A transcriptomic and molecular approach uncovering ASCL2 as a novel tumourigenic gene in breast cancer

Faramarzi, N.

A transcriptomic and molecular approach uncovering ASCL2 as a novel tumourigenic gene in breast cancer

Nicola Faramarzi

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

May 2019

Abstract

Breast cancer is highly heterogeneous and is considered a collection of molecularly distinct tumour subtypes. Substantial efforts have been made to explore the gene expression profiles underlying the subtypes, and to elucidate possible markers associated with clinical outcomes. However, research in this area has been met with significant challenges and despite ongoing advancements in diagnostics and targeted therapeutics, incidence and mortality continues to rise. Thus, there is a need for greater molecular characterisation of breast tumours, to further understand the mechanistic roles of genes within their respective signalling pathways.

With the advent of high-throughput technologies in transcriptomics, as well as the use of open databases and bioinformatics analysis tools, it is now possible to examine thousands of genes in parallel, generating an unprecedented amount of information. This provides a means for researchers to identify novel genes and targets from large volumes of gene expression data. However, the task of extracting clinically relevant results, is a prominent challenge. Therefore, the aim of this study was to use a streamlined *in silico* pipeline, integrated with *in vitro* methods to identify and functionally investigate a novel genetic marker demonstrating a key role in breast carcinogenesis.

Gene expression profiles from breast cancer cell lines were obtained from public databases (Array Express and Gene Expression Omnibus). Data was filtered and subjected to an extreme variation analysis to generate a list of differentially expressed genes. Subsequently, multiple pathway analysis tools were used to identify a novel candidate gene for further investigation. Achaete-scute complex homolog 2 (*ASCL2*) is a transcription factor and Wnt-target gene, recognised as a regulator of stem cell identity and embryogenesis. Gene expression was validated *in vitro* by Reverse Transcription Quantitative Polymerase Chain Reaction (RT-qPCR), and to assess the tumourigenic potential of *ASCL2,* siRNA knockdown was performed; assays were employed to measure proliferation, wound-healing and apoptosis. Data mining of patient tumours obtained from the

METABRIC study was also undertaken to ascertain the potential of *ASCL2* as a prognostic indicator.

This work utilised a systematic pipeline used by the wider scientific community for the identification of candidate genes from transcriptomic data. Differential expression of *ASCL2* was observed across multiple breast cancer cell lines, with largest the expression seen in MCF7 cells. Although evidence did not support the usage of *ASCL2* as a prognostic indicator in patient tumours, data integrated from multiple lines of investigation suggested that this gene may influence the migratory capacity of breast tumour cells, whilst exercising its tumourigeneic function via the Wnt signalling pathway in breast cancer. Thus, this potential novel role of *ASCL2* in breast tumourigenesis highlights a prominent area for further exploration.

Acknowledgements

Completing this PhD has not only allowed me to grow as a scientist, but also as a person, and would not have been possible without the support and kindness from many people.

Firstly, I would like to acknowledge and extend my deepest gratitude to my supervisor, Dr Nadège Presneau, for providing me with this incredible opportunity. Thank you for continuously nurturing, encouraging and trusting me. My sincerest thanks also go to Dr Miriam Dwek, for providing clarity and significant guidance during the writing process. To both, thank you for generously donating your time to support the completion of this project. Thank you to the University of Westminster for providing me with a scholarship to undertake this work within such a fantastic and vibrant department.

I have made some wonderful friends during my PhD, and no conversation went unappreciated. Nasrin, you have been the most incredible, kind and warm lab mentor from the start; I wouldn't have wanted to share our secluded lab with anyone else! Louise, thank you for looking out for me and for suggesting helpful revisions to the final thesis. Nadeen, Camille, Q, Karima, Tayebeh, Ted and Tom - thank you for offering your advice, providing your expertise, lending your ears, and keeping me sane!

To Amy and Rhys – you are both such special people and have become very dear to me. Thank you both for providing an infinite number of laughs along the way, and for being the smartest and coolest people to share this experience with. Rhys, an extra shout out for lending your lab skills when it was needed!

Finally, but most importantly, I'd like to dedicate this thesis to my family. Mum and Dad – you are the most accommodating, selfless and thoughtful parents. Thank you for your unwavering guidance, support and generosity, and for knowing what I'm capable of before I do. Thank you for always providing me with direction, but also space when I've needed it, and for being my constants during times of change. To my sister Emma, thank you for never failing to provide laughs through my hopeless tears.

I'm so grateful to be surrounded by so many supportive and loyal family and friends, to which I deeply thank.

Declaration & Acknowledgment of Contribution

I declare that the present work was carried out in accordance with the Guidelines and Regulations of the University of Westminster. I declare that the material presented in this thesis is original and is my own work, unless otherwise acknowledged or referenced.

I would like to acknowledge the contribution made to this thesis by Dr Rifat Hamoudi (University of Sharjah & University College London) and Dr Nadège Presneau (University of Westminster), for downloading and normalising the raw Affymetrix microarray data for the human breast cancer cell lines, presented in Section 2.1.1. Dr Rifat Hamoudi also designed and executed the extreme variation filtering analysis.

Published Work

Faramarzi N, Dwek M, and Presneau N., (2018) PO-461 A transcriptomic and molecular approach to uncovering achaete-scute complex homolog 2 (ASCL2) as a potential novel driver gene in breast cancer. *ESMO Open 2018*; 3.

