transcription of the plasmid containing the shRNA and the Puromycin resistance.

4.3.4. Cell proliferation and apoptosis

Cellular metabolic activity reduces Tetrazolium salts to formazan[232] which is the basis of the XTT assay. The coloured formazan product in the XTT assay is soluble, so exits the cell without the need for lysis, unlike the product of the MTT assay which is crystallised and requires a lysis step to release. The colorimetric change that occurs can be measured using a wavelength of 470 nm with a subtraction wavelength over 650nm.

There are certain limitations of using tetrazolium salt assays, one such limitation is the dynamic range; for the MTT assay there is a relatively small dynamic range, in which samples may appear similar to one another, especially as the colour becomes more saturated. The XTT has a larger dynamic range in comparison the MTT assay, meaning it is easier to discern the different metabolic activity of samples across the range.

Although the colorimetric changes are correlated to cell number, it may be unreliable to directly compare between different cell-lines, within the same cell line should be consistent[232]. Despite the fact direct quantitative comparisons cannot be drawn between cell lines, this is a crucial tool for comparing the impact of treatments within a cell line. Additionally, additives in the media can cause the reduction of tetrazolium salt independently of cellular metabolic activity[233], therefore regulation of the media and additives in crucial to getting a reliable result.

4.3.4.1. Impact of reduced CD109 expression on growth rate

High expression of CD109 has been linked to negative outcomes in various cancers, it is crucial to determine if CD109 is associated with increased proliferation of if it has another roll in cell survival. An XTT assay was implemented in each of the CD109 knockdown cell lines to assess the growth rate when CD109 is reduced compared to the normal high expression. The cell lines with the shRNA plasmid were seeded in at a density of $4x_{10}^{5}$ in 96-well plates, the XTT reagents were added to a separate plate daily and after a 4-hour incubation the absorbance was measured.

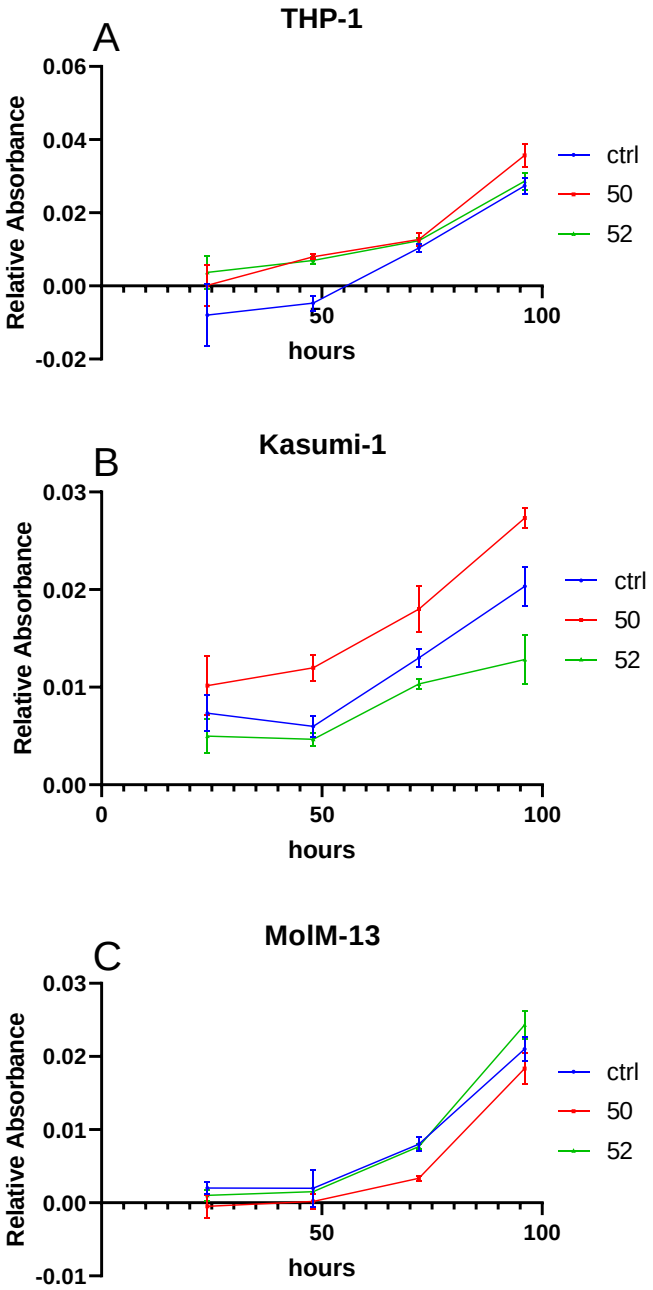

FIGURE 4.12. GROWTH RATE OF CD109 KNOCKDOWN CELL LINES. THE CELL LINES TRANSDUCED WITH PLASMID DNA CONFERRING CD109 SHRNA WERE SEEDED AND THE GROWTH RATE WAS MEASURED BY THE XTT ASSAY EVERY 24HOURS. EACH FIGURE REPRESENTS A DIFFERENT CELL LINE WITH THE SAME PLASMID TRANSFECTION. N=3

The XTT assay uses a reference wavelength of 650nm, the reading for which was subtracted from the measurement wavelength of 470nm to give the relative absorbance. In addition, a medium blank was measured in the same way to allow the reading to be background corrected as displayed in figure 4.12. It is clear from figure 4.12 the reduced expression of CD109 does not affect growth rate, all cell lines show the same trend in cell growth despite the introduction of the plasmid DNA. As CD109 does not appear to be critical to the cell growth it must have another function within the cell.

4.3.4.2. The effect of TGFβ1 on cell proliferation

The initial titrations for TGFβ1 treatment in AML cell lines indicated lower concentrations of TGFβ1 do not elicit a response. To gage the impact of TGFβ1 in the transfected AML cell lines a 25ng/mL concentration was used, as shown in figure 4.13. Although this concentration is higher than expected based on literature, all previous cell lines treated with TGFβ1 were adherent cell lines which respond to TGFβ1 differently. An XTT assay was performed in the CD109 knockdown AML cell lines shown in figure 4.13, it is clear TGFβ1 induces cell cycle arrest in the cell lines at the 25ng/mL concentration. Figure 4.13C shows the THP-1 cell line, although the cell cycle arrest was not as dramatic as the MolM-13 (figure 4.13A) or Kasumi-1 cell line (figure 4.13B), there is a reduced cell proliferation. It is important to note that despite the effects of TGFβ1 on the cell lines, the CD109 knockdown conditions did not exhibit any further differences in the cell proliferation rate compared to the control conditions.

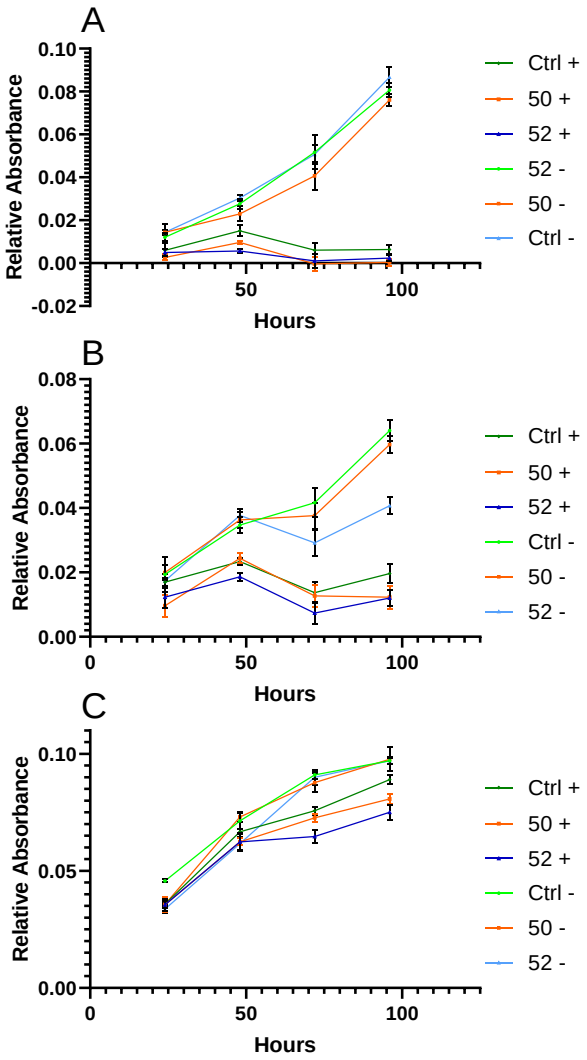

FIGURE 4.13. ASSAY OF CD109 KNOCKDOWN CELL LINES TREATED WITH TGFB1. THE CD109 KNOCKDOWN CELL LINES WERE TREATED WITH TGFB1 AND THE ABSORBANCE WAS MEASURED EVERY 24 HOURS. FIGURE A SHOWS THE MOLM-13 CELL LINE, FIGURE B SHOWS THE KASUMI-1 CELL LINE AND FIGURE C SHOWS THE THP-1 CELL LINE. N=3

4.3.4.3. Cytarabine and TGFβ1 treatment in AML cell lines

It was found that a TGFβ1 neutralising antibody had the ability to enhance cytarabine induced apoptosis in AML[220], from this it would be expected that in the CD109 knockdown cell lines the response to cytarabine may be reduced as the TGFβ1 signalling is uninterrupted. Despite the different expression of CD109 in the AML cell lines and the knockdown cell lines, it has been established that TGFβ1 causes cell cycle inhibition regardless. Looking at the combination of cytarabine and TGFβ1 may indicate how CD109 affects treatment response and how high CD109 promotes cancer progression and relapse. To do this the AML cell lines transduced with different plasmids were seeded and treated with TGFβ1 and cytarabine at the concentration corresponding to their calibration titration. An XTT assay was use and the cell line were measured every 24 hours to determine the difference in cell respiration.

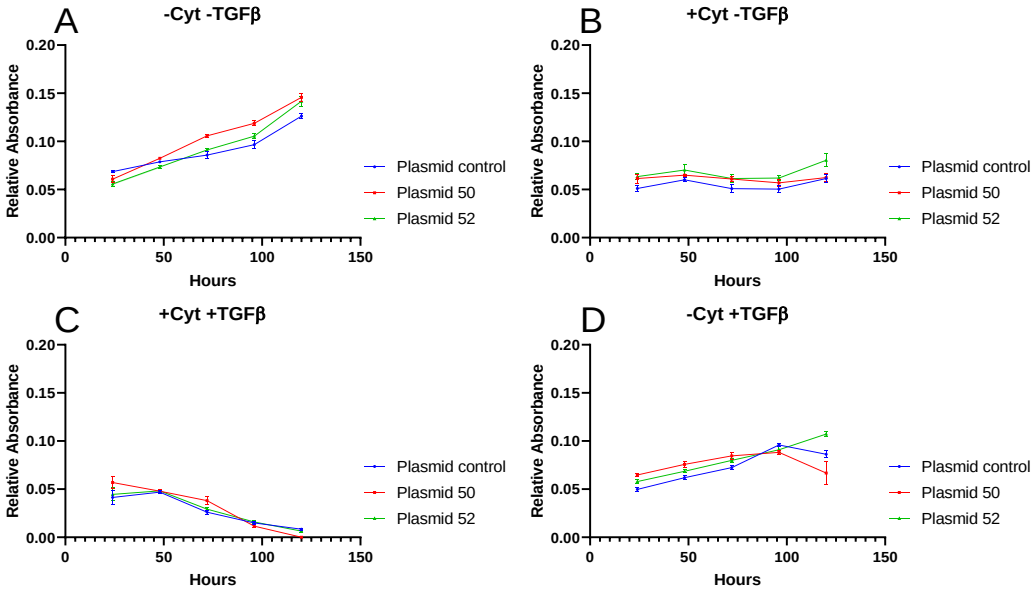

FIGURE 4.14. XTT ASSAY OF THE THP-1 CELL LINE TREATED WITH TGFB1 AND CYTARABINE. THE THP-1 CELL LINE WAS TRANSDUCED WITH DIFFERENT PLASMIDS, TWO CONFERRING CD109 SHRNA TO REDUCE THE EXPRESSION OF CD109. THESE CELL LINES WERE TREATED WITH TGFB1 AND CYTARABINE TO DETERMINE THE EFFECT OF CELL PROLIFERATION AND APOPTOSIS, AS MEASURED BY THE XTT ASSAY. N=3

Figure 4.14 displays the THP-1 cell line and the response to TGFβ1 and cytarabine, overall, the different expression of CD109 does not appear to affect the response to cytarabine or TGFβ1 with all plasmids following the same trend each time. Without any treatment the different plasmid conditions grow continually over the course of the experiment (figure

4.14A), however, with the introduction of TGFβ1 (figure 4.14D) the growth rate was reduced. When the THP-1 cells were treated with cytarabine there was a total cell cycle arrest with no increase in the cellular population, as shown in figure 4.14B. Interestingly when the cell line was treated with both TGFβ1 and cytarabine the cell population decreased over the course of the experiment (figure 4.14C), likely due to the increase apoptosis. In indication that TGFβ1 increases cellular apoptosis in response to cytarabine is in direct contrast so the research[220] by Tabe. Y et.al (2013) which indicated reducing the TGFβ1 signalling increased cytarabine related apoptosis.

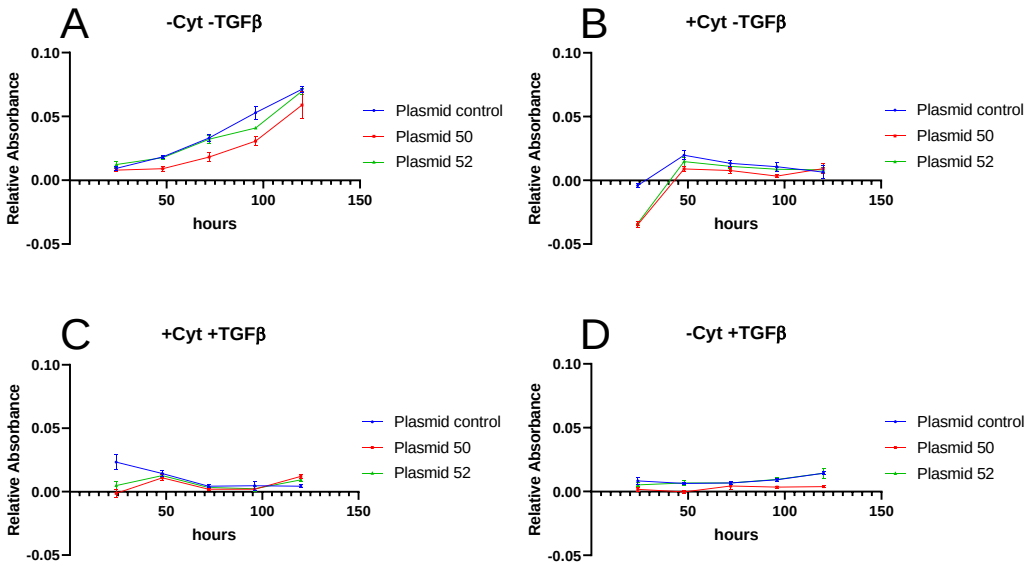

FIGURE 4.15. XTT ASSAY OF THE MOLM-13 CELL LINE TREATED WITH TGFB1 AND CYTARABINE. THE MOLM-13 CELL LINE WAS TRANSDUCED WITH DIFFERENT PLASMIDS, TWO CONFERRING CD109 SHRNA TO REDUCE THE EXPRESSION OF CD109. THESE CELL LINES WERE TREATED WITH TGFB1 AND CYTARABINE TO DETERMINE THE EFFECT OF CELL PROLIFERATION AND APOPTOSIS, AS MEASURED BY THE XTT ASSAY. N=3

Similarly, the MolM-13 cell line was treated with a combination of TGFβ1 and cytarabine in figure 4.15. MolM-13 cells without treatment showed continued cell growth with little or no difference between the plasmid conditions (figure 4.15A). When the cells were treated with cytarabine (figure 4.15B) the cell cycle was inhibited and showed some signs of apoptosis by the slight downward trajectory of the relative absorbance. In figure 4.15D, TGFβ1 induces cell cycle arrest as displayed by the completely flat relative absorbance in each of the plasmid conditions, although plasmid 50 showed a lower relative absorbance compared to the other conditions the difference is minimal. When the cell line was treated with both TGFβ1 and cytarabine (figure 4.15C) there is a downward trajectory of the relative absorbance indicating

the combined condition induces cell cycle arrest and apoptosis. Overall, in figure 4.15, there was no difference between the plasmid conditions, indicating CD109 expression does not impact treatment response.

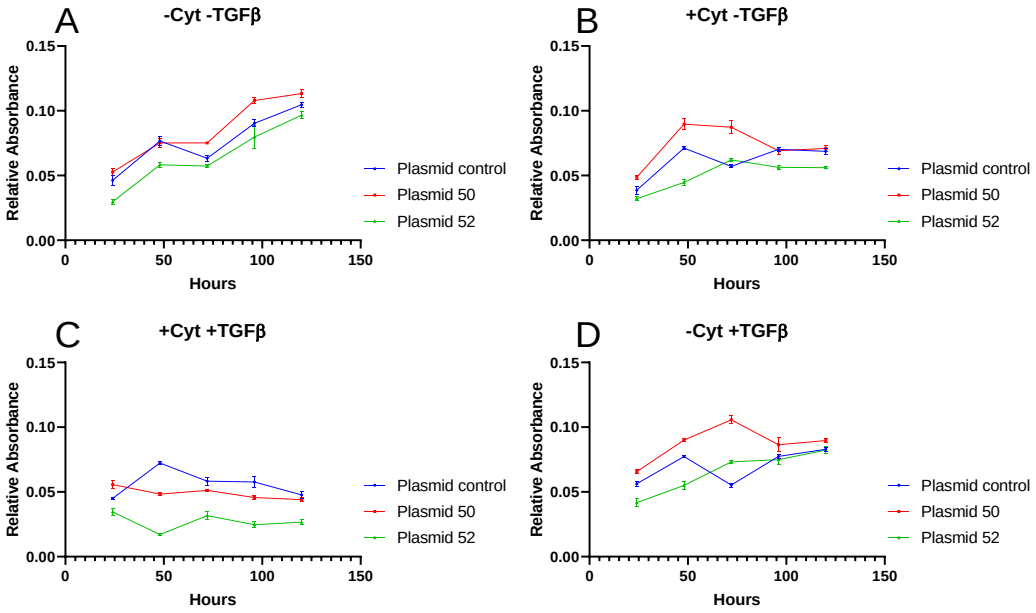

FIGURE 4.16. XTT ASSAY OF THE KASUMI-1 CELL LINE TREATED WITH TGFB1 AND CYTARABINE. THE KASUMI-1 CELL LINE WAS TRANSDUCED WITH DIFFERENT PLASMIDS, TWO CONFERRING CD109 SHRNA TO REDUCE THE EXPRESSION OF CD109. THESE CELL LINES WERE TREATED WITH TGFB1 AND CYTARABINE TO DETERMINE THE EFFECT OF CELL PROLIFERATION AND APOPTOSIS, AS MEASURED BY THE XTT ASSAY. N=3

In figure 4.16 the cell line Kasumi-1 was treated with combinations of cytarabine and TGFβ1. In figure 4.16A the cells were not treated with TGFβ1 or cytarabine as a baseline measurement, it is clear there is no growth inhibition and the cells continue to proliferate over the course of the experiment. In figure 4.16B the cells were treated with cytarabine which inhibited the cell proliferation to a similar degree in each of the plasmid conditions as shown by the flatter relative absorbance rate compared to the untreated condition. When the cells were treated with TGFβ1 (figure 4.16D) there is inhibition of the cell cycle shown by the flatter rate of the relative absorbance compared to untreated cells. In figure 4.16C the Kasumi-1 cell line was treated with both TGFβ1 and cytarabine, the cell proliferation was inhibited and there is an indication that apoptosis was induced by the downward trajectory of the relative absorbance over the course of the experiment.

Overall, treating AML cell lines with either TGFβ1 or Cytarabine induces cell cycle arrest, when both Cytarabine and TGFβ1 are used in conjunction the cell cycle is inhibited and apoptosis was induced. Although in this instance the different plasmid conditions did not behave differently to the treatments used, suggesting that CD109 expression did not affect the treatment response.

4.3.5. NanoString in CD109 knockdown cell lines

The cell line THP-1 was transduced with three plasmids, one conferring Puromycin resistance, and two conferring the CD109 knockdown shRNA. As the CD109 protein binds to and promotes the degradation of TGFβ1 receptor protein[144–146], the cells with reduced expression of CD109 should have increase signalling through TGFBR1. The transduced THP-1 cell lines were treated with cytarabine and TGFβ1 to determine the effects on the TGFβ1 pathways, and treatment response, using the NanoString platform. The human cancer pathology probe set was selected as this contains the TGFBR1 gene and the downstream signalling molecules associated with the TGFβ1 pathway. In table 4.4 there is an outline of the different treatments used and the combination of TGFβ1 and Cytarabine for each plasmid condition. The cells were seeded at a density of $1x_{10}^{6}$ cells/mL, TGFβ1 was added to the appropriate conditions at a concentration of 25ng/mL and incubated for 24 hours before cytarabine was added to the corresponding conditions at a concentration of 10μM. After a further incubation of 48 hours the cells were harvested, and the RNA was extracted.

	THP plasmid control				THP plasmid 50				THP plasmid 52			
Cytarabine	+	+	-	-	+	+	-	-	+	+	-	-
TGFβ1	+	-	+	-	+	-	+	-	+	-	+	-

TABLE 4.4. EXPERIMENTAL DESIGN FOR THE TREATMENT OF CD109 KNOCKDOWN CELLS. THE THP-1 CELL LINE WAS TRANSDUCED WITH SHRNA CONFERRING A CD109 KNOCKDOWN, THE DIFFERENT PLASMID CONDITIONS WERE TREATED WITH A COMBINATION OF TGFB1 AND CYTARABINE.

After treatment the cells were harvested, RNA was extracted, and cDNA was synthesised. A qPCR was performed for each sample looking at the Genes YWHAZ and a housekeeper and CD109 as the gene of interest. From the results of the qPCR the knockdown efficiency was calculated for each sample as displayed in table 4.5, comparing the knockdown plasmid condition to the corresponding condition for the control plasmid[165]. Although all cells were seeded for the same flask of transduced cells each condition shows a different knockdown

efficiency. Interestingly in THP cells transduced with plasmid 50 the knockdown without treatment was 76% however with the introduction of cytarabine the knockdown was completely null, in comparison to TGFβ1 treatment which reduced the knockdown to 45% and with TGFβ1 and cytarabine the reduction was to 25% knockdown. In THP cells transduced with plasmid 52 the knockdown without treatment was 40%, the introduction of cytarabine had no effect but treatment with TGFβ1 reduced the knockdown to 25%, interestingly in the condition with both cytarabine and TGFβ1 the knockdown was null. The differences seen between the plasmids is due to the shRNA sequences which target different sections of the CD109 RNA, altering the efficiencies. The changes in knockdown efficiency between the different treatments is interesting and as yet unexplained: it may be due to treatments inducing apoptosis and inhibiting the transcription mechanisms, or CD109 may be a TGFβ1 inducible gene which contributes to a negative feedback loop, further investigation is needed to fully understand the reason for altered expression between conditions.

Once the CD109 expression had been determined for each of the conditions, the NanoString cancer pathology probe set was used to further understand the pathways that are active in the cells. Figure 4.17 shows the correlation between the log2 counts of the untreated cells compared to the treated cells from the same plasmid group. Figure 4.17A shows the correlation in THP-1 cells transfected with the control plasmid, comparing the untreated cells to the treated cells. Overall, the counts for the genes are correlated, as displayed in the R^2 value, although there is some variation between the treatment conditions. Treating the cell with both cytarabine and TGFβ1 shows the largest change in correlation out of all the treatment conditions. In figure 4.17B the correlation between treated and untreated THP-1 cells transfected with plasmid 50. All conditions are highly correlated with the untreated cells, however the condition treated with both cytarabine and TGFβ1 shows the largest variation. The same trend is shown in figure 4.17C where the least correlated treatment is the condition with both cytarabine and TGFβ1. Although all treatments are highly correlated there is variance from the untreated condition in every plasmid group.

	Cyt	TGFβ1	Average Ct YWHAZ	Average Ct CD109	ΔcT	ΔΔcT	Estimated Knockdown efficiency
TC	+	+	15.11	18.36	3.26	N/A	N/A
	+	-	15.05	18.35	3.30	N/A	N/A
	-	+	14.16	16.83	2.67	N/A	N/A
	-	-	14.91	18.08	3.17	N/A	N/A
T50	+	+	13.85	17.43	3.58	0.32	25%
	+	-	14.28	17.41	3.13	-0.18	0%
	-	+	15.16	19.25	4.08	1.41	45%
	-	-	14.14	19.58	5.44	2.27	76%
T52	+	+	15.11	17.97	2.86	-0.40	0%
	+	-	14.64	18.57	3.93	0.63	40%
	-	+	15.01	18.64	3.63	0.96	25%
	-	-	14.43	18.34	3.92	0.74	40%

TABLE 4.5. THE CALCULATED KNOCKDOWN EFFICIENCY POST TREATMENT FOR EACH CONDITION. THE THP-1 CELL LINE WAS TRANSDUCED WITH THREE PLASMIDS, ONE CONTROL PLASMID AND TWO SHRNA PLASMIDS FOR CD109. THE CYCLE THRESHOLD WAS DETERMINED BY A QPCR OF THE SAMPLES IN TRIPLICATE, FOR BOTH CD109 AND THE HOUSEKEEPING GENE YWHAZ. THE ΔCT WAS CALCULATED AS THE CHANGE IN THE CYCLE THRESHOLD BETWEEN THE TWO GENES, THEN THE ΔΔCT WAS CALCULATED AS CHANGE IN THE CYCLE THRESHOLD BETWEEN THE CONTROL PLASMID AND THE CORRESPONDING CONDITION OF THE KNOCKDOWN PLASMID.

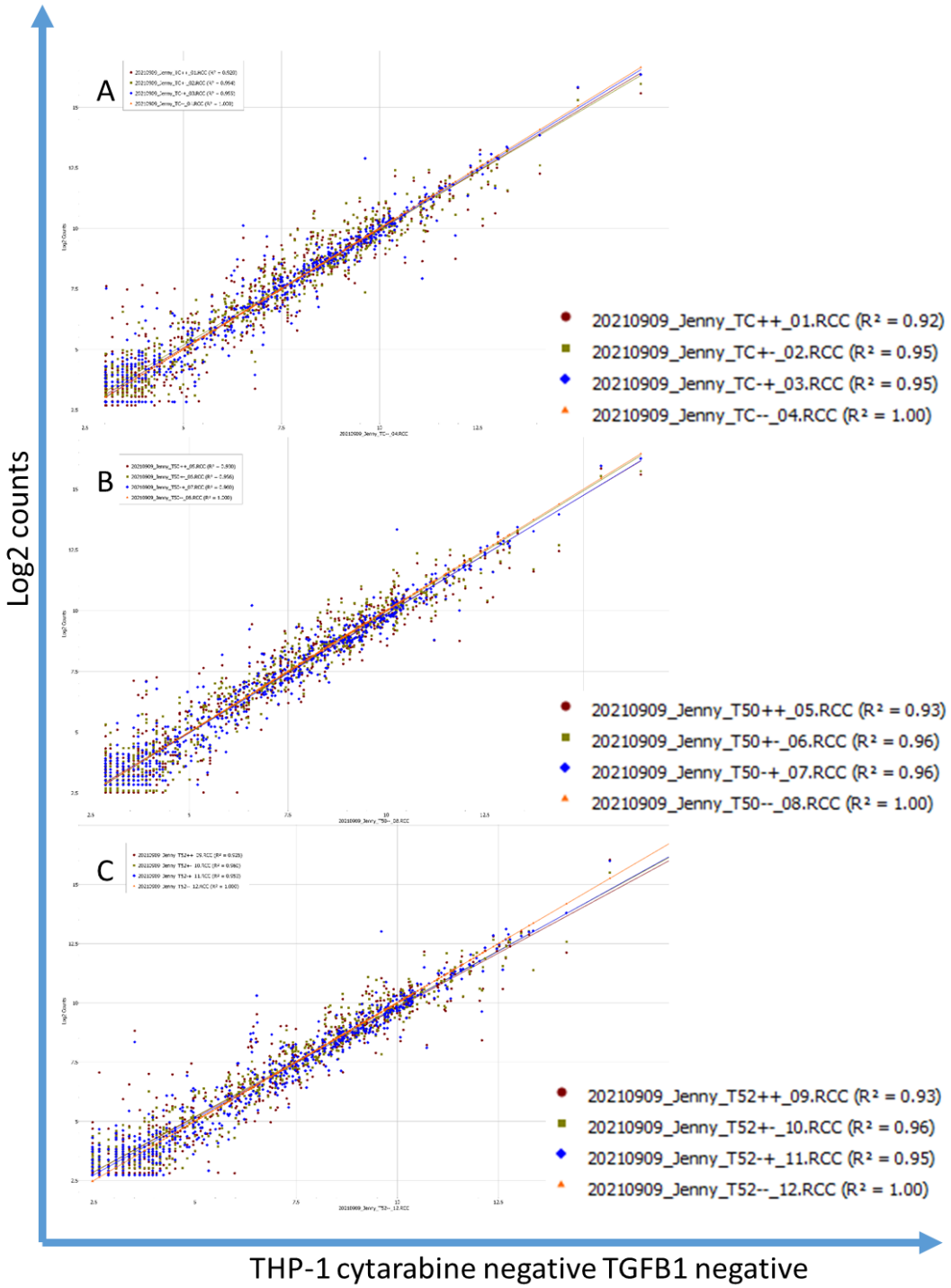

FIGURE 4.17. GENE EXPRESSION OF THP-1 CELLS TREATED WITH TGFB1 AND CYTARABINE. CORRELATION BETWEEN THP-1 CELLS TREATED WITH TGFB1 AND CYTARABINE COMPARED TO THP-1 CELLS WITHOUT CYTARABINE AND TGFB1, FROM THE SAME PLASMID CONDITION.

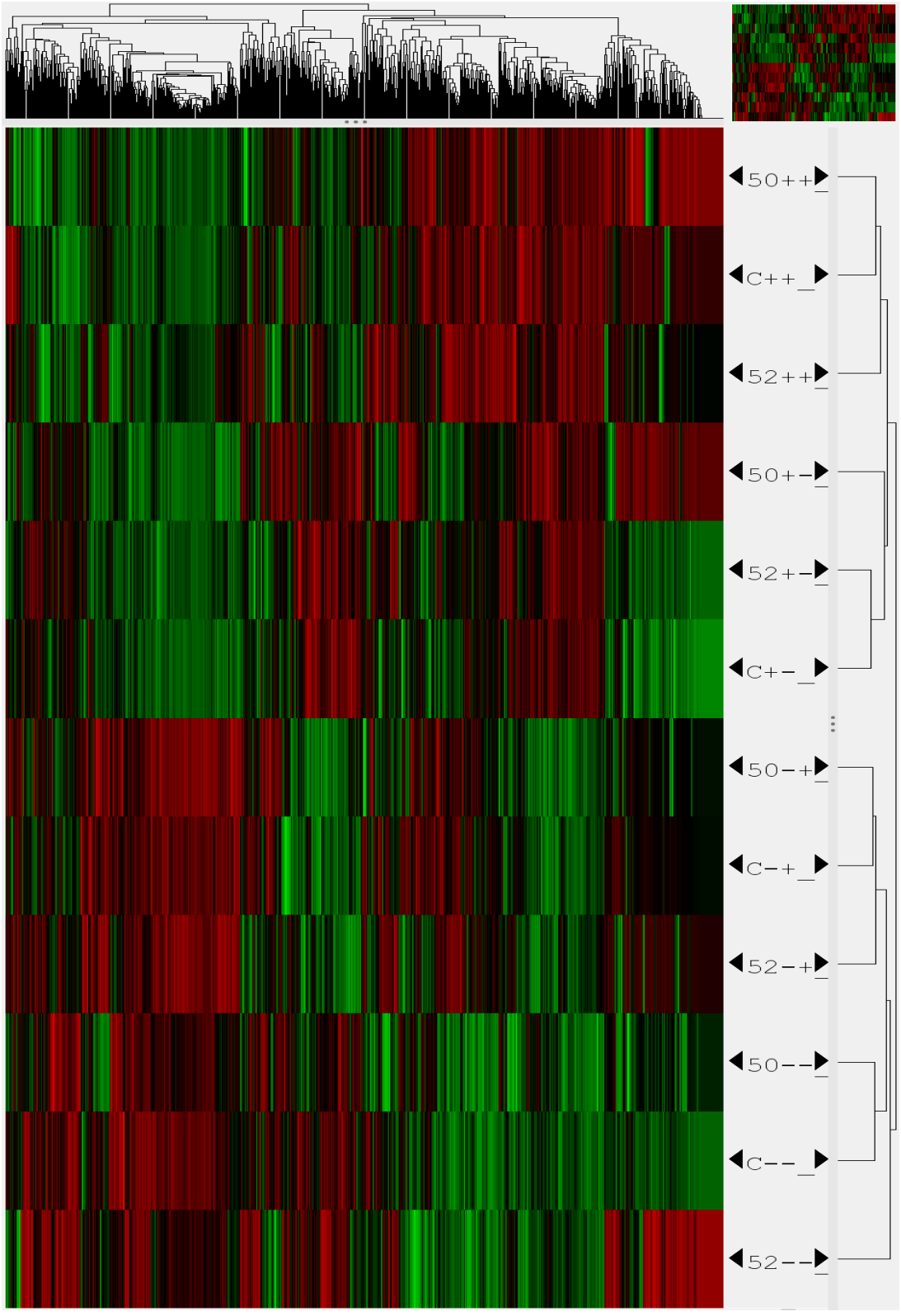

FIGURE 4.18. HEAT MAP WITH HIERARCHAL CLUSTERING OF THE THP-1 CELL LINE. THE NOMENCLATURE ON THE RIGHT INDICATES THE PLASMID USED TO TRANSDUCE THE SAMPLE FOLLOWED BY THE CYTARABINE STATUS THEN THE TGFB1 STATUS. HIERARCHAL CLUSTERING SHOWS THE SAMPLES MOST SIMILAR TO ONE ANOTHER.

To compare each of the different conditions the heat map in figure 4.18 was used. The hierarchal crusting across the top shows the relationship of each of the genes in the panel. Along the side the hierarchal clustering shows which of the conditions are most similar in their gene expression. The clustering shows overwhelmingly cytarabine treatment was the definitive factor for the gene expression profiles, as indicated by the first "+" or "-" in the sample name, then followed by the TGFβ treatment as indicated by the second "+" or "-" in the sample name. There are two main branches of the hierarchal clustering which are distinctly separated by the cytarabine treatment conditions, where all cytarabine positive samples are together and all cytarabine negative samples are together. Interestingly the plasmid 52 cytarabine negative and TGFβ1 negative sample clustered independently of all other samples. Within each subsection of clustering, there is no relationship between which sample had the highest CD109 knockdown (table 4.5) and which sample is more closely related to the control plasmid.

In figure 4.19 the fold change was calculated by comparing the CD109 knockdown condition to the corresponding condition in the control plasmid, from there the most down regulated genes for each of the treatment options for plasmid 50 was compared to the other conditions. In figure 4.19A, the top 10 downregulated genes associated with plasmid 50 treated with both cytarabine and TGFβ1 are displayed with a comparison to the other treatment conditions. Interestingly, looking at the gene RASGRP1, in the treatments with cytarabine there is a decrease in this gene, however in conditions without there is an upregulation of this gene, although plasmid 52 treated with cytarabine only and without treatment do not follow this pattern. The genes DKK1, THBS4, FGF22 and MAP2K6 are highly downregulated in plasmid 50 treated with both cytarabine and TGFβ1 but overall were upregulated in the other treatment conditions.

Looking at figure 4.19B, plasmid 50 with cytarabine showed a similar fold change in the genes compared to the other conditions. The gene UTY showed the least similar fold change values to the other conditions, where this gene was downregulated in the plasmid 50 and 52 cytarabine treated cells only, where the other conditions showed a slight upregulation or no major difference in fold change. In figure 4.19C all the top 10 downregulated genes show a similar pattern in the different treatment conditions, with the majority of them showing a downregulation and to a similar magnitude.

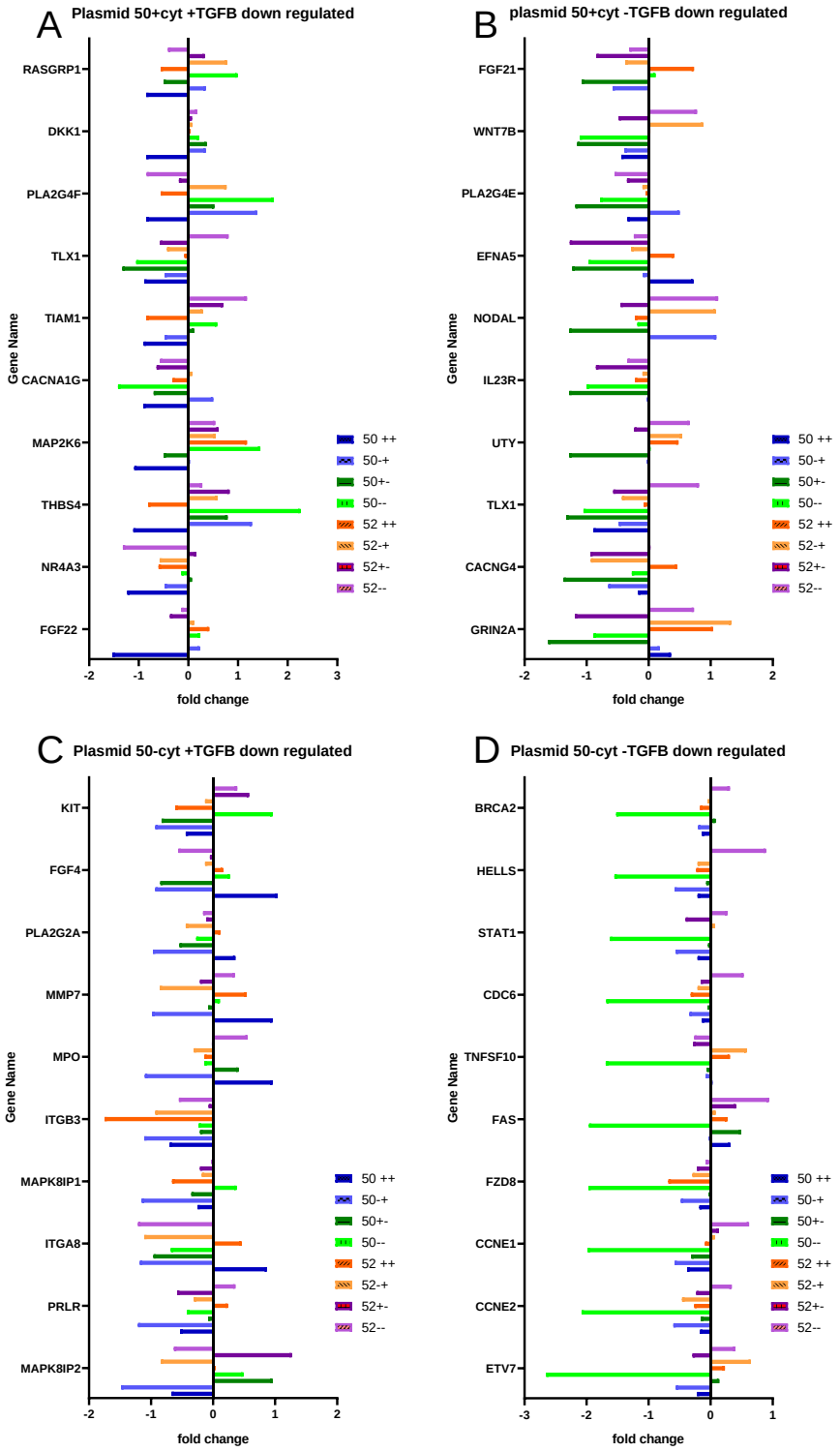

FIGURE 4.19. TOP 10 DOWNREGULATED GENES FOR PLASMID 50. THE THP-1 CELL LINE WAS TRANSDUCED WITH PLASMID DNA CONTAINING SHRNA TO REDUCE THE EXPRESSION OF CD109, THESE CELLS WERE TREATED WITH A COMBINATION OF CYTARABINE AND TGFB1. DIFFERENCE IN GENE EXPRESSION COMPARED TO THP-1 CELLS WITHOUT CD109 SHRNA WAS CALCULATED AS A FOLD-CHANGE, WHICH WAS COMPARED TO THE DIFFERENT TREATMENT CONDITIONS.