Contents

List of Figures

List of Tables

Abbreviations

AGO2	Protein Argonaute-2
Akt	Protein Kinase B
ALCAM	Activated Leukocyte Cell Adhesion Molecule, CD166
ANOVA	Analysis of Variance
ASCL2	Achaete-scute complex homolog 2
ATCC	American Type Culture Collection
ATM	Ataxia Telangiectasia Mutated Serine/Threonine Kinase
BCL-2	B-cell Lymphoma 2
bHLH	Basic Helix-Loop-Helix
BMP5/7	Bone Morphogenetic Protein 5/7
bp	Base Pair
BRAF	B-Raf Proto-Oncogene
BRCA1/2	Breast Cancer Type 1/2 Susceptibility Protein
BSA	Bovine Serum Albumin
C-MYC	Avian Myelocytomatosis Virus Oncogene Cellular Homolog
CCND1	Cyclin D1
CD133	Prominin-1
CD44	Cell Surface Glycoprotein CD44
CDH1	Cadherin 1, E-Cadherin
cDNA	complementary DNA
CHEK2	Checkpoint Kinase 2
CI	Confidence Interval
CO_2	Carbon Dioxide
CPT1A	Carnitine Palmitoyltransferase 1A
CRC	Colorectal Cancer
CSC	Cancer Stem Cell
Ct	Cycle Threshold
CTC	Circulating Tumour Cell
CTNNB1	Catenin Beta 1
CXCR4	Chemokine Receptor Type 4
DAVID	Database for Annotation, Visualization and Integrated Discovery
DMEM	Dulbecco`s Modified Eagle Media
DMSO	Dimethyl Sulfoxide
DNA	Deoxyribonucleic Acid
dNTP	Deoxyribonucleotide Triphosphate
DsiRNA	Dicer-Substrate siRNA
dsRNA	Double-strand RNA
DTX3	Deltex E3 Ubiquitin Ligase 3
EdU	5-ethynyl-2’-deoxyuridine
EGFR	Epidermal Growth Factor Receptor
EIF2S2	Eukaryotic Translation Initiation Factor 2 Subunit Beta
EIF6	Eukaryotic Translation Initiation Factor 6
EMT	Epithelial–mesenchymal Transition
ER	Oestrogen Receptor
ERBB	Epidermal Growth Factor Receptor Family
ERBB2	Tyrosine Kinase-Type Cell Surface Receptor HER2
FBS	Fetal Bovine Serum
FCS	Fetal Calf Serum
FDA	Food and Drug Administration
FDR	False Discovery Rate
FGD5	FYVE, RhoGEF And PH Domain Containing 5
FGF	Fibroblast Growth Factor
FGFR1	Fibroblast Growth Factor Receptor 1

FOXA1	Forkhead Box A1
FZD7	Frizzled Class Receptor 7
GATA3	GATA Binding Protein 3
GCRMA	GeneChip Robust Multi-array Averaging
GEO	Gene Expression Omnibus
GFP	Green Fluorescent Protein
GNAS	Guanine Nucleotide Binding Protein, Alpha Stimulating Activity Polypeptide 1
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
GWAS	Genome Wide Association Study
HER2	Human Epidermal Growth Factor Receptor 2
HPRT	Hypoxanthine Phosphoribosyltransferase
HR	Hazard Ratio
ID4	Inhibitor of DNA Binding 4
IDH1	Isocitrate Dehydrogenase 1
INDEL	Insertion or Deletion
JAK/Stat	Janus Kinase/Signal Transducer and Activator of Transcription
JUN	Jun Proto-Oncogene, AP-1 Transcription Factor Subunit
KEGG	Kyoto Encyclopedia of Genes and Genomes
KRAS	Kirsten Rat Sarcoma Viral Proto-Oncogene
L1CAM	L1 Cell Adhesion Molecule
LGR5	Leucine-rich repeat-containing G-protein coupled receptor 5
LRP6	Low-Density Lipoprotein Receptor-Related Protein 6
MAP2K4	Mitogen-Activated Protein Kinase Kinase 4
MAP3K1	Mitogen-Activated Protein Kinase Kinase Kinase 1
MAPK	Mitogen-Activated Protein Kinase
MAS5	Microarray Analysis Suite 5
MCL1	Myeloid Cell Leukemia Sequence 1 (BCL2-Related)
MDR	Multi-Drug Resistance
MEBM	Mammary Epithelial Cell Basal Medium
METABRIC	Molecular Taxonomy of Breast Cancer International Consortium
METTL6	Methyltransferase Like 6
$MgCl_2$	Magnesium Chloride
MIAME	Minimum Information About a Microarray Experiment
miRNA	MicroRNA
mRNA	Messenger RNA
MRPS23	Mitochondrial Ribosomal Protein S23
MSigDB	The Molecular Signatures Database
mTOR	Mammalian Target of Rapamycin
MUC18	Cell Surface Glycoprotein MUC18, CD146
NC	Negative Control
NCBI	National Center for Biotechnology Information
NES	Normalised Enrichment Score
NGS	Next Generation Sequencing
OCT4	Octamer-Binding Transcription Factor 4
ORA	Overrepresentation Analysis
PALB2	Partner and Localizer of BRCA2
PAM50	Prosigna Breast Cancer Prognostic Gene Signature Assay
PANTHER	Protein Analysis Through Evolutionary Relationships
PARP	Poly ADP Ribose Polymerase
PBS	Phosphate Buffered Saline
PCR	Polymerase Chain Reaction
PI3K	Phosphoinositide 3-Kinases
PR	Progesterone Receptor
PRS	Polygenic Risk Score
PTEN	Phosphatase and Tensin Homolog

R^2	Coefficient of Determination
RFP	Red Fluorescent Protein
RISC	RNA-Induced Silencing Complex
RNA	Ribonucleic Acid
RNAi	RNA Interference
RPII	RNA Polymerase II
RPMI	Roswell Park Memorial Institute
RT-qPCR	Reverse Transcription Quantitative Polymerase Chain Reaction
SEM	Standard Error of the Mean
sFRP1	Secreted Frizzled Related Protein 1
shRNA	Short Hairpin RNA
siRNA	Small Interfering RNA
SLC2A10	Solute Carrier Family 2 Member 10
SNP	Single-Nucleotide Polymorphism
SOX2	Sex Determining Region Y-Box 2
STK11	Serine/Threonine Kinase 11
STR	Short Tandem Repeat
SURV	Survivin/ Baculoviral Inhibitor of Apoptosis Repeat-containing 5
TCF	T-cell Factor
TCGA	The Cancer Genome Atlas
TGF-B	Transforming Growth Factor Beta
TGS	Third Generation Sequencing
TNBC	Triple Negative Breast Cancer
TP53	Tumour Protein 53
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
Wnt	Wingless/Integrated

Chapter I

Introduction

1.1 An Introduction to Breast Cancer

1.1.1 Introduction to Cancer

The term cancer refers to a group of widespread diseases caused by the uncontrolled growth of abnormal cells that have the potential to spread and invade surrounding healthy tissue. In the UK, 1 in 2 people have a risk of being diagnosed with cancer during their lifetime (Cancer Research UK, 2019).

Cancer is a highly complex and multi-step process now considered predominantly as a disease entailing genetic changes (Hanahan, & Weinberg, 2011; Vogelstein, *et al.*, 2013); mutations or alterations in the expression of certain genes ultimately sustain the development and growth of tumours. These genes can be largely classified as a loss-of-function tumour suppressor gene, responsible for preventing uncontrolled growth and encouraging cell cycle checkpoints and DNA repair, or a gain-of-function oncogene, which promotes cell proliferation and survival (Hanahan, & Weinberg, 2011; Lee, & Muller, 2010).