Figure 4.19D displayed the top 10 most downregulated genes of plasmid 50 without treatment, which has the most highly downregulated gene expression of all the conditions in figure 4.19. Although the genes displayed here are highly downregulated with reduced CD109 expression, the same genes are not downregulated to the same magnitude in plasmid 52. The genes displayed show a large fold change, compared to the other conditions where there is minimal difference from the control plasmid.

Figure 4.20 shows the top 10 upregulated genes for plasmid 50 in each of the treatment conditions compared to all other treatment conditions. In figure 4.20A there are four genes which are upregulated in plasmid 50 treated with both cytarabine and TGFβ1 that show a different pattern compared to other conditions. These genes are CDKN2A, IL12B, FGF4 and INHBA, which are mostly down regulated in the other conditions or have a very slight upregulation. Interestingly the gene IL12B is highly upregulated in plasmid 50 and 52 when treated with both TGFβ1 and cytarabine.

Figure 4.20B shows plasmid 50 treated with cytarabine, in general the top ten most upregulated genes in this condition behave similar to all other conditions. However, there are two genes that are exclusive to the plasmid 50 conditions, ITGB7 and IL1B, which in the plasmid 52 conditions are downregulated.

Figure 4.20C shows plasmid 50 treated with TGFβ1, overall, the upregulated genes in THP-1 cells transduced with plasmid 50 treated with just TGFβ1 showed a similar pattern to all other conditions. Many of the genes showed plasmid 50 and 52 both treated with TGFβ1 showed the same downregulation of the genes and often to the same magnitude. The genes RNF43 and SFN were highly upregulated compared to all other conditions, showing these are crucial to the response in this cell line.

The most upregulate genes of the THP-1 cell line transduced with plasmid 50 without treatment are shown in figure 4.20D. Similar to figure 4.19D the magnitude of the upregulated genes are very high, meaning they are very different to the THP-1 cell line transduced with plasmid control. The genes SOCS3 and FOS were both upregulated in the plasmid 50 conditions but downregulated in the plasmid 52 conditions. The genes HIST1H3H and HIST1H3B were both highly upregulated in this condition compared to all other conditions which showed minimal difference compared to the control plasmid conditions. Overall, there were only 5 genes that were commonly differentially expressed in figures 4.20 and 19. These were CRLF2, TLX1, NODAL, THBS4 and CACNB4, although there were genes from the same family that were common to different conditions.

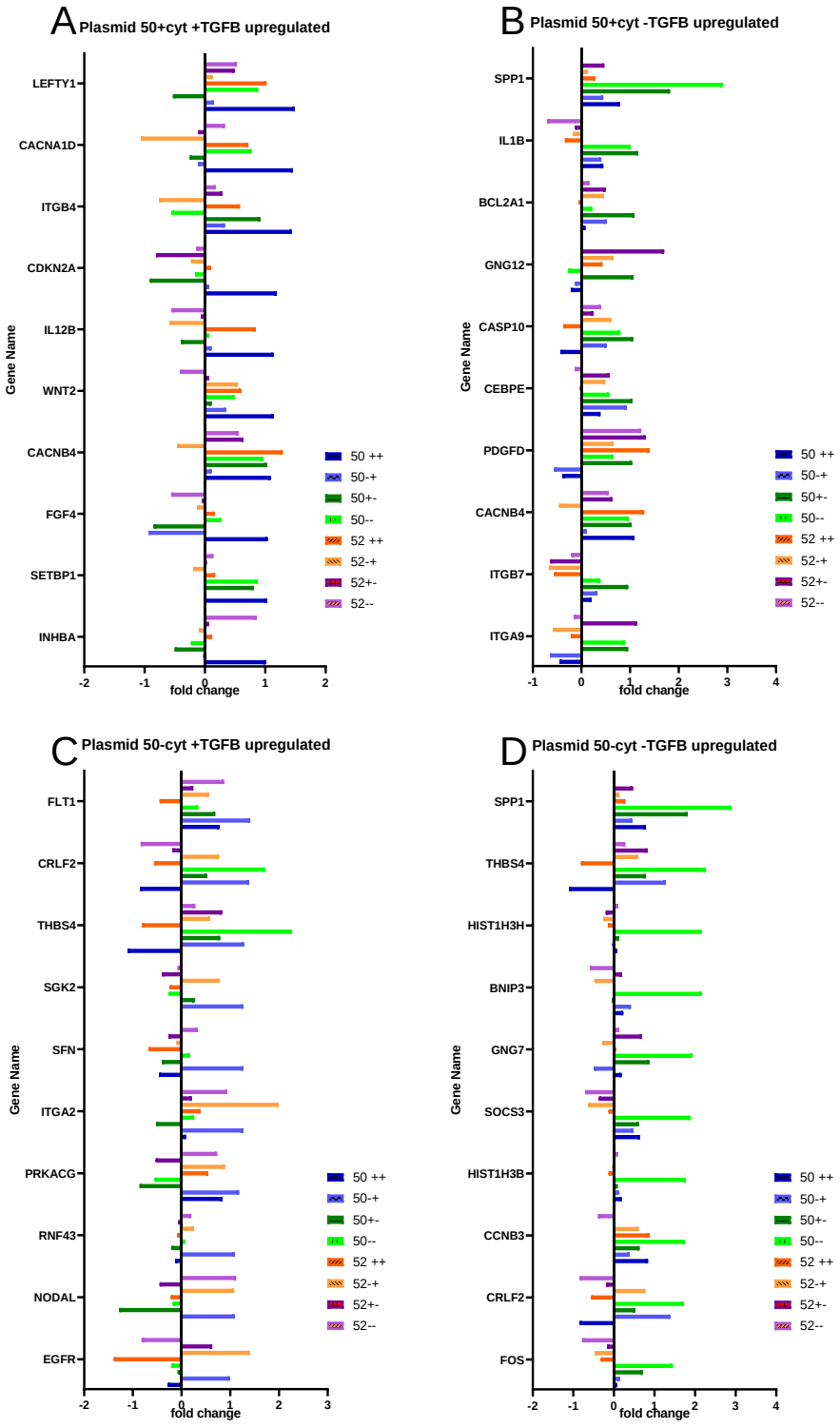

FIGURE 4.20. TOP 10 UPREGULATED GENES OF PLASMID 50. THE THP-1 CELL LINE WAS TRANSDUCED WITH PLASMID DNA CONTAINING SHRNA TO REDUCE THE EXPRESSION OF CD109, THESE CELLS WERE TREATED WITH A COMBINATION OF CYTARABINE AND TGFB1. DIFFERENCE IN GENE EXPRESSION COMPARED TO THP-1 CELLS WITHOUT CD109 SHRNA WAS CALCULATED AS A FOLD-CHANGE, WHICH WAS COMPARED TO THE DIFFERENT TREATMENT CONDITIONS.

In figure 4.21 shows the downregulation of genes in THP-1 cells transduced with plasmid 52. In figure 4.21A the downregulated genes when the cells were treated with both TGFβ1 and cytarabine are shown, notably MMP9 and RASFGRF1 are both down regulated in the plasmid 52 conditions compared to the plasmid 50 conditions. Overall, the genes shown in this figure are not correlated to a specific treatment or plasmid.

In figure 4.21B, the genes most downregulated in the THP-1 cells transduced with plasmid 52 treated with just cytarabine are shown. Interestingly, the gene GADD45G is highly downregulated all conditions without TGFβ1 but shows upregulation in those conditions with TGFβ1. The gene BMP2 was downregulated in both plasmid 50 and 52 treated with just cytarabine as opposed to all other conditions which showed slight upregulation. Both MMP9 and RASFGRF1 were part of the top 10 for this condition and plasmid 52 treated with both cytarabine and TGFβ1.

In figure 4.21C, the THP-1 cells transduced with plasmid 52 were treated with TGFβ1, the top 10 most downregulated genes are shown. The gene WNT7A was downregulated in both plasmid 50 and 52 treated with just TGFβ1 but was upregulated in all other conditions except for the untreated plasmid 52. Overall, the genes most downregulated in this condition are also downregulated in the other treatment conditions.

Figure 4.21D shows the top 10 downregulated genes associated with plasmid 52 transduced cells without treatment. The gene WNT2B was downregulated in both plasmid 50 and 52 untreated conditions but upregulated in all other conditions. Interestingly, the gene ITGA8 was down regulated in all conditions except for plasmid 50 and 52 treated with TGFβ1 and Cytarabine. Generally, the genes downregulated in this condition were also downregulated in the other conditions.

In figure 4.22, the genes most upregulated in THP-1 cells transduced with plasmid 52 in different treatment conditions are shown. In figure 4.22A the genes upregulated in cells treated with both TGFβ1 and cytarabine are shown. Overall, the genes upregulated in this condition are also upregulated in the other samples. The gene CNTFR is highly unregulated in this condition compared to the other samples which show a very small fold-change increase or decrease in comparison. In comparison to the other conditions, ACVRC1 also shows an upregulation although there were small fold-change increase or decrease in the other conditions. In all conditions with plasmid 52 LEF1 shows high upregulation, but a slight upregulation or a downregulation in the plasmid 50 conditions.

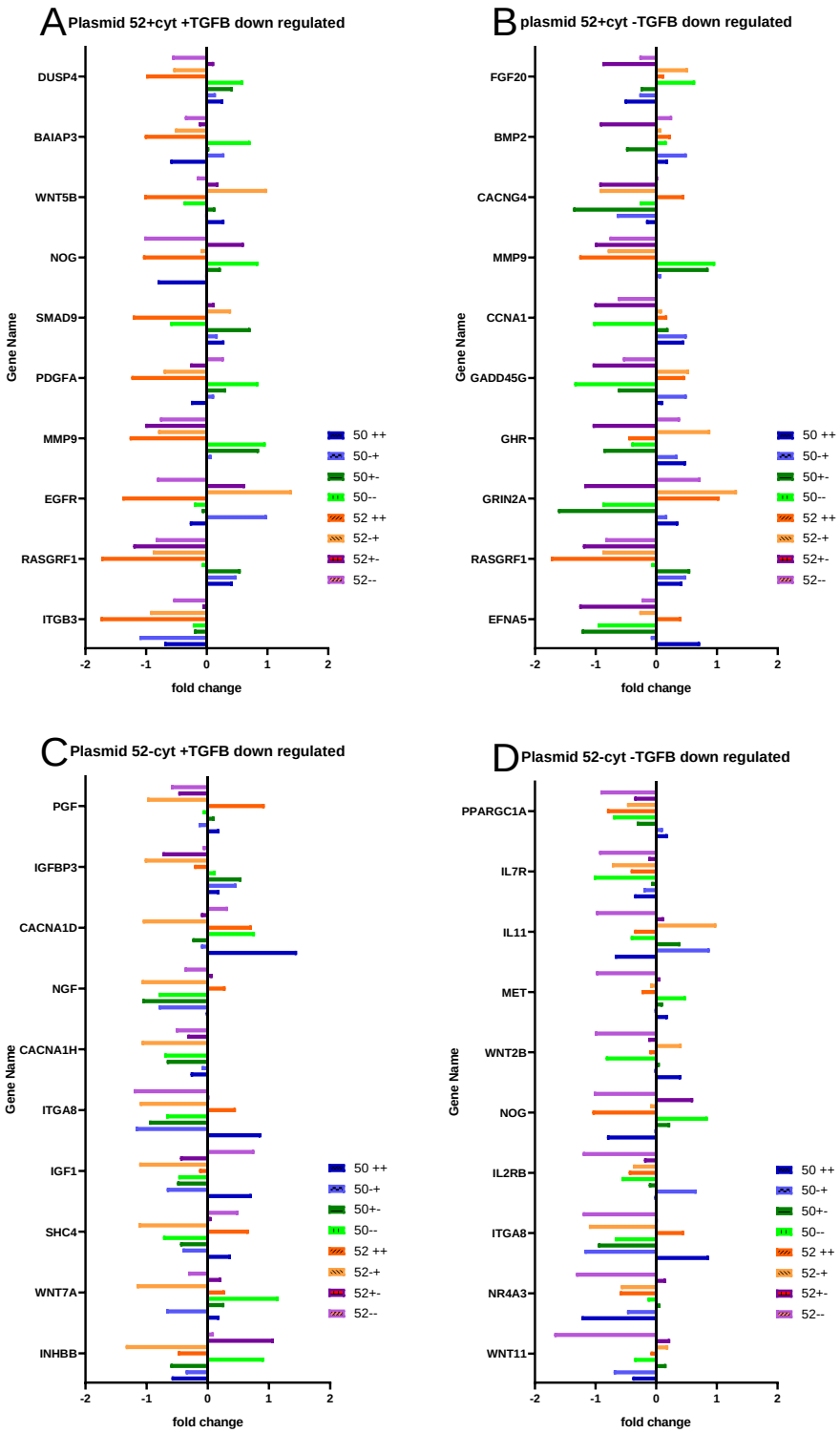

FIGURE 4.21. TOP 10 DOWNREGULATED GENES FOR PLASMID 52. THE THP-1 CELL LINE WAS TRANSDUCED WITH PLASMID DNA CONTAINING SHRNA TO REDUCE THE EXPRESSION OF CD109, THESE CELLS WERE TREATED WITH A COMBINATION OF CYTARABINE AND TGFB1. DIFFERENCE IN GENE EXPRESSION COMPARED TO THP-1 CELLS WITHOUT CD109 SHRNA WAS CALCULATED AS A FOLD-CHANGE, WHICH WAS COMPARED TO THE DIFFERENT TREATMENT CONDITIONS.

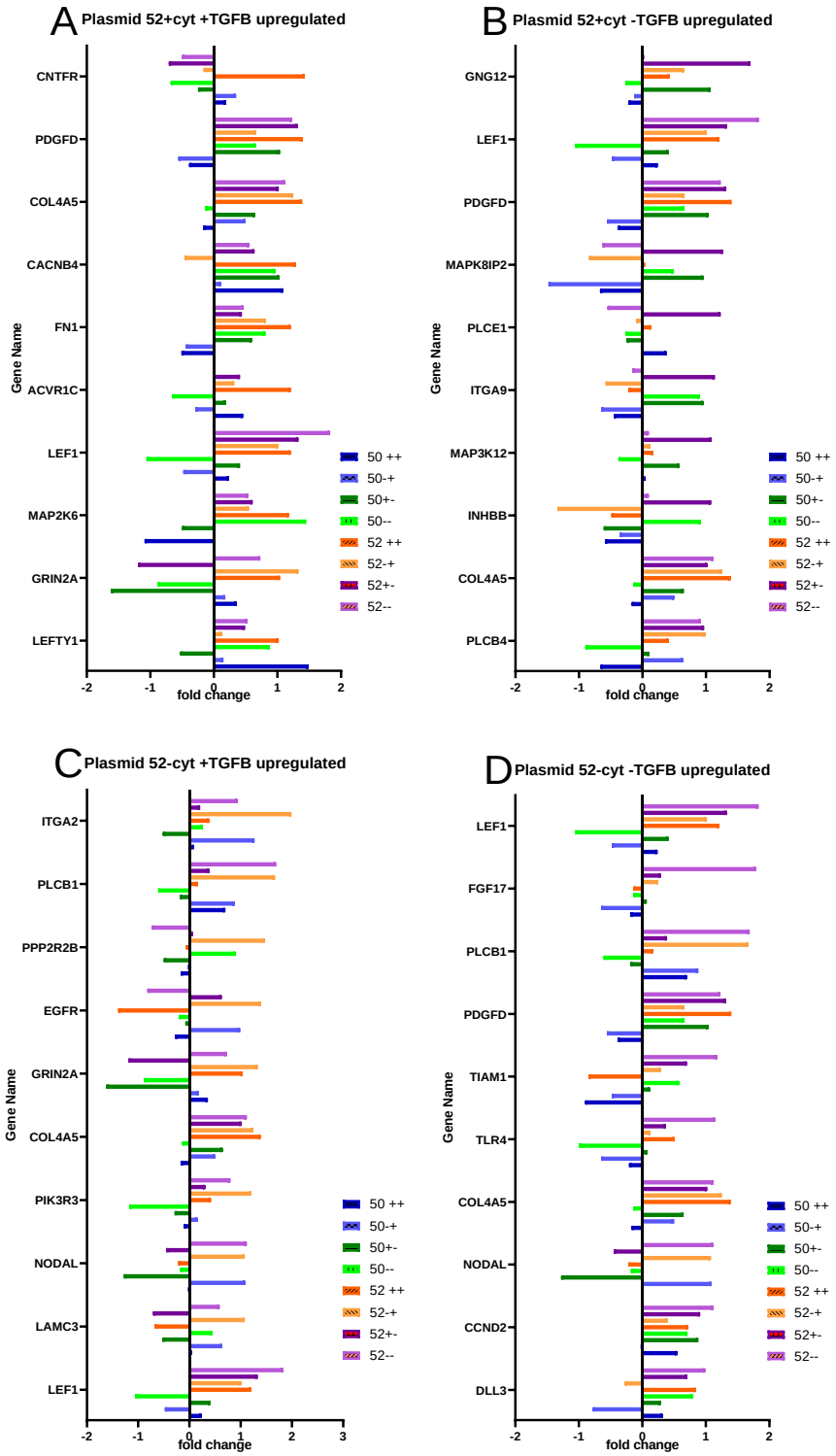

FIGURE 4.22. TOP 10 UPREGULATED GENES FOR PLASMID 52. THE THP-1 CELL LINE WAS TRANSDUCED WITH PLASMID DNA CONTAINING SHRNA TO REDUCE THE EXPRESSION OF CD109, THESE CELLS WERE TREATED WITH A COMBINATION OF CYTARABINE AND TGFB1. DIFFERENCE IN GENE EXPRESSION COMPARED TO THP-1 CELLS WITHOUT CD109 SHRNA WAS CALCULATED AS A FOLD-CHANGE, WHICH WAS COMPARED TO THE DIFFERENT TREATMENT CONDITIONS.