Despite its complexity, the development of cancer has been rationally classified into a handful of underpinning capabilities that tumours use to govern neoplastic growth. These are now universally known as the Hallmarks of Cancer (Figure 1.1), first described by Hanahan and Weinberg in 2000, and later revised in 2011. These key principles cause errors in cellular communication networks (signalling pathways), which ultimately affect cell growth, death, motility, metastasis, replicative immortality, evasion of the immune system, and instability of the genome (Hanahan, & Weinberg, 2011; Hanahan, & Weinberg, 2000).

Cancer can arise from familial (inherited) or acquired genetic mutations, but may also develop as a result of gene expression changes and other epigenetic modifications (Rizzolo, *et al.*, 2011). The majority of cancers arise from acquired somatic genetic mutations, which can result from factors such as external carcinogens, DNA instability or deregulation of gene expression (Greenman, *et al.*, 2007). These genetic changes are able to give cells a selective growth advantage, which accumulate over time and lead to disruptions in gene activity, and subsequently changes in cell behaviour. One could argue therefore, that all of the hallmarks affecting cellular behaviour depend on some type of genomic alteration in tumour cells, thus the ‘genomic instability and mutation’ hallmark could be considered to underpin all of the other hallmarks (Figure 1.1) (Hanahan and Weinberg, 2011).

Metastasis is the cause of approximately 90% of cancer related deaths, and occurs due to the genetic instability of primary cancer cells which have an invasive capacity; these may enter the circulation and eventually adapt to a distant tissue microenvironment (Gupta, & Massagué, 2006). Owing to the deaths related to invasion, this calls for greater characterisation of the genetic basis and markers leading to tumour metastasis (Bos, *et al.*, 2009).

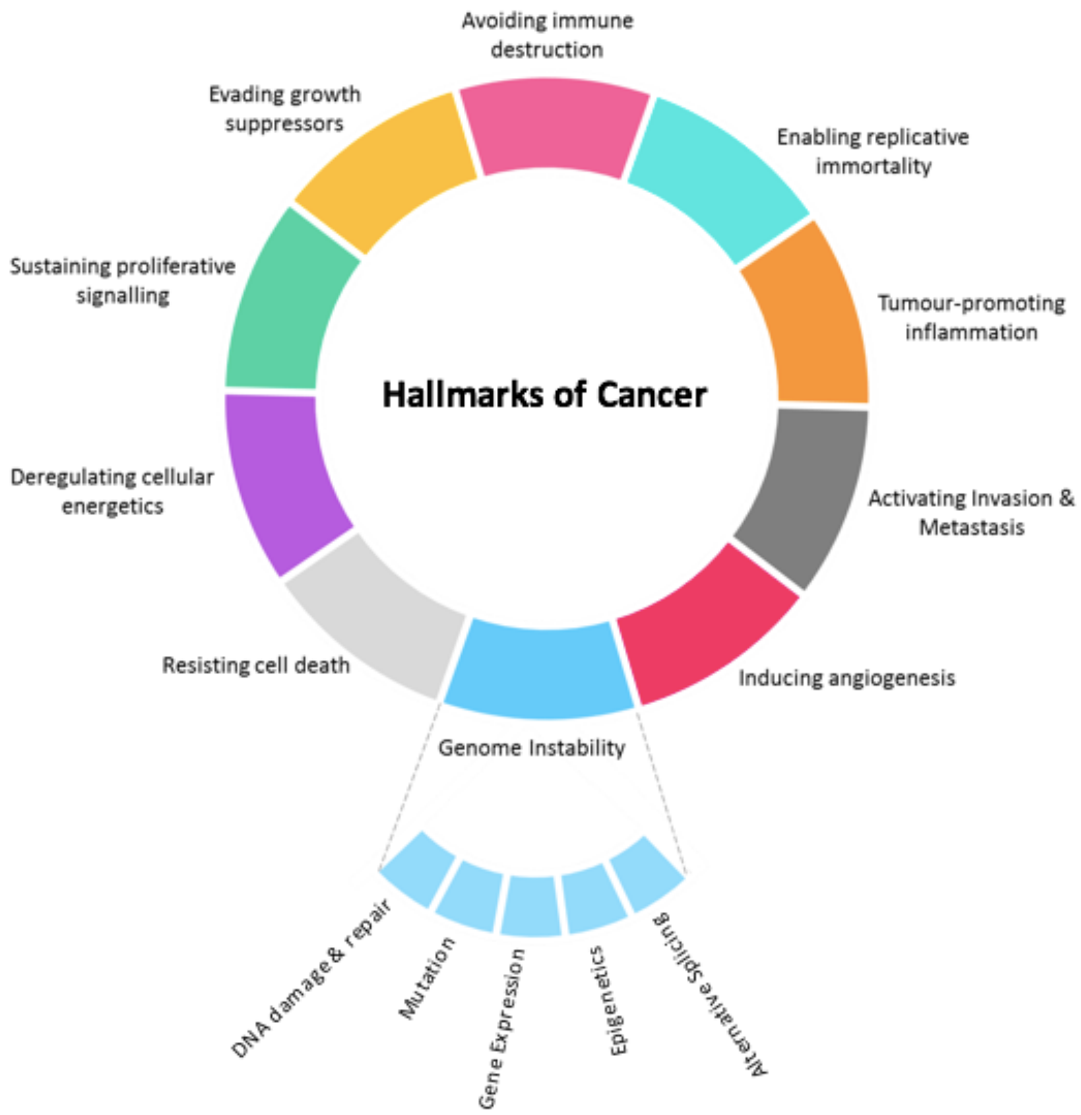

Avoiding immune destruction
Evading growth suppressors
Enabling replicative immortality
Sustaining proliferative signalling
Tumour-promoting inflammation
Hallmarks of Cancer
Deregulating cellular energetics
Activating Invasion & Metastasis
Resisting cell death
Inducing angiogenesis
Genome Instability
DNA damage & repair
Mutation
Gene Expression
Epigenetics
Alternative Splicing

Figure 1.1. The 10 Hallmarks of Cancer as described by Hanahan & Weinberg (2011), with focus on the 'Genome Instability' hallmark, responsible for the genetic variation seen in cancer cells.

1.1.2 Breast Cancer

Breast cancer was classically considered as three predominant subtypes with distinct features and behaviours – hormone receptor positive (HR+) (expressing oestrogen receptors, ER, and progesterone receptors, PR), human epidermal growth factor receptor positive (HER2+) and triple negative breast cancer (TNBC). These subtypes represent somewhat clear therapeutic groups. Broadly speaking, hormone receptor positive tumours are the most common, typically presenting with a good prognosis due to treatment using endocrine therapy. Up to approximately 10% of breast cancers are diagnosed as HER2+, and previously indicated a poor prognosis. Yet, treatment with anti-HER2 targeted agents has since provided a large benefit to patients in this group. Breast tumours that do not express any of the aforementioned receptors, are referred to as triple negative. These tumours are aggressive in nature, with limited treatment options, resulting in a high mortality rate (Karagoz, *et al.*, 2015; Perou, & Børresen-Dale, 2011; Reis-Filho, & Pusztai, 2011; Yeo, & Guan, 2017).

However, it has become well-known that breast cancer is highly heterogeneous, and the viewpoint that it is a single disease with varying histology is extremely outdated. It is now accepted that the term 'breast cancer' refers to a collection of tumour subtypes with distinctive aetiologies, origins, genetic signatures and clinical outcomes; this complex picture encompassing a number of tumour entities is considered in further detail in Section 1.1.4 (Karagoz, *et al.,* 2015; Perou, & Børresen-Dale, 2011; Reis-Filho, & Pusztai, 2011; Yeo, & Guan, 2017). This heterogeneity is mirrored in its complex genomic landscape, and despite advancements in subtype-specific therapeutics, many patients are not responsive to therapies, or present with unexpected tumour behaviour due to underlying genetic mutations that require characterisation.

In the UK, breast cancer is the most common type of cancer, and incidence is projected to rise by 2% by 2035; over 11,500 women died of the disease in 2016 (Cancer Research UK, 2019). However, due to increasing knowledge of underlying molecular aberrations and the development of subsequent treatments, survival rates have doubled since the 1970s (Figure 1.2) (Cancer Research UK, 2019). In the western world, breast cancer follows lung cancer as the second biggest cause of cancer-related mortality (Fadoukhair, *et al.,* 2015). On a global scale, breast cancer is the most frequently diagnosed cancer in women; approximately 1.67 million women were diagnosed with invasive breast carcinoma worldwide, with an estimated 522,000 women dying from the disease in 2012 (Liu, *et al.,* 2015). This was estimated to reach 627,000 deaths in 2018 (Bray, *et al.,* 2018). Prominent sites of breast cancer metastasis are the brain, bones and lungs (Bos, *et al.,* 2009). These statistics necessitate research focussed on further genomic characterisation that nurture the growth of breast cancer and that may be beneficial as pharmaceutical targets; this can then be translated to patient-tumour-specific molecular diagnosis, partnered with a personalised treatment strategy in the near future.

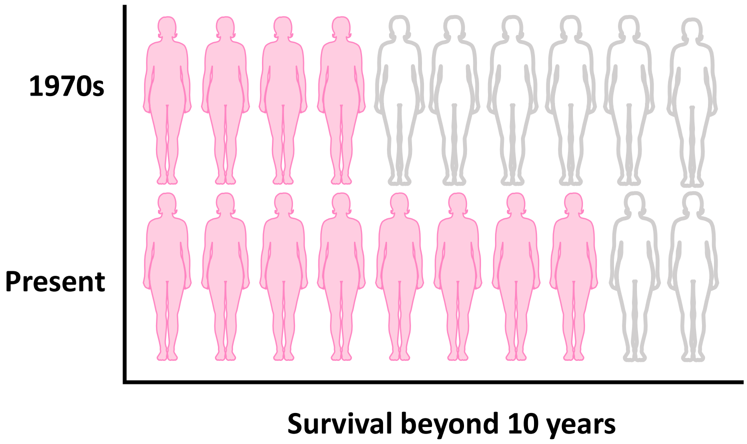


Figure 1.2. Cancer Research UK statistics illustrating that since the 1970s, breast cancer survival beyond 10 years has increased by half (statistics taken from Cancer Research UK, 2019).

1.1.3 Breast Cancer Risks

The risk factors associated with the growth of breast tumours are multifactorial. Primarily, inherited genetic variants have been identified as conferring a greater susceptibility to breast cancer in patients; these patients have a hereditary predisposition, and are considered to be high-risk, thus benefit greatly from screening and regular follow-up appointments. Notable genetic loci contributing to an increased familial risk are the high-penetrance *BRCA1* and *BRCA2* mutations on chromosome 17 and 13 respectively, responsible for disturbances in DNA damage and repair mechanisms (Vogelstein, *et al.*, 2013). Other rare, yet high-penetrance genes include *PTEN, CDH1, STK11*, and moderate-penetrance genes such as *CHEK2, ATM,* and *PALB2,* also screened in medical genetic practice (Antoniou, *et al.*, 2014; Michailidou, *et al.*, 2013; Shiovitz, & Korde, 2015; van der Groep, *et al.*, 2011). Although mutations in these genes result in a high-risk of developing breast cancer in the individual, within the general population, these only account for a small proportion of cases.

Yet, genome-wide association studies (GWAS) have now identified a number of breast cancer susceptibility variants (single-nucleotide polymorphisms, SNPs) that are of small risk individually, but may be grouped together to confer a substantial combined effect. These are known as a polygenic risk score (PRS) and can be utilised within preventative screening strategies to stratify women according to their risk of breast cancer, identifying patients most likely to benefit from intervention (Mavaddat *et al.,* 2019; Mina & Arun, 2019). A landmark study by Mavaddat *et al.,* (2019) reported the development and validation of subtype-specific polygenic risk scores for breast cancer, especially for the improved prediction of ER-negative breast cancer. The study identified a PRS of 313 SNPs significantly more predictive of risk (accounting for subtype, age and family history) than previously reported risk scores (Mavaddat *et al.,* 2019).

There are many hormonal and reproductive risk factors also associated with breast cancer, such as early menarche, late menopause, nulliparity and breast feeding. Early menarche in women and late menopause is associated with an increased breast cancer risk due to greater exposure to associated hormones (oestrogen and progesterone). However, earlier full term pregnancy in younger women lowers the risk of breast cancer. Women whom have not carried a pregnancy nor given birth are at a slightly increased risk of developing breast cancer after 40 years of age, but not at younger ages.