Figure 4.22B shows plasmid 52 cells treated with just cytarabine, although there was some variance the genes fold-change values were similar to the other conditions, with the exception of PLCE1 was upregulated in this condition but only showed minimal changes in fold-change for the different conditions.

In figure 4.22C the genes upregulated in plasmid 52 treated with TGFβ1 are shown, overall, the genes that are upregulated in this condition are also upregulated in the other conditions. The gene NODAL was upregulated in both plasmid 52 untreated, plasmid 52 treated with just TGFβ1 and the plasmid 50 condition also treated with just TGFβ1 but downregulated in the other conditions. The gene PIK3R3 was upregulated in all the plasmid 52 conditions but downregulated or show minimal fold change differences in the plasmid 50 conditions.

Figure 4.22D shows the genes most upregulated in cells transduced with plasmid 52 without treatment, generally, the genes upregulated in this condition are also upregulated in the other conditions. The gene FGF17 was highly upregulated in the untreated plasmid 52 cells but showed minimal fold-change differences in the other conditions. Both LEF1 and CPL4A5 shows upregulation in the plasmid 52 conditions but is downregulated or slightly upregulated in the plasmid 50 conditions.

Between figure 4.21 and 4.22 there were 13 genes which were common between the plasmid 52 up and down regulated genes, in comparison to just six for plasmid 50. Overall, there were 13 genes which appeared in both plasmid 50 and plasmid 52 analysis, showing there are common genes associated with the CD109 knock down cells.

4.4. Conclusion

CD109 has been associated with treatment resistance and relapse[218,234] in multiple cancers as well as cancer progression[139,140,187,188,215]. The results obtained in this chapter show reduced CD109 gene expression does not impact cell growth rate, however, as an upregulation has been shown to increase progression in patient samples, it is still crucial to understand the role of this gene and the associated protein. The expression of CD109 in AML cell lines has no quantitative comparison, it is unknown whether the expression is relatively high or low. An additional investigation looking at increased and decreased CD109 gene expression in AML, by knock-in and knock-out cell manipulation studies, may indicate if the gene has an impact at higher quantities. An alternative method of CD109 knockdown may decrease its expression to 0% which would highlight id this gene was critical to cellular function or an accessory to tumour progression.

Although increased CD109 expression is associated with relapse and treatment resistance[218], here the results indicate CD109 doesn't change the cells response to cytarabine treatment, therefor must have another role in cancer progression. In different cancers alternative drugs are used to treat the specific type of cancer, as this experiment was isolated to cytarabine in AML, a wider ranging study of different cancers and their associated treatments would give better perspective on the impact of CD109 and treatment response. Although the analysis focused on gene expression, the assays used focused on protein function, a better understanding of the relationship between transcript and protein is required.

Relatively little is known about CD109, only that it downregulates TGFBR1 protein[145,146,219] on the surface of cells, and the soluble[147] form also downregulates TGFβ1, any functionality of the gene transcript itself is unknown. As the results showed cells with reduced CD109 gene expression do not respond differently to TGFβ1 treatment, it is likely the protein expression of CD109 was unchanged or other pathways are also involved which have yet to be explained. Although the expression of CD109 did not alter the response to TGFβ1, the AML cell lines did respond as shown by the cell cycle arrest and the changes in the expression profiles of the different treatment conditions. To better understand the role of TGFβ1, treated cells could be analysed by mass spectrometry or through RNA sequencing to give a larger profile of the responsive cells.

Gene expression profiles do change depending on the treatment of the cell lines, with some commonality of the genes upregulated and downregulated in the cells with the same treatment but transduced with different plasmids. A larger comparison of the up and downregulated gene expression from the NanoString platform could show which pathways are active. To further investigate this, the cell lines should be treated again with a combination of cytarabine and TGFβ1, but the samples should be analysed in triplicate to ensure any differences are significant from the other conditions. In addition, the treated cells should be analysed using mass spectrometry to identify key changes in protein expression. Another interesting experiment would be to stain the treated cells with antibodies to identify changes in cellular location, as CD109 is a surface marker but is internalised with TGFBR1 when active.

4.4.1. Future work

During the analysis of gene expression in chapter three, CD109 was identified as a key marker in relapse AML. Through literature research, it was shown to be linked to treatment resistance, progression and poor overall survival in many[138–140,143,164,187,188,215,218] other cancers. Transcribed RNA is translated into protein; however, many RNA transcripts have functionality independent of their translated protein. The relationship between RNA expression and protein expression is not linear, meaning a gene transcribed in small quantities may show a high protein expression or vice versa. Relating the RNA expression to the protein expression for CD109 would be insightful in determining its influence on cancer progression and treatment response.

It is known that CD109 is a transmembrane protein[146], which promotes the internalisation and degradation of TGFBR1. It would be interesting to identify the cellular location of the CD109 protein in relation to the TGFBR1 protein, especially in different treatment conditions. By identifying the different cellular locations of CD109 and how this changes with treatment, the function of CD109 can be better explained.

This series of experiments focuses on a single dose of cytarabine without Daunorubicin: the standard treatment for AML includes two chemotherapeutics, cytarabine and Daunorubicin. The standard 7+3 treatment is a 7-day continuous infusion of Daunorubicin followed by a 3-day treatment of cytarabine, the two drugs are used in conjunction as they have two different rolls in the treatment of AML. Cytarabine integrates into the DNA and prevents DNA synthesis therefor halting cell cycle progression, whereas Daunorubicin is a DNA intercalator which prevents DNA replication, in addition it disrupts macromolecules, specifically preventing the re-joining of the DNA double helix once broken during DNA replication. One focuses on cell cycle arrest whereas the other promotes cell death. In future experiments, an adaption of the treatment schedule could be utilised to determine the full effects of CD109 on treatment resistance and minimal residual cell populations. Additionally, after the initial treatment schedule, any residual cell population should be harvested and expanded to see the differences in this resistant population, which would likely be responsible for relapse in a patient.

The three cell lines used in this investigation were from various different patients, but all were from relapse AML and all were male: The cell line THP-1 was line established from a one-year-old male at relapse, Kasumi-1 from a seven-year-old male at second relapse and MolM-13 from a 20-year-old male at relapse[235–237]. As the cell lines were from relapse

patients it is logical to assume, they may have some native resistance to cytarabine and may activate alternative pathways, owing to their survival. Additional cell lines which are from diagnostic samples would be ideal to test how CD109 expression changes treatment response; a comparison back to the relapse cell lines could highlight fundamental differences. To better investigate the impact of CD109, patient samples at different clinical time points could be analysed and correlated with the treatment received. Different clinical time points, for example diagnosis, relapse and remission, could elucidate changes of CD109 expression over time and understand its role in cancer progression.

Chapter 5 - Discussion and future work

5.1. Introduction

Cancer is a complex group of diseases that share the same fundamental characteristics of uncontrolled proliferation and the ability to metastasis, and as such here are many different types of cancer affecting every tissue in the body. It is said that 50% of the population will get cancer, but cancer affects everyone, including friends, family and the team caring for the patient.

Most cancers are solid tumours, however, there are also an array of blood cancers. There are four main subtypes of blood cancer, AML, CML, ALL and CLL; of all these, AML has the highest mortality rate and remains the second most common type of blood cancer. The survival rate for AML is very poor at just 30%[64,73,238]. Around 50% of AML patient achieve remission, of those patients, 50% will relapse within three years, which is associated with poor prognosis[239].

After treatment for AML, a sample is taken and tested for minimal residual disease[73,105,108,239], modern tests are far more sensitive compared to the older tests, allowing MRD to be detected even in small quantities, not just through the observation of abnormal cells but through molecular testing[240]. Relapse is associated with MRD, and resistant sub-populations of cells present at diagnosis[104,106,241,242]. The resistant sub-populations have an evolutionary survival advantage over other AML cells in the bone marrow[241–243], which means these cells resist treatment and can re-colonise the bone marrow. AML has many leukemic stem cell properties which affords them chemo-resistance and the ability to self-renew[241–243]. Cell can also acquire mutations during the treatment process, which provide resistance to treatment[241–243], and the new sub-clone can remain in-situ.

Biomarkers are key tools used in the clinical environment, which are a measurable biological component that is characteristic of a particular disease or disease state[166]. Of particular interest are molecular biomarkers, which can be obtained from a patient sample and measured by protein, DNA or RNA expression. There are seven main types of biomarker, all with different applications[111,120], although the primary biomarker type of interest in this investigation is the prognostic biomarker. A prognostic biomarker allows the prediction of a particular outcome, in this instance, determining which patients will relapse and which will not.

As relapse is so prevalent in AML[239,241,242] there is high demand for a clinical tool to predict relapse before it happens. Patients identified as high-risk would benefit from closer monitoring and alternative or more aggressive treatment option. As such the first key aim of this investigation was to develop a biomarker panel to predict relapse in AML. In the process of biomarker discovery, different methods were trialled, and a robust workflow was developed, which will speed up the discovery pipeline in the future. By combining machine learning and statistical approaches of biomarker discovery, a workflow that is quick and easy to understand has been developed, which was the second key aim of this investigation. As part of the biomarker discovery workflow clinically relevant genes were identified; the third aim of this investigation is to determine the impact of clinically relevant genes on AML.

5.2. Discussion

5.2.1. Methods of biomarker discovery

There are several methods of biomarker discovery used in industry, ranging from computational machine learning[129] to statistical analysis[111], each with its own advantages and disadvantages[179]. Each method of biomarker discovery is valid and has the potential to discover a robust biomarker panel. A drawback with heavily computational methods is the lack of understanding and trust by the wider scientific community, and as such the application of the discovered biomarkers is somewhat limited.

For this investigation a variety of biomarker discovery methods were implemented and combined to determine which gave the best overall panel of biomarkers, in the context of relapse in AML. A key aim of the workflows of biomarker discovery was to achieve transparency and ensure it was understandable by non-experts, this encourages trust in the prognostic panel discovered and allows for wider application.

Six different workflows were developed and tested, each returning a unique panel of biomarkers. The biomarker panels were discovered in the HOVON dataset and validated in two additional datasets, TCGA and TARGET, to determine which could be applied beyond the discovery dataset. Form the six different workflows tested in this investigation, the binary logistic regression workflow was the best overall: the panel was significant in both adult datasets, HOVON and TCGA, which was not seen in the other workflows. The BLR workflow was not significant in the TARGET dataset, which is a childhood dataset, and therefore some intrinsic differences are expected which cannot be controlled including increased survival and increased relapse free survival.

The panel discovered through the BLR workflow comprised of 9 different genes: SOCS2-AS1[191], HAL[192], NYNRIN[196], MXRA7[199], AGTPBP1[200], LAPTM4B[201], MPO[207], PYROXD1[208] and FEM1C[210]. Each of the genes in the panel can independently determine which patients are at risk of relapse, however in combination they provide a powerful prognostic tool which will determine this to a higher degree of accuracy. Through literature review, each of the genes or their associated protein has been linked to cancer. Many of the genes discovered in this panel are linked to cell cycle control, which is usually tightly regulated in proliferation, but dysregulated in cancer. The genes discovered here provides a key insight to target genes, and cellular processes even before a risk score has been calculated.

The elements involved in the BLR workflow are transparent and trusted statistical methods[244,245], which allows for a wider audience to understand the workflow and apply it. The workflow has the potential to be utilised for different clinical outcomes and for different conditions. As the constituting elements of the workflow don't require a time point, only a binary condition, many other clinical outcomes can be investigated. As the workflow is based on statistics this can be easily coded in R, which would ensure its correct application and the reproducibility.

The score developed in this investigation has the potential to indicate which patients are at risk of relapse and provide assist clinicians by identifying patients in need of additional monitoring and more aggressive treatment. However, as this score was only validated in patient datasets it is uncertain how it will perform in a clinical setting and which platform is best for identifying these patients. Additional testing is required in patient samples using equipment most commonly found in a biomedical setting, likely a qPCR machine. Firm boundaries are required when identifying patients who are likely to relapse, as such the score developed here would benefit from further refinement.

5.2.2. CD109 in AML

During the analysis of gene expression in relation to relapse in AML, the gene CD109 was identified as a key marker. The gene CD109 had the highest AUC of all the genes in the HOVON dataset, when analysed using ROC curves, indicating it can independently predict relapse in AML. Through literature research, CD109 has been linked to treatment resistance[218], progression[140,246] and poor overall survival[164,187,188] in many other cancers. It is known that CD109 is a transmembrane protein, which promotes the internalisation and degradation of TGFBR1[143,144,146], in turn reducing the signalling potential.

As increased CD109 expression was found to be associated with adverse outcomes in AML and other cancer, this gene was investigated further in the context of AML. In this investigation AML cells were transduced with plasmid DNA that confers CD109 shRNA to reduce the expression of CD109. The cells with reduced CD109 expression were treated with TGFβ1 and Cytarabine, at a dose dependent on the cell line. To determine the impact of CD109 expression on treatment response and survival, the growth rate was measured, and gene expression changes were analysed.

The results obtained show reduced CD109 gene expression does not impact cell growth rate, indicating is not a critical component for cell cycle progression. Because CD109 is a co-receptor for TGFBR1, the cells were treated with TGFβ1 to determine the impact on cell growth; this induced cell cycle arrest in the AML cell lines, however down regulation of CD109 had no impact on the magnitude of the response. The cell lines Kasumi-1 and MolM-13 had the most dramatic reduction in proliferation, which was completely arrested. In THP-1 cells the proliferation rate was reduced but not to the same degree as the other cell lines. It has previously been established that TGFβ1 induces cell cycle arrest[226,247,248], however, this cytokine has also been implicated in differentiation[225,249–251] and metastasis[227,228,252,253], these processes are both heavily reliant on gene expression changes[251,252,254–257]. Although TGFβ1 has many roles, the specific pathway activated will depend heavily on the microenvironment and other cytokines present. It is important to note the concentration of TGFβ1 was high in comparison to what would be expected in a biological context.

As CD109 has been associated with treatment resistance in the literature[218], the effect of CD109 expression and the response to cytarabine was investigated. As it is known that TGFβ1 has a roll in microenvironment a combination of Cytarabine and TGFβ1 were used to treat the cells. Four main conditions were investigated, untreated, cytarabine, TGFβ1 and Cytarabine with TGFβ1. It was shown that cytarabine induces cell cycle arrest in AML cell lines, but the expression of CD109 did not affect the magnitude of response. When cell lines were treated with cytarabine in combination with TGFβ1, the cell cycle was arrested and the cells began to undergo apoptosis, again the expression of CD109 did not impact the response to treatment. Although the expression of CD109 did not impact the response to treatment through the observation of cell growth curves, the gene expression profiles do change depending on the treatment conditions. Some commonality was observed between the genes upregulated and downregulated in the cells with the same treatment but transduced with different plasmids. Two key genes observed were NODAL and LEFTY1[258,259], both are secreted ligands of the TGFβ family of proteins but have largely different roles in the TGF

signalling pathway. NODAL is an activator of TGFβ signalling through the recruitment of SMAD, whereas LEFTY1 is an inhibitor of TGFβ signalling predominantly the inhibition of the NODAL/TGFβ interaction[260].

Although reducing CD109 expression did not alter the response to treatment in terms of growth rate in AML cell lines, the cells do respond through modified gene expression profiles. The changes in gene expression profiles indicate alternative pathways are activated through the treatment. A further analysis of the gene expression profiles will provide a better understanding of the role CD109 plays in cancer progression, relapse and treatment resistance. To better understand the role of TGFβ1 in treatment response, de-novo AML cell lines should be used as opposed to the cell lines obtained from relapse AML, which likely already have a resistance to treatment.

5.3. Conclusion

Overall, this investigation has generated a biomarker panel that can be used to predict relapse in AML with an accuracy of 76% in the HOVON dataset and 67% in the TCGA dataset. The panel consisted of 9 genes, many of which have been previously identified in different cancers and are heavily involved in cell cycle control. This panel can be used to aid clinicians in determining which patients should receive additional monitoring and alternative treatments, with an aim to improve overall survival in AML. Although this score provides an excellent foundation, the score should be validated further in additional datasets and patient samples, ensuring it can be used in a clinical setting.

Through the development of the biomarker discovery workflow a consistent and relatively fast method has been established for the discovery of clinically relevant biomarkers. Currently there are very few clinically relevant biomarkers available, however this workflow aims to speed up this process and provide an accurate panel for the desired clinical annotation. The binary logistic regression workflow is simple to follow and can be applied in a variety of situations provided there is a binary event.

CD109 was identified as a clinically relevant gene in relation to relapse, and though literature searching it was implicated in treatment resistance and cancer progression. This investigation established that a decreased expression of CD109 did not alter the response to TGFβ1 and cytarabine, in that the cell proliferation rate was no different to cells with normal levels of CD109 expression. However, there was altered gene expression in cells with reduced CD109 expression in relation to different treatment conditions. A more in-depth

investigation would provide a better understanding of the pathways activated and deactivate in response to TGFβ1 and cytarabine, and how these pathways are altered with CD109 expression.

It is important to recognise the benefits of the work already done, but also see the conclusions as a steppingstone on the way to a fuller understanding of relapse in AML and the impact of CD109 in cancer.

5.4. Future work

5.4.1. Reverse engineering the score

Although the score overall is consistent between the HOVON and TCGA datasets, a drawback of the score created is that it relies on median splits to determine whether a patient is "high" or "low" risk with arbitrary boundary determined by the individual dataset, meaning the score is less applicable to a clinical setting. The score needs to be more applicable to a clinical setting, and a fundamental change in the normalisation process will facilitate this.

The current process of normalisation begins after the data is collected from an array or RNA-sequencing platform. The data is quality checked and RMA normalised to ensure the data is consistent within a dataset, at this point the data is used for biomarker discovery. In future work the datasets can be merged, and a-quantile normalisation can be used to ensure consistency between datasets, this will facilitate direct comparison between platforms.

Using the processed data, the discovery workflows can be trialled again, and it will become apparent whether the score and workflow is good or if a different workflow and panel overtake its prognostic benefits. After the overall panel has been determined and validated, the original data, before a-quantile normalisation, could be used to re-calculate the score and determine boundaries for each of the patients in the datasets. This should give a higher predicting score with better more specific boundaries facilitating increased discrimination between high and low risk patients.

5.4.2. Application to other cancers

AML is a small part of the cancer story; although AML is the largest cause of leukaemia deaths in the UK, many other cancers have a higher incidence, mortality and relapse rate. The workflows developed for biomarker discovery should be applied to other cancers and other

clinically relevant points. This will allow for the workflows to be validated and a useful biomarker panel to be developed.

5.4.3. CD109 in AML

To fully understand the role of CD109 in AML native expression of CD109 gene and protein should be measured, this would indicate the relationship between gene and protein expression, whether it is solely an upregulation of the gen or if the protein is also over expressed. By indicating a base line expression, the cells can be transduced to again reduce the expression of CD109 and separately to increase the expression of CD109, again furthering the understanding of the translational relationship. The full range of differently expressed CD109 cells can be treated with TGFβ1 and cytarabine to determine the effect on cell behaviour and survival. Ideally the cells would be analysed with both RNA sequencing and mass-spectrometry to fully establish the changes that occur on a molecular level.