Other exogenous and lifestyle factors play a large role in the risk of developing breast cancer. Over the years, a number of studies have found an association between alcohol intake and increased breast cancer risk; compared to women who did not drink alcohol, the relative risk of breast cancer increased by approximately one third. This risk increased as consumption increased (McDonald *et al.,* 2013). Overall, evidence supports that lifetime moderate alcohol intake increases the risk of breast cancer (odds ratio [OR] = 2.13) (Terry *et al.,* 2006).

Smoking and alcohol consumption have been shown to be strongly associated with DNA methylation in breast tumourigenesis as well as in a range of other cancers (Catsburg, *et al.*, 2015; Christensen, *et al.*, 2010; Passarelli, *et al.*, 2016). Smoking is well known to cause direct and indirect DNA damage and instability, as well as changes in DNA damage responses, and it is estimated that heavy smoking for 40 years can cause up to 1,000 DNA aberrations in all cells (Alexandrov, *et al.*, 2016; Lord, & Ashworth, 2012). However, in spite of this, most prospective cohort studies have found no causal association between both active or passive smoking and incidence of breast cancer, highlighting little to no relationship between smoking and breast cancer risk (OR = 1.00) (Luo *et al.,* 2011; Prescott *et al.,* 2007; Lash & Aschengrau, 2002; Ahern *et al.,* 2009).

Another exogenous lifestyle factor specifically relating to breast cancer risk is the prolonged use of oral contraceptives, which have been proven to confer an increased risk of developing oestrogen receptor positive (ER+) breast cancer, depending on the variable formulation of the pill (Beaber, *et al.*, 2014; Mørch, *et al.*, 2017). Obesity and diet can also have an effect on breast cancer risk, as dietary changes have been known to prevent approximately one-third of cases (Brennan, *et al.*, 2010).

DNA damaging agents can also increase the risk of breast and other solid tumours. These agents can execute damage through occupational exposure, chemical warfare, or via genotoxic and mutagenic chemotherapeutic drugs. These can affect DNA repair pathways such as base excision and mismatch repair. Alkylating agents, such as melphalan, are used in the treatment of solid tumours and have the potential to give rise to secondary cancers. It is for this reason that treating patients with ineffective chemotherapy is to be avoided, as cytotoxic damage can occur in other tissues. These risks bombard DNA with damage over time, causing somatic aberrations that accumulate and eventually have the potential to drive carcinogenesis.

1.1.4 Classification of Breast Cancer

Over the years, the classification of breast cancer has advanced from being focussed on morphological features, to being consolidated with specific biomarkers and clinical features. As previously mentioned, breast cancer was traditionally categorised into three main subtypes – HR(ER/PR)+, HER2+ and TNBC. This classification has now been superseded by the 'intrinsic subtypes' taxonomy, first classified in 2000, and was developed to better reflect the gene expression and molecular patterns of the tumours (Perou, & Børresen-Dale, 2011). A summary of this is shown in Table 1.1 and Figure 1.3 encompassing the five intrinsic subtypes known as Luminal A, Luminal B, HER2-enriched, Claudin-low and Basal-like (Eroles, *et al.*, 2012).

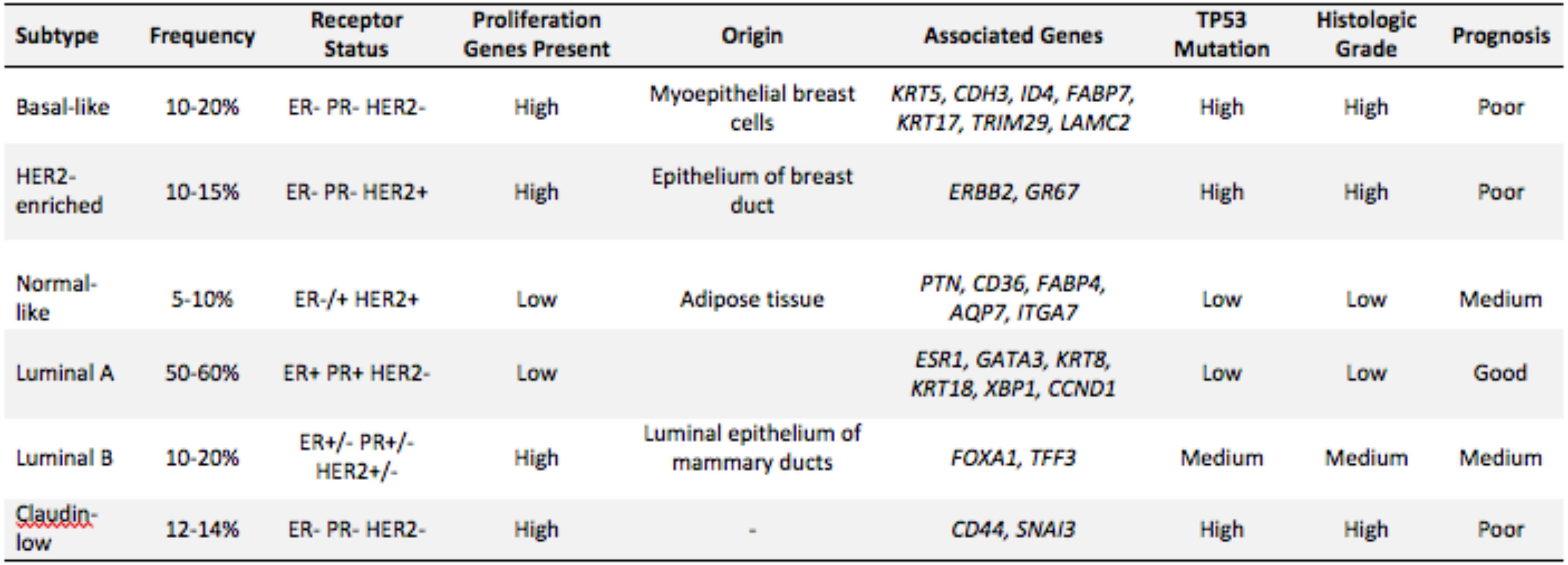
Luminal tumours are known to express hormone receptors (ER/PR) and are (broadly speaking) divided into types A (ER+, PR+, HER2-) and B (ER+, PR+, HER2+). Luminal tumours hold the status of the most common of the breast cancer subtypes, with Luminal A representing the majority. Although generally prognosis is good for these tumours, Luminal A tumours have a significantly better prognosis than Luminal B; these patients can be treated with and respond well to endocrine therapy in most cases, such as tamoxifen, as traditional chemotherapy can be less effective. Altered gene expression patterns of these tumours are commonly associated with ER activating genes (Dai, *et al.*, 2015).

Tumours within the HER2-enriched subtype over-express the HER2 (*ERBB2*) protein only and are ER and PR negative. Although these tumours are generally sensitive to some chemotherapies, they were traditionally known to be associated with a poorer prognosis than Luminal tumours due to their high risk of relapse (Dai, *et al.*, 2015). However, after the approval of trastuzumab in 2001, advances in the therapeutics used to treat these tumours have greatly improved the clinical management and survival outcomes of patients diagnosed with HER2+ breast cancer (Ortiz *et al.,* 2019).

As the name suggests, TNBC lacks the expression of hormone (ER/PR) and HER2 receptors, and because of the lack of evident target, is linked with an extremely poor prognosis. Its aggressive nature has attracted much research interest due to its lack of available molecular treatment targets (Karagoz, *et al.*, 2015). Lehmann, *et al.*, (2011) has classified TNBC into six further subtypes displaying distinctive gene expression patterns – basal-like 1 and 2, immunomodulatory, mesenchymal, mesenchymal stem-like, and luminal androgen receptor. Further characterising these tumours illustrates the complexity of breast cancer, and the notion that each subtype requires its own treatment strategy despite similar phenotypes, due to the unique gene expression patterns harboured by each individual tumour.

More recently in 2012, the METABRIC study, integrating genomic and transcriptomic sequencing, as well as long-term clinical follow-up, characterised 2000 primary breast tumours and identified 10 molecularly distinct subtypes, known as the ‘integrative clusters’ (Table 1.2) (Curtis, *et al.*, 2012; Dawson, *et al.*, 2013a). This ever expanding and scrutinised system of classifying breast cancer illustrates the complexity and heterogeneity of the disease, and exemplifies the importance of a specific classification system to aid the administration of efficient treatments in patients.

Table 1.1. A summary of the intrinsic subtype taxonomy and current molecular markers, adapted from Eroles *et al.* (2012).



Subtype	Frequency	Receptor Status	Proliferation Genes Present	Origin	Associated Genes	TP53 Mutation	Histologic Grade	Prognosis
Basal-like	10-20%	ER- PR- HER2-	High	Myoepithelial breast cells	*KRT5, CDH3, ID4, FABP7, KRT17, TRIM29, LAMC2*	High	High	Poor
HER2-enriched	10-15%	ER- PR- HER2+	High	Epithelium of breast duct	*ERBB2, GR67*	High	High	Poor
Normal-like	5-10%	ER-/+ HER2+	Low	Adipose tissue	*PTN, CD36, FABP4, AQP7, ITGA7*	Low	Low	Medium
Luminal A	50-60%	ER+ PR+ HER2-	Low		*ESR1, GATA3, KRT8, KRT18, XBP1, CCND1*	Low	Low	Good
Luminal B	10-20%	ER+/- PR+/- HER2+/-	High	Luminal epithelium of mammary ducts	*FOXA1, TFF3*	Medium	Medium	Medium
Claudin-low	12-14%	ER- PR- HER2-	High	-	*CD44, SNAI3*	High	High	Poor

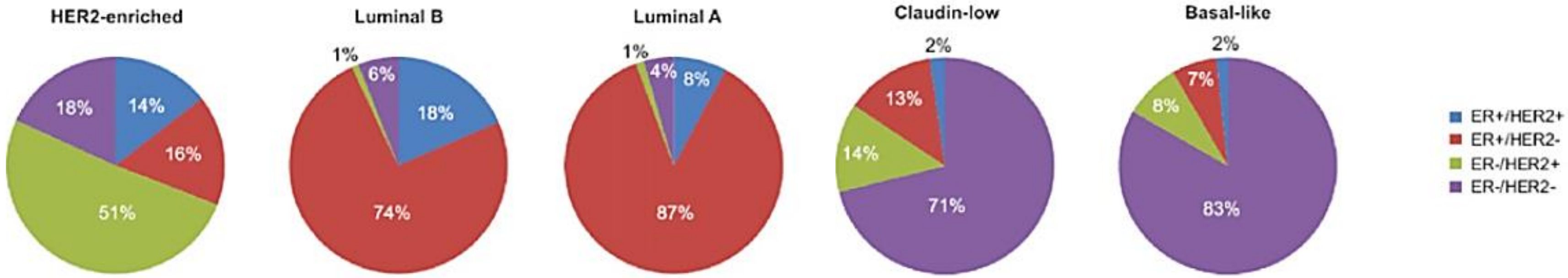


Figure 1.3. The mRNA expression of ER and HER2 across the breast cancer intrinsic subtypes, Eroles *et al.* (2012).

Table 1.2. A summary of the integrative cluster taxonomy as proposed by Dawson, *et al.,* (2013).

Subtype	Receptor Status	Genomic Instability	Associated Mutations	Pathobiology	Prognosis
IntClust 1	ER+, Luminal B predominantly	High	Amplification of 17q23 locus High prevalence of *GATA3* mutations	High proliferation	Intermediate
IntClust 2	ER+, Luminal A and B	High	Amplification of 11q13/14	Aggressive pathophysiology	Worst prognosis of all ER+ tumours
IntClust 3	Mainly Luminal A	Low		Small low-grade tumours	Good, best of all clusters
IntClust 4	ER +/- Can be triple negative. Mix of the intrinsic subtypes	Low			Favourable outcome
IntClust 5	ERBB2 amplified HER2+/ER- Luminal ER+	Intermediate		Presents at younger age High grade tumours	Poor
IntClust 6	ER+, Luminal A/B	High	Amplification of 8p12 locus Low levels of *PIK3CA* mutations		Intermediate
IntClust 7	ER+/PR+, Luminal A	Intermediate		Low grade well differentiated tumours	Good
IntClust 8	ER+, mainly Luminal A		1q gain/16q loss. High *PIK3CA, GATA3* mutations	Low grade well differentiated tumours	Good
IntClust 9	Mix of intrinsic subtypes Mainly ER+ Luminal B	High	Increased *TP53* mutations		Intermediate
IntClust 10	Mostly triple negative Basal-like tumours	Intermediate	High *TP53* mutations	Presents at young age High grade, poor differentiated	High risk pre-5 years Good post-5 years

1.1.5 Current Diagnostics & Therapeutics

At present, subtypes are diagnosed on the basis of immunohistochemical analysis of a tissue biopsy, and in situ hybridisation to detect a single gene amplification if results are equivocal (Hansen, & Bedard, 2013). However, it could be argued that this basic diagnosis using a panel of so few histopathological markers is not reflective of most tumour types and treatment requirements. Hence, there is a requirement for more widely available clinical molecular screening programmes and molecular diagnostic testing to shed light on patient specific genetic signatures that can be more efficiently targeted.