The standard treatment for AML uses a combination of Daunorubicin and cytarabine, whereas this series of experiments focused on a single dose of cytarabine. To improve on this existing work an adapted treatment schedule could be utilised, to determine the full effects of CD109 on treatment response. This would give a wider view of the effects of CD109 including treatment resistance, proliferation, apoptosis and activated signalling pathways, in a more accurate experimental design.

As CD109 is known to bind to and internalise TGFBR1, it would be interesting to identify the cellular location of the CD109 protein in relation to the TGFBR1 protein, especially in different treatment conditions. This could be achieved through antibody staining of each of the conditions, for both CD109 and TGFBR1. By identifying the different cellular locations of CD109 and how this change with treatment, the function of CD109 can be better explained. This experiment would determine if the over expression of CD109 causes increased function in the cancerous cells.

5.4.4. CD109 in other cancers

The gene CD109 was identified in most of the workflows used in this investigation. The gene CD109 had the highest AUC out of all the genes present in the HOVON dataset, indicate it has a key role in relapse. This gene has also been implicated in treatment resistance, relapse and progression in other cancers, showing it has an important influence on cancer as a whole, as this gene has such an impact in cancer it warrants further investigation. Through dataset

searching and literature review the larger story of CD109 can be shown, particularly looking at the association of CD109 to outcomes in different clinical datasets.

To fully investigate the role of CD109, cell lines from a variety of cancers can be transduced with knock-down plasmid DNA and knock-in DNA to determine how both under and over expression changes cellular behaviour. These cells can be treated with TGFβ1 and chemotherapeutic drugs to determine if increased or decreased CD109 expression changes response to treatment. Ideally the protein expression changes would be analysed using mass-spectrometry and the gene expression changes analysed using RNA sequencing.

REFERENCES

1. Moan, J., Grigalavicius, M., Baturaite, Z., Dahlback, A. & Juzeniene, A. The relationship between UV exposure and incidence of skin cancer. *Photodermatol. Photoimmunol. Photomed.* **31**, 26–35 (2015).

2. Singh, N. & Singh, S. Interstitial Lung Diseases and Air Pollution: Narrative Review of Literature. *Pulm. Ther. 2021 71* **7**, 89–100 (2021).

3. Lewtas, J. Air pollution combustion emissions: Characterization of causative agents and mechanisms associated with cancer, reproductive, and cardiovascular effects. *Mutat. Res. Mutat. Res.* **636**, 95–133 (2007).

4. Statistics on preventable cancers | Cancer Research UK. Available at: https://www.cancerresearchuk.org/health-professional/cancer-statistics/risk/preventable-cancers#heading-Two. (Accessed: 28th August 2021)

5. Anand, P. *et al.* Cancer is a Preventable Disease that Requires Major Lifestyle Changes. *Pharm. Res. 2008 259* **25**, 2097–2116 (2008).

6. Caporaso, N. *et al.* Genome-wide and candidate gene association study of cigarette smoking behaviors. *PLoS One* **4**, (2009).

7. Overweight and obesity statistics | Cancer Research UK. Available at: https://www.cancerresearchuk.org/health-professional/cancer-statistics/risk/overweight-and-obesity#heading-Zero. (Accessed: 28th August 2021)

8. Avgerinos, K. I., Spyrou, N., Mantzoros, C. S. & Dalamaga, M. Obesity and cancer risk: Emerging biological mechanisms and perspectives. *Metabolism* **92**, 121–135 (2019).

9. Teng, X. & Hardwick, J. M. Cell Death in Genome Evolution. *Semin. Cell Dev. Biol.* **39**, 3 (2015).

10. Teng, S., Wang, L., Srivastava, A. K., Schwartz, C. E. & Alexov, E. Structural Assessment of the Effects of Amino Acid Substitutions on Protein Stability and Protein-Protein Interaction. *Int. J. Comput. Biol. Drug Des.* **3**, 334 (2010).

11. Oliver, T. R. *et al.* New Insights into Human Nondisjunction of Chromosome 21 in Oocytes. *PLoS Genet.* **4**, 1000033 (2008).

12. Gottlieb, S. F., Tupper, C., Kerndt, C. C. & Tegay, D. H. Genetics, Nondisjunction. *StatPearls* (2021).

13. Kakizuka, A. *et al.* Chromosomal translocation t(15;17) in human acute promyelocytic leukemia fuses RARα with a novel putative transcription factor, PML. *Cell* **66**, 663–674 (1991).

14. Goldman, S. L. *et al.* Epigenetic Modifications in Acute Myeloid Leukemia: Prognosis, Treatment, and Heterogeneity. *Front. Genet.* **10**, 133 (2019).

15. Li, S. & Wu, X. Common fragile sites: protection and repair. *Cell Biosci. 2020 101* **10**, 1–9 (2020).

16. Kumar, R. *et al.* HumCFS: a database of fragile sites in human chromosomes. *BMC Genomics 2019 199* **19**, 1–8 (2019).

17. Garber, K. B., Visootsak, J. & Warren, S. T. Fragile X syndrome. *Eur. J. Hum. Genet. 2008 166* **16**, 666–672 (2008).

18. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).

19. Trendowski, M. The inherent metastasis of leukaemia and its exploitation by sonodynamic therapy. *Crit. Rev. Oncol. Hematol.* **94**, 149–163 (2015).

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

21. Lee, P.-H. & Osley, M. A. Who gets a license: DNA synthesis in quiescent cells re-entering the cell cycle. *Curr. Genet. 2021 674* **67**, 539–543 (2021).

22. Molinari, M. Cell cycle checkpoints and their inactivation in human cancer. *Cell Prolif.* **33**, 261–274 (2000).

23. Kastan, M. B. & Bartek, J. Cell-cycle checkpoints and cancer. *Nat. 2004 4327015* **432**, 316–323 (2004).

24. Miles, S. & Breeden, L. A common strategy for initiating the transition from proliferation to quiescence. *Curr. Genet. 2016 632* **63**, 179–186 (2016).

25. Amin, A. R. M. R. *et al.* Evasion of anti-growth signaling: a key step in tumorigenesis and potential target for treatment and prophylaxis by natural compounds. *Semin. Cancer Biol.* **35 Suppl**, S55 (2015).

26. McFarland, C. D., Korolev, K. S., Kryukov, G. V., Sunyaev, S. R. & Mirny, L. A. Impact of deleterious passenger mutations on cancer progression. *Proc. Natl. Acad. Sci.* **110**, 2910–2915 (2013).

27. Kroemer, G. *et al.* Classification of cell death: Recommendations of the Nomenclature Committee on Cell Death 2009. *Cell Death and Differentiation* **16**, 3–11 (2009).

28. Lowe, S. W. & Lin, A. W. Apoptosis in cancer. *Carcinogenesis* **21**, 485–495 (2000).

29. Fernald, K. & Kurokawa, M. Evading apoptosis in cancer. *Trends Cell Biol.* **23**, 620 (2013).

30. Pitolli, C. *et al.* Do Mutations Turn p53 into an Oncogene? *Int. J. Mol. Sci. 2019, Vol. 20, Page 6241* **20**, 6241 (2019).

31. Prior, I. A., Lewis, P. D. & Mattos, C. A Comprehensive Survey of Ras Mutations in Cancer. *Cancer Res.* **72**, 2457–2467 (2012).

32. Banker, D. E. *et al.* The t(8;21) translocation is not consistently associated with high Bcl-2 expression in de novo acute myeloid leukemias of adults. *Clin. Cancer Res.* **4**, (1998).

33. Zhou, J. *et al.* BCL2 overexpression: clinical implication and biological insights in acute myeloid leukemia. *Diagn. Pathol.* **14**, (2019).

34. Konopleva, M. & Letai, A. BCL-2 inhibition in AML: an unexpected bonus? *Blood* **132**, 1007–1012 (2018).

35. McBride, A. *et al.* The Role of Inhibition of Apoptosis in Acute Leukemias and Myelodysplastic Syndrome. *Front. Oncol.* **9**, 192 (2019).

36. Wei, Y. *et al.* Targeting Bcl-2 Proteins in Acute Myeloid Leukemia. *Front. Oncol.* **10**, (2020).

37. Bories, P. *et al.* Impact of TP53 mutations in acute myeloid leukemia patients treated with azacitidine. *PLoS One* **15**, e0238795 (2020).

38. Prokocimer, M., Molchadsky, A. & Rotter, V. Dysfunctional diversity of p53 proteins in adult acute myeloid leukemia: projections on diagnostic workup and therapy. *Blood* **130**, 699 (2017).

39. AM, H. & DA, S. Current status and new treatment approaches in TP53 mutated AML. *Best Pract. Res. Clin. Haematol.* **32**, 134–144 (2019).

40. Kadia, T. M. *et al.* TP53 mutations in newly diagnosed acute myeloid leukemia: Clinicomolecular characteristics, response to therapy, and outcomes. *Cancer* **122**, 3484–3491 (2016).

41. Quintás-Cardama, A. *et al.* p53 pathway dysfunction is highly prevalent in acute myeloid leukemia independent of TP53 mutational status. *Leuk. 2017 316* **31**, 1296–1305 (2016).

42. Shay, J. W. & Wright, W. E. Role of telomeres and telomerase in cancer. *Semin. Cancer Biol.* **21**, 349–353 (2011).

43. Artandi, S. E. & Attardi, L. D. Pathways connecting telomeres and p53 in senescence, apoptosis, and cancer. *Biochem. Biophys. Res. Commun.* **331**, 881–890 (2005).

44. Rajabi, M. & Mousa, S. A. The Role of Angiogenesis in Cancer Treatment. *Biomed. 2017, Vol. 5, Page 34* **5**, 34 (2017).

45. Carmeliet, P. VEGF as a Key Mediator of Angiogenesis in Cancer. *Oncology* **69**, 4–10 (2005).

46. Ferrara, N. VEGF and the quest for tumour angiogenesis factors. *Nat. Rev. Cancer 2002 210* **2**, 795–803 (2002).

47. Jeon, M. *et al.* HIF1A overexpression using cell-penetrating DNA-binding protein induces angiogenesis in vitro and in vivo. *Mol. Cell. Biochem. 2017 4371* **437**, 99–107 (2017).

48. Chatterjee, N. & Walker, G. C. Mechanisms of DNA damage, repair and mutagenesis. *Environ. Mol. Mutagen.* **58**, 235 (2017).

49. Comen, E. A., Bowman, R. L. & Kleppe, M. Underlying Causes and Therapeutic Targeting of the Inflammatory Tumor Microenvironment. *Front. Cell Dev. Biol.* **6**, 56 (2018).

50. Grivennikov, S. I., Greten, F. R. & Karin, M. Immunity, Inflammation, and Cancer. *Cell* **140**, 883 (2010).

51. Houghton, A. N. & Guevara-Patiño, J. A. Immune recognition of self in immunity against cancer. *J. Clin. Invest.* **114**, 468 (2004).

52. Anderson, K. G., Stromnes, I. M. & Greenberg, P. D. Obstacles posed by the tumor microenvironment to T cell activity: a case for synergistic therapies. *Cancer Cell* **31**, 311 (2017).

53. Cornel, A. M., Mimpen, I. L. & Nierkens, S. MHC Class I Downregulation in Cancer: Underlying Mechanisms and Potential Targets for Cancer Immunotherapy. *Cancers (Basel).* **12**, 1–33 (2020).

54. Jones, R. G. & Thompson, C. B. Tumor suppressors and cell metabolism: a recipe for cancer growth. *Genes Dev.* **23**, 537 (2009).

55. Orang, A. V., Petersen, J., McKinnon, R. A. & Michael, M. Z. Micromanaging aerobic respiration and glycolysis in cancer cells. *Mol. Metab.* **23**, 98 (2019).

56. Adekola, K., Rosen, S. T. & Shanmugam, M. Glucose transporters in cancer metabolism. *Curr. Opin. Oncol.* **24**, 650 (2012).

57. Ancey, P.-B., Contat, C. & Meylan, E. Glucose transporters in cancer – from tumor cells to the tumor microenvironment. *FEBS J.* **285**, 2926–2943 (2018).

58. Barea, F. & Bonatto, D. Relationships among carbohydrate intermediate metabolites and DNA damage and repair in yeast from a systems biology perspective. *Mutat. Res. Mol. Mech. Mutagen.* **642**, 43–56 (2008).

59. Cancer Research UK. Cancer Statistics for the UK. Available at: https://www.cancerresearchuk.org/health-professional/cancer-statistics-for-the-uk. (Accessed: 15th November 2018)

60. Cancer Research UK. Cancer incidence for common cancers. (2016). Available at: http://www.cancerresearchuk.org/health-professional/cancer-statistics/incidence/common-cancers-compared#heading-Three. (Accessed: 23rd February 2018)

61. Cancer mortality for all cancers combined | Cancer Research UK. Available at: https://www.cancerresearchuk.org/health-professional/cancer-statistics/mortality/all-cancers-combined#heading-Zero. (Accessed: 24th August 2021)

62. Cancer risk statistics | Cancer Research UK. Available at: https://www.cancerresearchuk.org/health-professional/cancer-statistics/risk#heading-One. (Accessed: 24th August 2021)

63. Seita, J. & Weissman, I. L. Hematopoietic stem cell: self-renewal versus differentiation. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **2**, 640–53 (2010).

64. Cancer Research UK. Acute myeloid leukaemia (AML) incidence statistics | Cancer Research UK. Available at: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/leukaemia-aml/incidence. (Accessed: 5th August 2019)

65. Cancer Research UK. Acute lymphoblastic leukaemia (ALL) incidence statistics | Cancer Research UK. Available at: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/leukaemia-all/incidence. (Accessed: 5th August 2019)

66. Cancer Research UK. Chronic lymphocytic leukaemia (CLL) incidence statistics | Cancer Research UK. Available at: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/leukaemia-cll/incidence. (Accessed: 5th August 2019)

67. Cancer Research UK. Chronic myeloid leukaemia (CML) incidence statistics | Cancer Research UK. Available at: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/leukaemia-cml/incidence. (Accessed: 5th August 2019)

68. J, B. & ES, W. Gemtuzumab ozogamicin for the treatment of acute myeloid leukemia. *Expert Rev. Clin. Pharmacol.* **11**, 549–559 (2018).

69. Stein, E. M. *et al.* Enasidenib in mutant IDH2 relapsed or refractory acute myeloid leukemia. *Blood* **130**, 722–731 (2017).

70. Medinger, M. & Passweg, J. R. Acute myeloid leukaemia genomics. *Br. J. Haematol.* **179**, 530–542 (2017).

71. Arber, D. A. *et al.* The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood* **127**, 2391–2405 (2016).

72. Döhner, H. *et al.* Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet. *Blood* **115**, 453–74 (2010).

73. Döhner, H. *et al.* Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood* **129**, 424–447 (2017).

74. Acute myeloid leukaemia - Symptoms - NHS. Available at: https://www.nhs.uk/conditions/acute-myeloid-leukaemia/symptoms/. (Accessed: 4th June 2020)

75. Döhner, H., Weisdorf, D. J. & Bloomfield, C. D. Acute Myeloid Leukemia. *N. Engl. J. Med.* **373**, 1136–1152 (2015).

76. Cancer Research UK. About acute myeloid leukaemia. Available at: https://www.cancerresearchuk.org/about-cancer/acute-myeloid-leukaemia-aml/about-acute-myeloid-leukaemia. (Accessed: 7th August 2019)

77. Shafer, D. & Grant, S. Update on rational targeted therapy in AML. *Blood Rev.* **30**, 275–283 (2016).

78. Huynh, T. T. *et al.* Retinoic acid and arsenic trioxide induce lasting differentiation and demethylation of target genes in APL cells. *Sci. Rep.* **9**, 9414 (2019).

79. Mueller, B. U. *et al.* ATRA resolves the differentiation block in t(15;17) acute myeloid leukemia by restoring PU.1 expression. *Blood* **107**, 3330–8 (2006).

80. Grignani, F. *et al.* The acute promyelocytic leukemia-specific PML-RAR alpha fusion protein inhibits differentiation and promotes survival of myeloid precursor cells. *Cell* **74**, 423–31 (1993).

81. Cull, E. H. & Altman, J. K. Contemporary treatment of APL. *Curr. Hematol. Malig. Rep.* **9**, 193–201 (2014).

82. Burnett, A. K. ACUTE MYELOID LEUKEMIA: TREATMENTOF ADULTS UNDER 60 YEARS. *Rev. Clin. Exp. Hematol.* **6**, 26–45 (2002).

83. Ley, T. J. *et al.* Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.* **368**, 2059–2074 (2013).

84. Sievers, E. L. *et al.* Efficacy and safety of gemtuzumab ozogamicin in patients with CD33-positive acute myeloid leukemia in first relapse. *J. Clin. Oncol.* **19**, 3244–54 (2001).

85. Amadori, S. *et al.* Gemtuzumab ozogamicin (Mylotarg®) as single-agent treatment for frail patients 61 years of age and older with acute myeloid leukemia: final results of AML-15B, a phase 2 study of the European Organisation for Research and Treatment of Cancer and Gruppo Italiano Malattie Ematologiche dell'Adulto

Leukemia Groups. *Leukemia* **19**, 1768–1773 (2005).

86. R, D., R, B., R, S., P, B. & RK, R. Enasidenib: First Mutant IDH2 Inhibitor for the Treatment of Refractory and Relapsed Acute Myeloid Leukemia. *Anticancer. Agents Med. Chem.* **18**, 1936–1951 (2018).

87. Issa, G. C. & DiNardo, C. D. Acute myeloid leukemia with IDH1 and IDH2 mutations: 2021 treatment algorithm. *Blood Cancer J. 2021 116* **11**, 1–7 (2021).

88. Yang, H., Ye, D., Guan, K.-L. & Xiong, Y. IDH1 and IDH2 mutations in tumorigenesis: mechanistic insights and clinical perspectives. *Clin. Cancer Res.* **18**, 5562 (2012).

89. FDA approves ivosidenib as first-line treatment for AML with IDH1 mutation | FDA. Available at: https://www.fda.gov/drugs/resources-information-approved-drugs/fda-approves-ivosidenib-first-line-treatment-aml-idh1-mutation. (Accessed: 27th August 2021)

90. Tibsovo: Withdrawn application | European Medicines Agency. Available at: https://www.ema.europa.eu/en/medicines/human/withdrawn-applications/tibsovo. (Accessed: 27th August 2021)

91. Roboz, G. J. *et al.* Ivosidenib induces deep durable remissions in patients with newly diagnosed IDH1-mutant acute myeloid leukemia. *Blood* **135**, 463–471 (2020).

92. Wei, Y. *et al.* Targeting Bcl-2 Proteins in Acute Myeloid Leukemia. *Front. Oncol.* **10**, (2020).

93. FDA approves venetoclax in combination for AML in adults | FDA. Available at: https://www.fda.gov/drugs/fda-approves-venetoclax-combination-aml-adults. (Accessed: 27th August 2021)

94. Pollyea, D. A., Amaya, M., Strati, P. & Konopleva, M. Y. Venetoclax for AML: changing the treatment paradigm. *Blood Adv.* **3**, 4326 (2019).