Although systemic therapy for the treatment of breast tumours is determined by subtype, few approved drugs have emerged in the past few decades in the hope of specifically targeting biomarkers of breast cancer. Those demonstrating success include endocrine therapies, aromatase inhibitors, or tamoxifen directed at treating ER+ breast cancers, and the anti-HER2 class of monoclonal antibodies (such as trastuzumab or pertuzumab), and tyrosine kinase inhibitors (TKI's, such as lapatinib), used for treatment of HER2+ tumours. Trastuzumab was the first therapy to specifically target HER2+ metastatic tumours, and is considered a leading example of using patient genomic profiles to direct treatment decisions (Fadoukhair, *et al.*, 2015; Hansen, & Bedard, 2013).

While targeted treatments such as tamoxifen and trastuzumab arose to improve the survival of patients with hormone receptor positive and HER2+ cancers respectively, these were initially met with unexpected inefficiency in a large number of patients who were thought to possess the corresponding molecular markers for these treatments (Rexer, & Arteaga, 2013; Vu, & Claret, 2012). Some patients with HER2+ breast cancer fail to respond to trastuzumab and ultimately develop progressive disease, likely due to the manifestation of a resistant phenotype. However, new strategies have been developed to circumvent the different acquired therapeutic resistance mechanisms observed in some HER2+ tumours; for example, utilising a combination of PI3K inhibitors and cyclin D1-cyclin-dependent kinase 4/6 (CDK 4/6) inhibitors alongside anti-HER2 agents, to target the alterations that lead to hyperactivation of downstream signalling in the PI3K/AKT/mTOR axis, which otherwise instigate and perpetuate resistance to

HER2 targeting therapies (Goel *et al.,* 2016; Ortiz *et al.,* 2019; Vernieri *et al.,* 2019).

TNBC is treated with often ineffective traditional or adjuvant chemotherapy, however recent efforts to produce targeted therapies has brought about the advent of poly (ADP) ribose polymerase inhibitors (PARP) (for *BRCA1* tumours and TNBC), epidermal growth factor receptor inhibitors (*EGFR*) and antiangiogenic agents (Karagoz, *et al.*, 2015). Despite this, adjuvant chemotherapy is still administered to approximately 60% of all early-stage breast cancer patients, with only up to 15% of these patients benefitting (Reis-Filho, & Pusztai, 2011). It is important to only treat tumours likely to be responsive to genotoxic chemotherapy, as all patients are at risk of the highly toxic side effects (Reis-Filho, & Pusztai, 2011).

1.1.6 The Problem with Current Therapies

It is now accepted that breast cancer as a disease requires the interplay of multiple signalling pathways that are capable of sustaining proliferative signalling and evading apoptosis, among the other hallmarks of cancer (Hanahan, & Weinberg, 2011). These cause distinct tumours in each patient likely to benefit from a more personalised treatment approach. Factors such as nodal status or tumour burden are no longer thought to be the only determinants of treatment response, but rather the molecular characteristics of the tumour (Reis-Filho, & Pusztai, 2011). However, current therapies are not patient specific and are usually targeted at the 'average' population rather than smaller, targeted groups of patients with certain biomarkers or molecular signatures. Nevertheless, before new pharmaceuticals can be designed and administered to patients for a more 'personalised' approach, new and robust molecular markers need to be identified, thoroughly examined and validated (Eroles, *et al.*, 2012).

Ali, *et al.*, (2016) investigated metastatic TNBC, a subtype for which there is no effective targeted therapy after conventional platinum-based and anthracycline chemotherapies become ineffective. This study used in-depth genomic profiling of the patient's tumour to identify a commonly mutated anti-apoptotic gene, *MCL1*, in TNBC, which then dictated personalised treatment using sorafenib and vorinostat (preclinical evidence demonstrated the efficacy of these drugs for TNBC with *MCL1* amplification) (Ali, *et al.*, 2016).

This study highlights that despite the failure of many rounds of differing conventional therapies, treatment based on the patient's tumour genetic profile prevailed and extended survival. Therefore, characterising further subtype-specific genetic markers driving breast cancer tumorigenesis is extremely valuable in developing a greater selection of targeted drugs to aid in designing tailored treatment regimens for patients. This has the potential to limit the amount of toxic chemotherapeutic agents administered to patients who may have tumours resistant to conventional therapies.

However, the differing molecular characteristics per subtype is not the only challenge. The dissemination of cancer cells that migrate away from the primary tumour via the lymphatic and circulatory system, and metastasise as secondary tumours in other regions of the body, is a main cause of death in women (Braune, *et al.*, 2018). Another significant hurdle for researchers and clinicians alike is multidrug resistance (MDR); in these cases, tumour cells develop the ability to acquire resistance and escape the cytotoxic effects of drugs and chemotherapeutic agents, leading to increased cell survival and thus resulting in ineffective treatment and inevitable relapse (Dewangan, *et al.*, 2017).

Cancer stem cells (CSCs) also present another prominent challenge to current treatment strategies. These cells have multiple characteristics that allow them to evade cytotoxic treatment. Firstly, their ability to self-renew, regenerate and differentiate permits the accumulation of genetic mutations. Secondly, these cells possess a quiescent nature, which is protective against conventional treatments that target rapidly diving cells. In addition, their capacity to self-renew gives rise to the production of multiple heterogeneous cancer cell lineages that make up a tumour (Dewangan, *et al.*, 2017). With these challenges in mind, research efforts have taken a shift towards selectively targeting therapy-resistant cells and CSCs, by interfering with the mechanisms and signalling pathways that these cells rely on.

1.2 Breast Cancer Signalling Networks

It is well known that genetic and epigenetic changes are responsible for driving the development of breast cancer. These changes most often correspond to a signalling pathway that controls cellular functions and allows cells to communicate; hyperactivation or inactivation of these pathways disrupts the homeostasis of cells and can alter downstream signalling networks, leading to a cascade of disturbances and gives rise to the hallmarks of cancer (Sever, & Brugge, 2015).