95. FDA grants regular approval to venetoclax in combination for untreated acute myeloid leukemia | FDA. Available at: https://www.fda.gov/drugs/resources-information-approved-drugs/fda-grants-regular-approval-venetoclax-combination-untreated-acute-myeloid-leukemia. (Accessed: 27th August 2021)

96. Abbas, H. A., Alfayez, M., Kadia, T., Ravandi-Kashani, F. & Daver, N. Midostaurin in acute myeloid leukemia: An evidence-based review and patient selection. *Cancer*

Management and Research **11**, 8817–8828 (2019).

97. Levis, M. Midostaurin approved for FLT3-mutated AML. *Blood* **129**, 3403–3406 (2017).

98. Stone, R. M. *et al.* Midostaurin plus Chemotherapy for Acute Myeloid Leukemia with a FLT3 Mutation. *http://dx.doi.org/10.1056/NEJMoa1614359* **377**, 454–464 (2017).

99. Daver, N., Schlenk, R. F., Russell, N. H. & Levis, M. J. Targeting FLT3 mutations in AML: review of current knowledge and evidence. *Leukemia* **33**, 299–312 (2019).

100. Tzogani, K. *et al.* The European Medicines Agency Review of Gilteritinib (Xospata) for the Treatment of Adult Patients with Relapsed or Refractory Acute Myeloid Leukemia with an FLT3 Mutation. (2020). doi:10.1634/theoncologist.2019-0976

101. Levis, M. & Perl, A. E. Gilteritinib: potent targeting of FLT3 mutations in AML. *Blood Adv.* **4**, 1178 (2020).

102. FDA approves gilteritinib for relapsed or refractory acute myeloid leukemia (AML) with a FLT3 mutatation | FDA. Available at: https://www.fda.gov/drugs/fda-approves-gilteritinib-relapsed-or-refractory-acute-myeloid-leukemia-aml-flt3-mutatation. (Accessed: 27th August 2021)

103. Overview | Gilteritinib for treating relapsed or refractory acute myeloid leukaemia | Guidance | NICE.

104. Ravandi, F., Walter, R. B. & Freeman, S. D. Evaluating measurable residual disease in acute myeloid leukemia. *Blood Adv.* **2**, 1356–1366 (2018).

105. Ossenkoppele, G. & Schuurhuis, G. J. MRD in AML: Time for redefinition of CR? *Blood* **121**, 2166–2168 (2013).

106. Bachas, C. *et al.* The role of minor subpopulations within the leukemic blast compartment of AML patients at initial diagnosis in the development of relapse. *Leukemia* **26**, 1313–1320 (2012).

107. Tallman, M. S. *et al.* Acute Myeloid Leukemia, Version 3.2019, NCCN Clinical Practice Guidelines in Oncology. *J. Natl. Compr. Cancer Netw.* **17**, 721–749 (2019).

108. Hantel, A., Stock, W. & Kosuri, S. Molecular Minimal Residual Disease Testing in Acute Myeloid Leukemia: A Review for the Practicing Clinician. *Clin. Lymphoma Myeloma Leuk.* **18**, 636–647 (2018).

109. Kreuzer, K. A. *et al.* Preexistence and evolution of imatinib mesylate-resistant clones in chronic myelogenous leukemia detected by a PNA-based PCR clamping technique. *Ann. Hematol.* **82**, 284–289 (2003).

110. Roche-Lestienne, C. & Preudhomme, C. Mutations in the ABL kinase domain pre-exist the onset of imatinib treatment. in *Seminars in Hematology* **40**, 80–82 (W.B. Saunders, 2003).

111. McDermott, J. E. *et al.* Challenges in biomarker discovery: combining expert insights with statistical analysis of complex omics data. *Expert Opin. Med. Diagn.* **7**, 37–51 (2013).

112. Xi, X. *et al.* RNA biomarkers: Frontier of precision medicine for cancer. *Non-coding RNA* **3**, (2017).

113. Stamey, T. A. *et al.* Prostate-Specific Antigen as a Serum Marker for Adenocarcinoma of the Prostate. *N. Engl. J. Med.* **317**, 909–916 (1987).

114. Martin, R. M. *et al.* Effect of a low-intensity PSA-based screening intervention on prostate cancer mortality: The CAP randomized clinical trial. *JAMA - J. Am. Med. Assoc.* **319**, 883–895 (2018).

115. Burgdorf, S. K., Claesson, M. H., Nielsen, H. J. & Rosenberg, J. Changes in cytokine and biomarker blood levels in patients with colorectal cancer during dendritic cell-based vaccination. *Acta Oncol. (Madr).* **48**, 1157–1164 (2009).

116. Monastero, R. N. & Pentyala, S. Cytokines as Biomarkers and Their Respective Clinical Cutoff Levels. *Int. J. Inflam.* **2017**, (2017).

117. Scott, D. W. *et al.* Determining cell-of-origin subtypes of diffuse large B-cell lymphoma using gene expression in formalin-fixed paraffin-embedded tissue. *Blood* **123**, 1214–1217 (2014).

118. Davies, J. C. *et al.* Efficacy and safety of ivacaftor in patients aged 6 to 11 years with cystic fibrosis with a G551D mutation. *Am. J. Respir. Crit. Care Med.* **187**, 1219–1225 (2013).

119. Kyle, R. A. *et al.* A Long-Term Study of Prognosis in Monoclonal Gammopathy of Undetermined Significance. *N. Engl. J. Med.* **346**, 564–569 (2002).

120. Group, F.-N. B. W. *BEST (Biomarkers, EndpointS, and other Tools) Resource*. (Food

and Drug Administration (US), 2018).

121. Mayeux, R. Biomarkers: Potential Uses and Limitations. *NeuroRx* **1**, 182 (2004).

122. McDermott, J. E. *et al.* Challenges in Biomarker Discovery: Combining Expert Insights with Statistical Analysis of Complex Omics Data. *Expert Opin. Med. Diagn.* **7**, 37–51 (2013).

123. Prada-Arismendy, J., Arroyave, J. C. & Röthlisberger, S. Molecular biomarkers in acute myeloid leukemia. *Blood Rev.* **31**, 63–76 (2017).

124. FLT3 fms related receptor tyrosine kinase 3 [Homo sapiens (human)] - Gene - NCBI. Available at: https://www.ncbi.nlm.nih.gov/gene/2322. (Accessed: 2nd March 2022)

125. Pezaro, C., Woo, H. H. & Davis, I. D. Prostate cancer: measuring PSA. *Intern. Med. J.* **44**, 433–440 (2014).

126. Castellsagué, X. Natural history and epidemiology of HPV infection and cervical cancer. *Gynecol. Oncol.* **110**, S4–S7 (2008).

127. Burd, E. M. Human Papillomavirus and Cervical Cancer. *Clin. Microbiol. Rev.* **16**, 1 (2003).

128. Cervical cancer risk | Cancer Research UK. Available at: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/cervical-cancer/risk-factors#heading-Two. (Accessed: 27th August 2021)

129. Zafeiris, D., Rutella, S. & Ball, G. R. An Artificial Neural Network Integrated Pipeline for Biomarker Discovery Using Alzheimer's Disease as a Case Study. (2018). doi:10.1016/j.csbj.2018.02.001

130. Khan, S. R., Manialawy, Y., Wheeler, M. B. & Cox, B. J. Unbiased data analytic strategies to improve biomarker discovery in precision medicine. *Drug Discovery Today* **24**, 1735–1748 (2019).

131. Barrett, T. *et al.* NCBI GEO: Archive for functional genomics data sets - Update. *Nucleic Acids Res.* **41**, (2013).

132. Athar, A. *et al.* ArrayExpress update-from bulk to single-cell expression data. *Nucleic Acids Res.* **47**, 711–715 (2018).

133. Massagué, J. TGFβ in Cancer. *Cell* **134**, 215–230 (2008).

134. Furler, R. L., Nixon, D. F., Brantner, C. A., Popratiloff, A. & Uittenbogaart, C. H. TGF-β Sustains Tumor Progression through Biochemical and Mechanical Signal Transduction. *Cancers (Basel).* **10**, (2018).

135. Pickup, M., Novitskiy, S. & Moses, H. L. The roles of TGFβ in the tumour microenvironment. *Nat. Rev. Cancer* **13**, 788 (2013).

136. Levy, L. & Hill, C. S. Alterations in components of the TGF-β superfamily signaling pathways in human cancer. *Cytokine Growth Factor Rev.* **17**, 41–58 (2006).

137. Tsai, Y.-L. *et al.* Endoplasmic reticulum stress activates SRC, relocating chaperones to the cell surface where GRP78/CD109 blocks TGF-β signaling. doi:10.1073/pnas.1714866115

138. Yokoyama, M. *et al.* CD109, a negative regulator of TGF-β signaling, is a putative risk marker in diffuse large B-cell lymphoma. *Int. J. Hematol.* **105**, 614–622 (2017).

139. Hasegawa, M. *et al.* CD109 expression in basal-like breast carcinoma. *Pathol. Int.* **58**, 288–294 (2008).

140. Hagiwara, S. *et al.* Up-regulation of CD109 expression is associated with carcinogenesis of the squamous epithelium of the oral cavity. *Cancer Sci.* **99**, 1916–23 (2008).

141. Emori, M. *et al.* High Expression of CD109 Antigen Regulates the Phenotype of Cancer Stem-Like Cells/Cancer-Initiating Cells in the Novel Epithelioid Sarcoma Cell Line ESX and Is Related to Poor Prognosis of Soft Tissue Sarcoma. *PLoS One* **8**, e84187 (2013).

142. Hashimoto, M. *et al.* Expression of CD109 in human cancer. *Oncogene* **23**, 3716–3720 (2004).

143. Zhang, J.-M. *et al.* CD109 attenuates TGF-β1 signaling and enhances EGF signaling in SK-MG-1 human glioblastoma cells. *Biochem. Biophys. Res. Commun.* **459**, 252–258 (2015).

144. Bizet, A. A. *et al.* CD109-mediated degradation of TGF-β receptors and inhibition of TGF-β responses involve regulation of SMAD7 and Smurf2 localization and function. *J. Cell. Biochem.* **113**, 238–246 (2012).

145. Hagiwara, S. *et al.* Processing of CD109 by furin and its role in the regulation of TGF-β signaling. *Oncogene* **29**, 2181–2191 (2010).

146. Bizet, A. A. *et al.* The TGF-β co-receptor, CD109, promotes internalization and degradation of TGF-β receptors. *Biochim. Biophys. Acta - Mol. Cell Res.* **1813**, 742–753 (2011).

147. Li, C. *et al.* Soluble CD109 binds TGF-β and antagonizes TGF-β signalling and responses. *Biochem. J.* **473**, 537–47 (2016).

148. Valk, P. J. M. *et al.* Prognostically Useful Gene-Expression Profiles in Acute Myeloid Leukemia. *N. Engl. J. Med.* **350**, 1617–1628 (2004).

149. Sanders, M. A. & Valk, P. J. M. E-MTAB-3444. *ArrayExpress* (2016). Available at: https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-3444/. (Accessed: 11th July 2018)

150. Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. affy analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307–315 (2004).

151. Huber, A. K. and R. G. and W. arrayQualityMetrics--a bioconductor package for quality assessment of microarray data. *Bioinformatics* **25**, 415–6 (2009).

152. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. *N. Engl. J. Med.* **368**, 2059–2074 (2013).

153. Silva, T. C. TCGAWorkflowData: Data for TCGA Workflow. (2019).

154. Brunner, A. M. & Graubert, T. A. Genomics in childhood acute myeloid leukemia comes of age. *Nature Medicine* **24**, 7–9 (2018).

155. Lancashire, L. J. *et al.* A validated gene expression profile for detecting clinical outcome in breast cancer using artificial neural networks. *Breast Cancer Res. Treat.* **120**, 83–93 (2010).

156. Sakai, T. Two sample T-tests for IR evaluation: Student or welch? *SIGIR 2016 - Proc. 39th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.* 1045–1048 (2016). doi:10.1145/2911451.2914684

157. McKnight, P. E. & Najab, J. Mann-Whitney U Test. *Corsini Encycl. Psychol.* 1–1 (2010). doi:10.1002/9780470479216.CORPSY0524

158. Goel, M. K., Khanna, P. & Kishore, J. Understanding survival analysis: Kaplan-Meier estimate. *Int. J. Ayurveda Res.* **1**, 274 (2010).

159. Nettleman, M. D. Receiver Operator Characteristic (ROC) Curves. *Infect. Control Hosp. Epidemiol.* **9**, 374–377 (1988).

160. Jean-Charles Sanchez, M. M. X. R. N. T. A. H. N. T. F. L. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).

161. Harrell, F. E. Binary Logistic Regression. 219–274 (2015). doi:10.1007/978-3-319-19425-7_10

162. Schemper, M. & Henderson, R. Predictive Accuracy and Explained Variation in Cox Regression. *Biometrics* **56**, 249–255 (2000).

163. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).

164. Wagner, S. *et al.* A parsimonious 3-gene signature predicts clinical outcomes in an acute myeloid leukemia multicohort study. *Blood Adv.* **3**, 1330–1346 (2019).

165. ThermoFisher. Understanding Calculations for siRNA Data | Thermo Fisher Scientific - UK. Available at: https://www.thermofisher.com/uk/en/home/references/ambion-tech-support/rnai-sirna/tech-notes/understanding-calculations-for-sirna-data.html. (Accessed: 23rd April 2020)

166. Definition of biomarker - NCI Dictionary of Cancer Terms - National Cancer Institute. Available at: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/biomarker?redirect=true. (Accessed: 1st July 2021)

167. Henry, N. L. & Hayes, D. F. Cancer biomarkers. *Molecular Oncology* **6**, 140–146 (2012).

168. Pathology Reports - National Cancer Institute. Available at: https://www.cancer.gov/about-cancer/diagnosis-staging/diagnosis/pathology-reports-fact-sheet. (Accessed: 1st July 2021)

169. Meric-Bernstam, F. *et al.* Advances in HER2-targeted therapy: Novel agents and opportunities beyond breast and gastric cancer. *Clin. Cancer Res.* **25**, 2033–2041 (2019).

170. Jayaraman, S., Reid, J. M., Hawse, J. R. & Goetz, M. P. Endoxifen, an Estrogen Receptor Targeted Therapy: From Bench to Bedside. *Endocrinology* **162**, (2021).

171. Siziopikou, K. P., Ariga, R., Proussaloglou, K. E., Gattuso, P. & Cobleigh, M. The Challenging Estrogen Receptor-Negative/ Progesterone Receptor-Negative/HER-2-Negative Patient: A Promising Candidate for Epidermal Growth Factor Receptor-Targeted Therapy? *Breast J.* **12**, 360–362 (2006).

172. Allred, D. C. Issues and updates: evaluating estrogen receptor-α, progesterone receptor, and HER2 in breast cancer. *Mod. Pathol. 2010 232* **23**, S52–S59 (2010).

173. Mohamed, A., Krajewski, K., Cakar, B. & Ma, C. X. Targeted Therapy for Breast Cancer. *Am. J. Pathol.* **183**, 1096–1112 (2013).

174. Giulianelli, S., Molinolo, A. & Lanari, C. Targeting progesterone receptors in breast cancer. *Vitam. Horm.* **93**, 161–184 (2013).

175. Liedtke, C. *et al.* Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer. *J. Clin. Oncol.* **26**, 1275–1281 (2008).

176. Kumar, P. & Aggarwal, R. An overview of triple-negative breast cancer. doi:10.1007/s00404-015-3859-y

177. De Ruijter, T. C., Veeck, J., De Hoon, J. P. J., Van Engeland, M. & Tjan-Heijnen, V. C. Characteristics of triple-negative breast cancer. *J. Cancer Res. Clin. Oncol.* **137**, 183–192 (2011).

178. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **1**, 206–215 (2019).

179. Loyola-Gonzalez, O. Black-box vs. White-Box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access* **7**, 154096–154113 (2019).

180. Wu, M., Li, C. & Zhu, X. FLT3 inhibitors in acute myeloid leukemia. *Journal of Hematology and Oncology* **11**, 1–11 (2018).

181. Chen, Y. *et al.* Tyrosine kinase inhibitors targeting FLT3 in the treatment of acute myeloid leukemia. *Stem Cell Investigation* **4**, (2017).

182. J, G. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, (2013).

183. Cerami, E. *et al.* The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discov.* **2**, 401–404 (2012).

184. Zheng, S., Kim, H. & Verhaak, R. G. W. Silent Mutations Make Some Noise. *Cell* **156**, 1129–1131 (2014).

185. AH, L. & IO, E. The Nottingham prognostic index for invasive carcinoma of the breast. *Pathol. Oncol. Res.* **14**, 113–115 (2008).

186. Fong, Y. *et al.* The Nottingham Prognostic Index: five- and ten-year data for all-cause Survival within a Screened Population. *Ann. R. Coll. Surg. Engl.* **97**, 137 (2015).

187. Sato, T. *et al.* High-level expression of CD109 is frequently detected in lung squamous cell carcinomas. *Pathol. Int.* **57**, 719–724 (2007).

188. Hasegawa, M. *et al.* CD109, a new marker for myoepithelial cells of mammary, salivary, and lacrimal glands and prostate basal cells. *Pathol. Int.* **57**, 245–250 (2007).

189. Misawa, A., Takayama, K.-I., Urano, T. & Inoue, S. Androgen-induced Long Noncoding RNA (lncRNA) SOCS2-AS1 Promotes Cell Growth and Inhibits Apoptosis in Prostate Cancer Cells. *J. Biol. Chem.* **291**, 17861–80 (2016).

190. Jian, F. *et al.* The long-noncoding RNA SOCS2-AS1 suppresses endometrial cancer progression by regulating AURKA degradation. *Cell Death Dis. 2021 124* **12**, 1–13 (2021).

191. Zheng, Z., Li, X., You, H., Zheng, X. & Ruan, X. LncRNA SOCS2-AS1 inhibits progression and metastasis of colorectal cancer through stabilizing SOCS2 and sponging miR-1264. *Aging (Albany. NY).* **12**, 10517–10526 (2020).

192. Frezza, C. Histidine degradation boosts cancer therapy. *Nature* **559**, 484 (2018).

193. Petrova, B. & Kanarek, N. Potential Benefits and Pitfalls of Histidine Supplementation for Cancer Therapy Enhancement. *J. Nutr.* **150**, 2580S-2587S (2020).

194. Wei, L. *et al.* Transcriptome Analysis of Pediatric AML Reveals Non Protein-Coding RNAs Associated with Poor Survival Outcome and Treatment Resistance. *Blood* **134**, 2675–2675 (2019).

195. Leonard, M. K. *et al.* Comprehensive molecular profiling of UV-induced metastatic

melanoma in Nme1/Nme2-deficient mice reveals novel markers of survival in human patients. *Oncogene 2021* 1–14 (2021). doi:10.1038/s41388-021-01998-w

196. Mahamdallie, S. *et al.* Identification of new Wilms tumour predisposition genes: an exome sequencing study. *Lancet. Child Adolesc. Heal.* **3**, 322 (2019).

197. Xiao, H. *et al.* Mutations in epigenetic regulators are involved in acute lymphoblastic leukemia relapse following allogeneic hematopoietic stem cell transplantation. *Oncotarget* **7**, 2696 (2016).

198. Pihur, V., Datta, S. & Datta, S. Finding common genes in multiple cancer types through meta–analysis of microarray experiments: A rank aggregation approach. *Genomics* **92**, 400–403 (2008).

199. Nguyen, N. K. *et al.* Abstract 2535: Leukemic Cell Proteomic Profiling in Pediatric AML. *Cancer Res.* **81**, 2535–2535 (2021).

200. Kwak, H. J. *et al.* Expression of ATP/GTP Binding Protein 1 Has Prognostic Value for the Clinical Outcomes in Non-Small Cell Lung Carcinoma. *J. Pers. Med.* **10**, 1–22 (2020).

201. Meng, Y. *et al.* LAPTM4B: an oncogene in various solid tumors and its functions. *Oncogene 2016 3550* **35**, 6359–6365 (2016).

202. Li, Y., Iglehart, J. D., Richardson, A. L. & Wang, Z. C. The amplified cancer gene LAPTM4B promotes tumor growth and tolerance to stress through the induction of autophagy. *Autophagy* **8**, 273 (2012).

203. Li, L. *et al.* LAPTM4B: A novel cancer-associated gene motivates multidrug resistance through efflux and activating PI3K/AKT signaling. *Oncogene 2010 2943* **29**, 5785–5795 (2010).

204. L, R. *et al.* Myeloperoxidase-positive cell infiltration in colorectal carcinogenesis as indicator of colorectal cancer risk. *Cancer Epidemiol. Biomarkers Prev.* **17**, 2291–2297 (2008).

205. Kiyohara, C., Yoshimasu, K., Takayama, K. & Nakanishi, Y. NQO1, MPO, and the risk of lung cancer: A HuGE review. *Genet. Med. 2005 77* **7**, 463–478 (2005).

206. Ambrosone, C. B. *et al.* Polymorphisms in Genes Related to Oxidative Stress (MPO, MnSOD, CAT) and Survival After Treatment for Breast Cancer. *Cancer Res.* **65**,

(2005).

207. He, C., Tamimi, R. M., Hankinson, S. E., Hunter, D. J. & Han, J. A prospective study of genetic polymorphism in MPO, antioxidant status, and breast cancer risk. *Breast Cancer Res. Treat. 2008 1133* **113**, 585–594 (2008).

208. Shabani, S., Khayer, N., Motalebzadeh, J., zadeh, T. M. & Mahjoubi, F. Characterization of pathways involved in colorectal cancer using real-time RT-PCR gene expression data. *Gastroenterol. Hepatol. From Bed to Bench* **14**, 123 (2021).

209. Huang, C. *et al.* MicroRNA-17 promotes cell proliferation and migration in human colorectal cancer by downregulating SIK1. *Cancer Manag. Res.* **11**, 3521 (2019).

210. Dankert, J. F., Pagan, J. K., Starostina, N. G., Kipreos, E. T. & Pagano, M. FEM1 proteins are ancient regulators of SLBP degradation. *http://dx.doi.org/10.1080/15384101.2017.1284715* **16**, 556–564 (2017).

211. Yan, X. *et al.* Molecular basis for ubiquitin ligase CRL2FEM1C-mediated recognition of C-degron. *Nat. Chem. Biol. 2021 173* **17**, 263–271 (2021).

212. *Sample & Assay Technologies Critical Factors for Successful Real-Time PCR 2 www.qiagen.com Real-Time PCR Brochure 07/2010 Contents*.

213. Castillo, D. *et al.* Leukemia multiclass assessment and classification from Microarray and RNA-seq technologies integration at gene expression level. *PLoS One* **14**, e0212127 (2019).

214. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).

215. Zhang, J. *et al.* CD109 expression in squamous cell carcinoma of the uterine cervix. *Pathol. Int.* **55**, 165–169 (2005).

216. Mancuso, P. *et al.* A Subpopulation of Circulating Endothelial Cells Express CD109 and is Enriched in the Blood of Cancer Patients. *PLoS One* **9**, e114713 (2014).

217. Metzeler, K. H. *et al.* An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia. *Blood* **112**, 4193–201 (2008).

218. Heuser, M. *et al.* Gene-expression profiles and their association with drug resistance in adult acute myeloid leukemia. *Haematologica* **90**, 1484–92 (2005).

219. Finnson, K. W. *et al.* Identification of CD109 as part of the TGF-β receptor system in human keratinocytes. *FASEB J.* **20**, 1525–1527 (2006).

220. Tabe, Y. *et al.* TGF-β-Neutralizing Antibody 1D11 Enhances Cytarabine-Induced Apoptosis in AML Cells in the Bone Marrow Microenvironment. *PLoS One* **8**, e62785 (2013).

221. Hamada, A., Kawaguchi, T. & Nakano, M. Clinical Pharmacokinetics of Cytarabine Formulations. *Clin. Pharmacokinet. 2002 4110* **41**, 705–718 (2012).

222. Chemotherapy for acute myeloid leukaemia (AML) | Acute myeloid leukaemia | Cancer Research UK. Available at: https://www.cancerresearchuk.org/about-cancer/acute-myeloid-leukaemia-aml/treating-aml/chemotherapy/chemotherapy-for-aml. (Accessed: 28th February 2022)

223. Murphy, T. & Yee, K. W. L. Cytarabine and daunorubicin for the treatment of acute myeloid leukemia. *Expert Opin. Pharmacother.* **18**, 1765–1780 (2017).

224. Dombret, H. & Gardin, C. An update of current treatments for adult acute myeloid leukemia. *Blood* **127**, 53 (2016).

225. Watabe, T. & Miyazono, K. Roles of TGF-β family signaling in stem cell renewal and differentiation. *Cell Research* **19**, 103–115 (2009).

226. Iordanskaia, T. & Nawshad, A. Mechanisms of transforming growth factor β induced cell cycle arrest in palate development. *J. Cell. Physiol.* **226**, 1415–1424 (2011).

227. Barrett, C. S. X., Millena, A. C. & Khan, S. A. TGF-β Effects on Prostate Cancer Cell Migration and Invasion Require FosB. *Prostate* **77**, 72–81 (2017).

228. XU, Z., SHEN, M. X., MA, D. Z., WANG, L. Y. & ZHA, X. L. TGF-β1 -promoted epithelial-to-mesenchymal transformation and cell adhesion contribute to TGF-β1 -enhanced cell migration in SMMC-7721 cells. *Cell Res. 2003 135* **13**, 343–350 (2003).

229. KD, B., KK, S., CD, B. & FR, C. The housekeeping gene YWHAZ remains stable in a model of developmentally primed non-alcoholic fatty liver disease. *Liver Int.* **32**, 1315–1321 (2012).

230. Gubern, C. *et al.* Validation of housekeeping genes for quantitative real-time PCR in in-vivo and in-vitro models of cerebral ischaemia. *BMC Mol. Biol. 2009 101* **10**, 1–10 (2009).

231. Lemma, S., Avnet, S., Salerno, M., Chano, T. & Baldini, N. Identification and Validation of Housekeeping Genes for Gene Expression Analysis of Cancer Stem Cells. *PLoS One* **11**, (2016).

232. Kuhn, D. M., Balkis, M., Chandra, J., Mukherjee, P. K. & Ghannoum, M. A. Uses and Limitations of the XTT Assay in Studies of Candida Growth and Metabolism. *J. Clin. Microbiol.* **41**, 506 (2003).

233. Funk, D., Schrenk, H.-H. & Frei, D. E. Serum albumin leads to false-positive results in the XTT and the MTT assay. *https://doi.org/10.2144/000112528* **43**, 178–186 (2018).

234. Lee, K. *et al.* Elevation of CD109 promotes metastasis and drug resistance in lung cancer via activation of EGFR-AKT-mTOR signaling. *Cancer Sci.* **111**, 1652 (2020).

235. German Collection of Microorganisms and Cell Cultures GmbH: Details. Available at: https://www.dsmz.de/collection/catalogue/details/culture/ACC-220. (Accessed: 21st September 2021)

236. German Collection of Microorganisms and Cell Cultures GmbH: Details. Available at: https://www.dsmz.de/collection/catalogue/details/culture/ACC-16. (Accessed: 21st September 2021)

237. German Collection of Microorganisms and Cell Cultures GmbH: Details. Available at: https://www.dsmz.de/collection/catalogue/details/culture/ACC-554. (Accessed: 21st September 2021)

238. Papaemmanuil, E. *et al.* Genomic Classification and Prognosis in Acute Myeloid Leukemia. *N. Engl. J. Med.* **374**, 2209–2221 (2016).

239. Thol, F. & Ganser, A. Treatment of Relapsed Acute Myeloid Leukemia. *Curr. Treat. Options Oncol.* **21**, 66 (2020).

240. Gaksch, L. *et al.* Residual disease detection using targeted parallel sequencing predicts relapse in cytogenetically normal acute myeloid leukemia. *Am. J. Hematol.* (2017). doi:10.1002/ajh.24922

241. Shlush, L. I. *et al.* Tracing the origins of relapse in acute myeloid leukaemia to stem cells. *Nat. 2017 5477661* **547**, 104–108 (2017).

242. Jonas, B. A. On the origin of relapse in AML. *Sci. Transl. Med.* **9**, (2017).

243. Shlush, L. I. & Mitchell, A. AML evolution from preleukemia to leukemia and relapse. *Best Practice and Research: Clinical Haematology* **28**, (2015).

244. Irala-EstØvez, J. de & Navajas, D. M.-C. C. D.-M. J. M.-C. A. S. del C. R. F.-C. Comparison of different methodological approaches to identify risk factors of nosocomial infection in intensive care units. *Intensive Care Med* 1254±1262 (2001). doi:10.1007/s001340101007

245. Staley, J. R. *et al.* A comparison of Cox and logistic regression for use in genome-wide association studies of cohort and case-cohort design. *Eur. J. Hum. Genet.* **25**, 854–862 (2017).

246. Murray, L. J. *et al.* CD109 is expressed on a subpopulation of CD34+ cells enriched in hematopoietic stem and progenitor cells. *Exp. Hematol.* **27**, 1282–1294 (1999).

247. Migdalska, A. *et al.* Growth inhibitory effects of transforming growth factor-β1 in vivo. *Growth Factors* **4**, 239–245 (1991).

248. Mukherjee, P., Winter, S. L. & Alexandrow, M. G. Cell Cycle Arrest by Transforming Growth Factor β1 near G1/S Is Mediated by Acute Abrogation of Prereplication Complex Activation Involving an Rb-MCM Interaction. *Mol. Cell. Biol.* **30**, 845–856 (2010).

249. Solomon, L. A. *et al.* Coordination of Myeloid Differentiation with Reduced Cell Cycle Progression by PU.1 Induction of MicroRNAs Targeting Cell Cycle Regulators and Lipid Anabolism. *Mol. Cell. Biol.* **37**, (2017).

250. RUSCETTI, F. W. *et al.* Role of Transforming Growth Factor-β1 in Regulation of Hematopoiesis. *Ann. N. Y. Acad. Sci.* **628**, 31–43 (1991).

251. Massagué, J. & Xi, Q. TGF-β control of stem cell differentiation genes. *FEBS Letters* **586**, 1953–1958 (2012).

252. TGF-beta Signaling Regulates the Expression of Multiple Target Genes with Tumor-Suppressing & Tumor-Promoting Effects: R&D Systems. Available at: https://www.rndsystems.com/resources/posters/tgf-beta-signaling-regulates-expression-multiple-target-genes-tumor-suppressing-tumor. (Accessed: 4th August 2020)

253. Huang, F. *et al.* HER2/EGFR–AKT signaling switches TGFb from inhibiting cell proliferation to promoting cell migration in breast cancer. *Cancer Res.* **78**, 6073–

6085 (2018).

254. Kannan, K., Amariglio, N., Rechavi, G. & Givol, D. Profile of gene expression regulated by induced p53: Connection to the TGF-β family. *FEBS Lett.* **470**, 77–82 (2000).

255. Ranganathan, P. *et al.* Expression profiling of genes regulated by TGF-beta: Differential regulation in normal and tumour cells. *BMC Genomics* **8**, 98 (2007).

256. Peng, D., Fu, L. & Sun, G. Expression analysis of the TGF-β/SMAD target genes in adenocarcinoma of esophagogastric junction. *Open Med.* **11**, 83–86 (2016).

257. Du, Y. *et al.* Mapping gene expression patterns during myeloid differentiation using the EML hematopoietic progenitor cell line. *Exp. Hematol.* **30**, 649–658 (2002).

258. Zabala, M. *et al.* LEFTY1 Is a Dual-SMAD Inhibitor that Promotes Mammary Progenitor Growth and Tumorigenesis. *Cell Stem Cell* **27**, 284-299.e8 (2020).

259. Zhang, L. *et al.* Lefty-1 inhibits renal epithelial–mesenchymal transition by antagonizing the TGF-β/Smad signaling pathway. *J. Mol. Histol. 2020 511* **51**, 77–87 (2020).

260. Gambichler, T. *et al.* Expression of Lefty predicts Merkel cell carcinoma-specific death. *J. Eur. Acad. Dermatology Venereol.* **34**, 2016–2020 (2020).

SUPPLEMENTARY TABLES

5.4.5. Supplementary table 1

Gene name	B	S.E.	Wald	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
CD109	0.379	0.080	22.349	0.000	1.461	1.248	1.709
ADGRG1	0.384	0.084	20.833	0.000	1.468	1.245	1.731
NYNRIN	0.611	0.142	18.489	0.000	1.842	1.394	2.433
SIRPB2	-0.359	0.084	18.351	0.000	0.698	0.593	0.823
ACVR1B	-0.590	0.142	17.208	0.000	0.554	0.419	0.733
FUT4	-0.438	0.108	16.518	0.000	0.645	0.522	0.797
ZC3H12C	0.365	0.091	15.931	0.000	1.440	1.204	1.723
SOCS2-AS1	0.507	0.128	15.637	0.000	1.660	1.291	2.134
HDAC4	-0.397	0.102	15.112	0.000	0.672	0.550	0.821
DNMT3B	0.363	0.094	14.925	0.000	1.437	1.196	1.727
HAL	-0.482	0.125	14.827	0.000	0.618	0.483	0.789
KIAA0125	0.397	0.106	14.051	0.000	1.488	1.209	1.831
RAB27A	-0.358	0.097	13.681	0.000	0.699	0.579	0.845
ADGRG5	0.391	0.106	13.615	0.000	1.479	1.201	1.820
PIWIL4	-0.375	0.103	13.211	0.000	0.687	0.561	0.841
SOCS2	0.221	0.063	12.370	0.000	1.247	1.103	1.411
SVIP	0.436	0.134	10.516	0.001	1.546	1.188	2.012
ARHGAP22	0.318	0.098	10.483	0.001	1.374	1.134	1.665
MAP7	0.245	0.076	10.392	0.001	1.278	1.101	1.484
MYO1B	0.366	0.114	10.242	0.001	1.442	1.152	1.805
PXK	-0.393	0.125	9.802	0.002	0.675	0.528	0.863
SPNS3	0.318	0.105	9.234	0.002	1.375	1.120	1.688
TRIM16	0.329	0.110	8.891	0.003	1.389	1.119	1.725
MID1IP1	-0.364	0.127	8.169	0.004	0.695	0.541	0.892
SPNS2	0.323	0.113	8.138	0.004	1.381	1.106	1.723
FNDC3B	-0.337	0.121	7.810	0.005	0.714	0.564	0.904

MAP1A	0.320	0.117	7.521	0.006	1.377	1.096	1.731
ABCG1	0.412	0.154	7.189	0.007	1.510	1.117	2.041
ST3GAL6	-0.287	0.108	7.103	0.008	0.750	0.608	0.927
MSI2	0.180	0.069	6.821	0.009	1.197	1.046	1.370
CXXC5	0.231	0.089	6.784	0.009	1.260	1.059	1.498
SV2A	0.291	0.114	6.552	0.010	1.337	1.070	1.670
TLR2	-0.234	0.093	6.400	0.011	0.791	0.660	0.949
CD53	-0.349	0.139	6.326	0.012	0.706	0.538	0.926
LCT	0.354	0.143	6.100	0.014	1.425	1.076	1.887
CCDC102A	0.322	0.136	5.581	0.018	1.380	1.056	1.803
DEAF1	0.341	0.146	5.493	0.019	1.407	1.057	1.871
HOMER3	-0.188	0.081	5.309	0.021	0.829	0.707	0.972
ADPGK	-0.434	0.191	5.141	0.023	0.648	0.445	0.943
DAPK1	0.205	0.091	5.012	0.025	1.227	1.026	1.468
SLC31A2	-0.226	0.103	4.818	0.028	0.797	0.652	0.976
GNA12	0.296	0.139	4.518	0.034	1.344	1.023	1.766
MAST4	0.277	0.131	4.501	0.034	1.319	1.021	1.703
FKBP15	-0.282	0.134	4.457	0.035	0.754	0.580	0.980
USP20	0.287	0.137	4.363	0.037	1.333	1.018	1.745
GPSM1	0.213	0.103	4.285	0.038	1.237	1.011	1.514
TMEM167A	-0.295	0.144	4.207	0.040	0.744	0.561	0.987
LILRA2	-0.172	0.084	4.167	0.041	0.842	0.713	0.993
EMILIN2	-0.191	0.095	4.044	0.044	0.826	0.686	0.995
MFSD1	-0.278	0.140	3.935	0.047	0.757	0.576	0.997
PHC2	-0.352	0.178	3.928	0.047	0.703	0.496	0.996
FRMD4B	0.168	0.085	3.889	0.049	1.183	1.001	1.399
P4HB	-0.301	0.156	3.715	0.054	0.740	0.544	1.005
AGTRAP	-0.131	0.106	1.519	0.218	0.878	0.713	1.080
ARID3B	0.318	0.169	3.539	0.060	1.374	0.987	1.913
ARRDC1	-0.168	0.150	1.262	0.261	0.845	0.630	1.134
ATG7	-0.242	0.136	3.178	0.075	0.785	0.602	1.024
CPD	-0.183	0.112	2.673	0.102	0.833	0.669	1.037
CRADD	0.106	0.174	0.373	0.541	1.112	0.791	1.563

DNAJC5	-0.191	0.143	1.779	0.182	0.826	0.624	1.094
DPP4	0.283	0.148	3.676	0.055	1.328	0.994	1.774
ERMP1	0.234	0.124	3.544	0.060	1.264	0.990	1.613
FHL1	0.147	0.078	3.579	0.059	1.159	0.995	1.350
GALNS	-0.277	0.162	2.940	0.086	0.758	0.552	1.040
GLA	-0.295	0.157	3.532	0.060	0.745	0.547	1.013
HSF5	0.097	0.138	0.494	0.482	1.101	0.841	1.442
LRBA	0.171	0.120	2.033	0.154	1.187	0.938	1.502
MCU	0.003	0.165	0.000	0.987	1.003	0.726	1.385
MFSD10	-0.182	0.134	1.853	0.173	0.834	0.641	1.083
MLX	-0.240	0.156	2.356	0.125	0.787	0.579	1.069
NFATC2	0.272	0.141	3.713	0.054	1.312	0.995	1.730
NIPSNAP1	0.170	0.166	1.051	0.305	1.185	0.857	1.640
PDK2	0.171	0.160	1.137	0.286	1.186	0.867	1.623
PEX5	0.163	0.158	1.063	0.303	1.177	0.864	1.603
PHKA2	-0.223	0.169	1.739	0.187	0.800	0.574	1.115
PRDM15	0.114	0.170	0.448	0.504	1.120	0.803	1.562
PRKCH	0.053	0.091	0.342	0.559	1.055	0.882	1.261
PRKD2	0.197	0.160	1.516	0.218	1.217	0.890	1.664
RNF185	-0.096	0.168	0.331	0.565	0.908	0.654	1.261
SEMA4F	0.086	0.194	0.198	0.657	1.090	0.745	1.595
SPPL2B	0.218	0.144	2.305	0.129	1.244	0.938	1.649
TCF4	0.138	0.072	3.663	0.056	1.148	0.997	1.323
TLR5	-0.048	0.095	0.251	0.616	0.953	0.791	1.149
ZNF438	-0.273	0.162	2.826	0.093	0.761	0.554	1.046

SUPPLEMENTARY TABLE 5.1. UNIVARIATE BLR RESULTS OF THE ANN WORKFLOW. AFTER THE ANN WAS PERFORMED THE 84 UNIQUE GENES, IDENTIFIED AS RELATIVE TO RELAPSE, WERE VALIDATED USING BLR, FROM WHICH 52 WERE SIGNIFICANTLY ASSOCIATED WITH RELAPSE AS SHOWN IN GREEN.

5.4.6. Supplementary table 2

Gene name	B	S.E.	Wald	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
CD109	0.255	0.101	6.370	0.012	1.291	1.059	1.573
SLC31A2	0.380	0.183	4.291	0.038	1.462	1.021	2.095
ADPGK	-0.894	0.342	6.821	0.009	0.409	0.209	0.800
ACVR1B	-0.763	0.238	10.277	0.001	0.466	0.292	0.743
DNMT3B	0.587	0.173	11.511	0.001	1.799	1.281	2.526
MSI2	-0.255	0.127	4.019	0.045	0.775	0.604	0.994
SOCS2-AS1	0.548	0.177	9.571	0.002	1.729	1.222	2.446
CEBPA double mutations	-0.802	0.497	2.603	0.107	0.449	0.169	1.188

SUPPLEMENTARY TABLE 5.2. MULTIVARIATE BLR OF THE ANN WORKFLOW INCLUDING CEBPA. THE GENES WHICH SIGNIFICANTLY CONTRIBUTED TO THE SCORE BASED ON THE MULTIVARIATE BLR IN THE ANN WORKFLOW WERE ANALYSED AGAIN USING A BLR BUT INCLUDED CLINICAL ANOTATIONS.

5.4.7. Supplementary table 3

Gene name	B	S.E.	Wald	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
CD109	0.302	0.098	9.408	0.002	1.352	1.115	1.640
SLC31A2	0.397	0.182	4.752	0.029	1.487	1.041	2.124
ADPGK	-0.949	0.340	7.787	0.005	0.387	0.199	0.754
ACVR1B	-0.762	0.237	10.337	0.001	0.467	0.293	0.743
DNMT3B	0.544	0.170	10.242	0.001	1.723	1.235	2.405
MSI2	-0.277	0.128	4.704	0.030	0.758	0.590	0.974
SOCS2-AS1	0.553	0.176	9.882	0.002	1.738	1.231	2.454
flt3-itd	0.252	0.277	0.826	0.363	1.287	0.747	2.217

SUPPLEMENTARY TABLE 5.3. MULTIVARIATE BLR OF THE ANN WORKFLOW INCLUDING FLT3-ITD. THE GENES WHICH SIGNIFICANTLY CONTRIBUTED TO THE SCORE BASED ON THE MULTIVARIATE BLR IN THE ANN WORKFLOW WERE ANALYSED AGAIN USING A BLR BUT INCLUDED CLINICAL ANOTATIONS.

5.4.8. Supplementary table 4

Gene name	B	S.E.	Wald	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
CD109	0.275	0.099	7.742	0.005	1.317	1.085	1.598
SLC31A2	0.369	0.184	4.021	0.045	1.446	1.008	2.073
ADPGK	-0.996	0.344	8.399	0.004	0.369	0.188	0.724
ACVR1B	-0.680	0.242	7.893	0.005	0.506	0.315	0.814
DNMT3B	0.547	0.171	10.221	0.001	1.727	1.236	2.415
MSI2	-0.282	0.129	4.796	0.029	0.754	0.586	0.971
SOCS2-AS1	0.605	0.180	11.332	0.001	1.831	1.287	2.603
t(8;21)(q22;q22)	-0.682	0.402	2.882	0.090	0.505	0.230	1.111

SUPPLEMENTARY TABLE 5.4. MULTIVARIATE BLR OF THE ANN WORKFLOW INCLUDING T(8;21)(Q22;Q22).

THE GENES WHICH SIGNIFICANTLY CONTRIBUTED TO THE SCORE BASED ON THE MULTIVARIATE BLR IN THE ANN WORKFLOW WERE ANALYSED AGAIN USING A BLR BUT INCLUDED CLINICAL ANOTATIONS.

5.4.9. Supplementary table 5

Gene name	B	S.E.	Wald	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
CD109	0.273	0.100	7.409	0.006	1.313	1.079	1.598
SLC31A2	0.374	0.183	4.169	0.041	1.454	1.015	2.083
ADPGK	-0.853	0.343	6.168	0.013	0.426	0.217	0.835
ACVR1B	-0.735	0.238	9.509	0.002	0.479	0.301	0.765
DNMT3B	0.496	0.173	8.230	0.004	1.643	1.170	2.306
MSI2	-0.253	0.128	3.917	0.048	0.776	0.604	0.998
SOCS2-AS1	0.548	0.176	9.677	0.002	1.730	1.225	2.443
Cytogenetic Risk	0.332	0.206	2.603	0.107	1.394	0.931	2.088

SUPPLEMENTARY TABLE 5.5. MULTIVARIATE BLR OF THE ANN WORKFLOW INCLUDING CYTOGENETIC RISK. THE GENES WHICH SIGNIFICANTLY CONTRIBUTED TO THE SCORE BASED ON THE MULTIVARIATE BLR IN THE ANN WORKFLOW WERE ANALYSED AGAIN USING A BLR BUT INCLUDED CLINICAL ANOTATIONS.

5.4.10. Supplementary Table 6

Gene	FDR	Original
CD109	0.000	0.000
ADGRG1	0.001	0.000
NYNRIN	0.001	0.000
SIRPB2	0.001	0.000
FUT4	0.002	0.000
ACVR1B	0.003	0.000
SOCS2-AS1	0.003	0.000
ZC3H12C	0.004	0.000
RAB27A	0.007	0.000
HDAC4	0.007	0.000
DNMT3B	0.007	0.000
HAL	0.008	0.000
PIWIL4	0.009	0.000
NGFRAP1	0.009	0.000
TPM2	0.009	0.000
TES	0.010	0.000
KIAA0125	0.010	0.000
ADGRG5	0.010	0.000
EVI2B	0.010	0.000
SOCS2	0.011	0.000

SUPPLEMENTARY TABLE 5.6. CORRECTED P VALUE FOR THE T-TEST METHOD. ALL THE GENES IN THE HOVON DATASET WERE ANALYSED USING A T-TEST, THE P-VALUE WAS CORRECTED FOR THE FALSE DISCOVERY RATE. THE TOP 20 SIGNIFICANT GENES ARE SHOWN HERE

5.4.11. Supplementary table 7

Gene Name	B	S.E.	Wald	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
CD109	0.379	0.080	22.349	0.000	1.461	1.248	1.709
ADGRG1	0.384	0.084	20.833	0.000	1.468	1.245	1.731
NYNRIN	0.611	0.142	18.489	0.000	1.842	1.394	2.433
FUT4	-0.438	0.108	16.518	0.000	0.645	0.522	0.797
ACVR1B	-0.590	0.142	17.208	0.000	0.554	0.419	0.733
SOCS2-AS1	0.507	0.128	15.637	0.000	1.660	1.291	2.134
ZC3H12C	0.365	0.091	15.931	0.000	1.440	1.204	1.723
RAB27A	-0.358	0.097	13.681	0.000	0.699	0.579	0.845
HDAC4	-0.397	0.102	15.112	0.000	0.672	0.550	0.821
DNMT3B	0.363	0.094	14.925	0.000	1.437	1.196	1.727
HAL	-0.482	0.125	14.827	0.000	0.618	0.483	0.789
PIWIL4	-0.375	0.103	13.211	0.000	0.687	0.561	0.841
NGFRAP1	0.261	0.069	14.275	0.000	1.298	1.134	1.485
TPM2	0.415	0.110	14.128	0.000	1.514	1.219	1.879
TES	-0.397	0.116	11.779	0.001	0.672	0.536	0.843
KIAA0125	0.397	0.106	14.051	0.000	1.488	1.209	1.831
ADGRG5	0.391	0.106	13.615	0.000	1.479	1.201	1.820
EVI2B	-0.391	0.111	12.391	0.000	0.676	0.544	0.841
SOCS2	0.221	0.063	12.370	0.000	1.247	1.103	1.411

SUPPLEMENTARY TABLE 5.7. UNIVARIATE BLR FOR THE T-TEST METHOD. ALL THE GENES IN THE HOVON DATASET WERE ANALYSED USING A T-TEST, THE TOP 20 SIGNIFICANT GENES WERE ANALYSED USING A UNIVARIATE BLR.

5.4.12. Supplementary table 8

Gene name	B	S.E.	Wald	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
CD109	0.379	0.080	22.349	0.000	1.461	1.248	1.709
ADGRG1	0.384	0.084	20.833	0.000	1.468	1.245	1.731
SIRPB2	-0.359	0.084	18.351	0.000	0.698	0.593	0.823
FUT4	-0.438	0.108	16.518	0.000	0.645	0.522	0.797
ACVR1B	-0.590	0.142	17.208	0.000	0.554	0.419	0.733
NYNRIN	0.611	0.142	18.489	0.000	1.842	1.394	2.433
HAL	-0.482	0.125	14.827	0.000	0.618	0.483	0.789
ZC3H12C	0.365	0.091	15.931	0.000	1.440	1.204	1.723
HDAC4	-0.397	0.102	15.112	0.000	0.672	0.550	0.821
RAB27A	-0.358	0.097	13.681	0.000	0.699	0.579	0.845
DNMT3B	0.363	0.094	14.925	0.000	1.437	1.196	1.727
PIWIL4	-0.375	0.103	13.211	0.000	0.687	0.561	0.841
SOCS2-AS1	0.507	0.128	15.637	0.000	1.660	1.291	2.134
KIAA0125	0.397	0.106	14.051	0.000	1.488	1.209	1.831
NGFRAP1	0.261	0.069	14.275	0.000	1.298	1.134	1.485
ADGRG5	0.391	0.106	13.615	0.000	1.479	1.201	1.820
ALDH1A1	0.278	0.083	11.196	0.001	1.320	1.122	1.554
ST6GALNAC4	0.393	0.116	11.483	0.001	1.481	1.180	1.859
CD58	-0.384	0.108	12.747	0.000	0.681	0.552	0.841
TPM2	0.415	0.110	14.128	0.000	1.514	1.219	1.879

SUPPLEMENTARY TABLE 5.8. UNIVARIATE BLR FOR THE ANN SECOND WORKFLOW. A T-TEST WAS USED TO IDENTIFY THE TOP 1000 MOST SIGNIFICANT GENES WHICH WERE THEN TAKEN TO THE ANN ALGORITHM, WHICH WAS USED TO IDENTIFY THE TOP 20 GENES WITH THE LOWEST AVERAGE TEST ERROR. THE TOP 20 GENES WERE VALIDATED USING A BLR AS SHOWN HERE.

5.4.13. Supplementary table 9

	Gene	AUC	Threshold	Specificity	Sensitivity	Accuracy
1	CD109	0.64327284	6.56919522	0.69536424	0.54978355	0.60732984
2	ACVR1B	0.63435681	9.20827611	0.53642384	0.68831169	0.62827225
3	HAL	0.63214931	8.79603168	0.53642384	0.69264069	0.63089005
4	ADGRG1	0.63183395	9.68445381	0.69536424	0.51515152	0.58638743
5	SIRPB2	0.63008515	8.82584584	0.61589404	0.61038961	0.61256545
6	NYNRIN	0.62711792	7.6329554	0.52980132	0.66233766	0.60994764
7	FUT4	0.62610017	10.314327	0.65562914	0.56709957	0.60209424
8	HDAC4	0.6256128	7.63938101	0.63576159	0.61038961	0.62041885
9	ZC3H12C	0.62501075	5.7619593	0.60264901	0.61038961	0.60732984
10	RAB27A	0.62288925	10.6348025	0.51655629	0.70995671	0.63350785
11	SOCS2-AS1	0.61592271	6.17108324	0.68874172	0.51515152	0.58376963
12	NGFRAP1	0.61503397	10.0423651	0.64900662	0.55411255	0.59162304
13	CD58	0.61483329	9.14070847	0.54304636	0.66666667	0.61780105
14	TMEM243	0.61431725	8.37580431	0.56953642	0.63203463	0.60732984
15	KIAA0125	0.61385855	8.88778748	0.52980132	0.67099567	0.61518325
16	ADGRG5	0.61377254	8.13890611	0.58940397	0.61038961	0.60209424
17	DNMT3B	0.61331384	8.41619179	0.62251656	0.58874459	0.60209424
18	MPO	0.61285514	12.5762519	0.61589404	0.57575758	0.59162304
19	TMSB10	0.61182306	12.921338	0.59602649	0.5974026	0.59685864
20	PIWIL4	0.61050429	8.80586137	0.63576159	0.53679654	0.57591623

SUPPLEMENTARY TABLE 5.9. TOP 20 AUC FOR THE BLR WORKFLOW. ALL THE GENES IN THE HOVON DATASET WERE ANALYSED USING A ROC CURVE, THE TOP 20 GENES WITH THE HIGHEST AUC ARE DISPLAYED HERE.

5.4.14. Supplementary table 10

Gene name	B	S.E.	Wald	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
CD109	0.379	0.080	22.349	0.000	1.461	1.248	1.709
ACVR1B	-0.590	0.142	17.208	0.000	0.554	0.419	0.733
HAL	-0.482	0.125	14.827	0.000	0.618	0.483	0.789
ADGRG1	0.384	0.084	20.833	0.000	1.468	1.245	1.731
SIRPB2	-0.359	0.084	18.351	0.000	0.698	0.593	0.823
NYNRIN	0.611	0.142	18.489	0.000	1.842	1.394	2.433
FUT4	-0.438	0.108	16.518	0.000	0.645	0.522	0.797
HDAC4	-0.397	0.102	15.112	0.000	0.672	0.550	0.821
ZC3H12C	0.365	0.091	15.931	0.000	1.440	1.204	1.723
RAB27A	-0.358	0.097	13.681	0.000	0.699	0.579	0.845
SOCS2-AS1	0.507	0.128	15.637	0.000	1.660	1.291	2.134
NGFRAP1	0.261	0.069	14.275	0.000	1.298	1.134	1.485
CD58	-0.384	0.108	12.747	0.000	0.681	0.552	0.841
TMEM243	-0.420	0.120	12.252	0.000	0.657	0.519	0.831
KIAA0125	0.397	0.106	14.051	0.000	1.488	1.209	1.831
ADGRG5	0.391	0.106	13.615	0.000	1.479	1.201	1.820
DNMT3B	0.363	0.094	14.925	0.000	1.437	1.196	1.727
MPO	-0.177	0.051	11.842	0.001	0.838	0.757	0.927
TMSB10	-0.541	0.160	11.393	0.001	0.582	0.425	0.797
PIWIL4	-0.375	0.103	13.211	0.000	0.687	0.561	0.841

SUPPLEMENTARY TABLE 5.10. UNIVARIATE BLR FOR THE ROC WORKFLOW. ALL THE GENES IN THE HOVON DATASET WERE ANALYSED USING A ROC CURVE, THE TOP 20 GENES WITH THE HIGHEST AUC WERE ANALYSED USING A UNIVARIATE BLR.

5.4.15. Supplementary table 11

Gene name	B	S.E.	Wald	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
CD109	0.379	0.080	22.349	0.000	1.461	1.248	1.709
ACVR1B	-0.590	0.142	17.208	0.000	0.554	0.419	0.733
HAL	-0.482	0.125	14.827	0.000	0.618	0.483	0.789
ADGRG1	0.384	0.084	20.833	0.000	1.468	1.245	1.731
SIRPB2	-0.359	0.084	18.351	0.000	0.698	0.593	0.823
NYNRIN	0.611	0.142	18.489	0.000	1.842	1.394	2.433
FUT4	-0.438	0.108	16.518	0.000	0.645	0.522	0.797
HDAC4	-0.397	0.102	15.112	0.000	0.672	0.550	0.821
ZC3H12C	0.365	0.091	15.931	0.000	1.440	1.204	1.723
RAB27A	-0.358	0.097	13.681	0.000	0.699	0.579	0.845
SOCS2-AS1	0.507	0.128	15.637	0.000	1.660	1.291	2.134
NGFRAP1	0.261	0.069	14.275	0.000	1.298	1.134	1.485
CD58	-0.384	0.108	12.747	0.000	0.681	0.552	0.841
TMEM243	-0.420	0.120	12.252	0.000	0.657	0.519	0.831
KIAA0125	0.397	0.106	14.051	0.000	1.488	1.209	1.831
ADGRG5	0.391	0.106	13.615	0.000	1.479	1.201	1.820
DNMT3B	0.363	0.094	14.925	0.000	1.437	1.196	1.727
MPO	-0.177	0.051	11.842	0.001	0.838	0.757	0.927
TMSB10	-0.541	0.160	11.393	0.001	0.582	0.425	0.797
PIWIL4	-0.375	0.103	13.211	0.000	0.687	0.561	0.841
TPM2	0.415	0.110	14.128	0.000	1.514	1.219	1.879
PYROXD1	-0.445	0.139	10.291	0.001	0.641	0.488	0.841
SVIP	0.436	0.134	10.516	0.001	1.546	1.188	2.012
SOCS2	0.221	0.063	12.370	0.000	1.247	1.103	1.411

LAPTM4B	0.171	0.050	11.570	0.001	1.186	1.075	1.309
SH3BP4	0.326	0.095	11.632	0.001	1.385	1.149	1.670
EVI2B	-0.391	0.111	12.391	0.000	0.676	0.544	0.841
PLIN2	-0.344	0.105	10.707	0.001	0.709	0.577	0.871
PXK	-0.393	0.125	9.802	0.002	0.675	0.528	0.863
YPEL2	-0.448	0.134	11.139	0.001	0.639	0.491	0.831
MXRA7	0.405	0.117	11.915	0.001	1.499	1.191	1.886
RNF130	-0.506	0.160	10.023	0.002	0.603	0.441	0.825
NPDC1	0.345	0.099	12.226	0.000	1.412	1.164	1.714
HOPX	0.203	0.066	9.459	0.002	1.225	1.077	1.395
TMSB4X	-0.643	0.197	10.644	0.001	0.526	0.357	0.774
AGTPBP1	-0.331	0.104	10.044	0.002	0.718	0.585	0.881
ALDH1A1	0.278	0.083	11.196	0.001	1.320	1.122	1.554
TES	-0.397	0.116	11.779	0.001	0.672	0.536	0.843
GIMAP7	-0.038	0.062	0.381	0.537	0.962	0.852	1.087
FEM1C	-0.337	0.134	6.356	0.012	0.714	0.550	0.928
ARHGAP22	0.318	0.098	10.483	0.001	1.374	1.134	1.665

SUPPLEMENTARY TABLE 5.11. UNIVARIATE BLR FOR THE BLR WORKFLOW. THE GENES IN THE HOVON DATASET WERE ANALYSED USING A ROC CURVE, ALL GENES WITH AN AUC OVER 0.6 WERE ANALYSED USING A UNIVARIATE BLR.