Frequent mutations in various parts of the PI3K pathway have been confirmed by many studies and represent a very relevant tumourigenic pathway in breast cancer (Yang, *et al.*, 2016). Sustained proliferative signalling in breast cancer may also be a result of deregulated HER2 signalling in some tumours. The human epidermal receptors have tyrosine kinase activity which can activate JAK/Stat and Ras/Raf/MAPK pathways, as well as PI3K/Akt/mTOR signalling (Eroles, *et al.,* 2012). Studies have also shown that Wnt/β-catenin signalling, a regulator of migration, differentiation and proliferation, is implicated in TNBC due to upregulated Wnt receptors (FZD7 and LRP6). Salinomycin has been identified as an inhibitor of Wnt/β-catenin signalling (primarily expression of *LRP6*) and a specific killer of CSCs in breast cancer, lending itself to be a potential treatment option in the future (King, *et al.*, 2012). Notch signalling has also been implicated in breast cancer, with Notch1 being shown to promote the epithelial-mesenchymal transition, thus stimulating proliferation and subsequent metastasis (Bolós, *et al.*, 2013). Other less characterised pathways, such as the stress-induced JUN-kinase pathway has been suggested to predict chemosensitivity of ER+ tumours via mutations in *MAP3K1* and *MAP2K4* (Hansen, & Bedard, 2013; Xue, *et al.*, 2018)

Knowledge of deregulated signalling pathways can aid treatment development or selection for breast cancer patients. For example, for patients with HER2+ breast cancer that are resistant to trastuzumab, a monoclonal antibody able to downregulate HER2 from the surface of the cell, there are multiple known mechanisms whereby cancer cells can escape the action of this drug. The PI3K-Akt pathway is now known to aid in the anti-tumour effect of trastuzumab. Hence in these cases, a PI3K inhibitor may be administered alongside trastuzumab for greater treatment efficiency (Rexer, & Arteaga, 2013).

There are many emerging pharmaceutical agents targeting the numerous aberrant signalling pathways in breast cancer (Table 1.3). The quantity of dysfunctional pathways is further evidence of the complex heterogeneity in breast cancer, and demands greater characterisation of the aberrant genes perturbing these transduction pathways. Many large-scale sequencing projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) have been undertaken, revealing the complexity of the full molecular profile of breast cancer. The most frequently mutated genes are *TP53* and *PIK3CA* (36-37%), but there are many other persisting gene mutations that occur at much lower frequencies which may prove to be significant targets for drugs (Hansen, & Bedard, 2013).

Table 1.3. Current aberrant signalling pathways identified to be involved in breast cancer and the agents in development to target them. Adapted from Fadoukhair *et al.*, (2015) and Liang, *et al.*, (2016).

Pathway	Targets	Agents
PI3K/Akt/mTOR	mTORC1 Isoform-selective PI3K inhibitors Dual mTOR/PI3K inhibitiors	Everolimus, temsirolimus, rapamycin GDC-0023 XL765
ErbB/HER2	EGFR inhibitor HER2 inhibitor	Erlotinib, gefitinib, cetuximab Pertuzumab, trastuzumab-DM1
MAPK	MEK1/2 inhibitors	Binimetinib, cobimetinib
FGF	Multi-targeted FGFR inhibitors	Dovitinib, lucitanib
Notch	Gamma-secretase inhibitor	PF-03084014
PARP	PARP1/2	Olaparib, niraparib
Wnt	NOP14, BKCa channels, Emilin2, WISP, NRBP1, TRAF4, Wntless	Salinomycin

1.2.1 Wnt Signalling in Breast Cancer

Wnt signalling is a highly important network for development, and is renowned for its involvement in cancer; this pathway is particularly considered a hallmark of colorectal cancer (Basu, *et al.*, 2016; Howe, & Brown, 2004). The pathway is divided into the canonical (β-catenin dependent) and non-canonical, (β-catenin independent) pathway – for the purpose of this thesis, the canonical Wnt signalling pathway will be of focus.

The main purpose of Wnt signalling in normal cells is to regulate embryonic development, adult homeostasis and tissue morphogenesis, chiefly using β-catenin to transduce Wnt signals from the extracellular membrane, into the nucleus from the cytoplasm (Figure 1.4). This pathway must be strictly regulated at each stage in the network, from ligand-receptor binding to transcriptional and post-transcriptional control, to maintain its finely-tuned normal activity (Zarkou, *et al.*, 2018). In synergy with other pathways, such as TGF-β and Notch, Wnt signalling maintains cellular homeostasis by regulating migration, differentiation, proliferation and apoptosis; it has also been reported that Wnt signalling encourages self-renewal in stem cells (Howe, & Brown, 2004; Zarkou, *et al.*, 2018).

Therefore, it is expected that abnormal activation of the Wnt pathway is responsible for tumour initiation and growth, motility and migration, and invasion, as well as playing an integral role in the epithelial-to-mesenchymal-transition (EMT) in cancer (Anastas, & Moon, 2013). In the past few decades, activated Wnt signalling involving β-catenin has become increasingly studied in breast cancer, and elevated β-catenin levels are commonly observed in most clinical breast tumour tissue samples (Braune, *et al.*, 2018; Howe, & Brown, 2004; Jang, *et al.*, 2015). Additionally, overexpression of many Wnt genes, such as *WNT1*, and abnormal expression of Wnt regulators (e.g. soluble Frizzled-related protein, *sFRP1*) have been seen in mouse model breast cancers in the past (Howe, & Brown, 2004). While the primary molecular performers of Wnt signalling have been extensively studied and characterised, many aspects of the pathway remain elusive (Zarkou, *et al.*, 2018).

Additionally, when considering the increased attention received by CSCs in breast cancer, their regulation by the Wnt pathway is extremely promising as a potential therapeutic target (Kazi, *et al.*, 2016). Aberrant Wnt signalling is considered to be important in the regulation of CSC migration and self-renewal.

1.2.2 Cancer Stemness

The epithelial-to-mesenchymal transition (EMT) is a mechanism harnessed in embryonic development and tumour progression. This process encourages epithelial cells to convert to mesenchymal stem cells by losing cell polarity and adhesion, and gaining motility and other invasive characteristics, by which tumour cells can migrate from primary sites and spread via blood vessels and the lymphathic system. In addition to this, EMT induces tumour cells with an invasive capacity to acquire ‘stemness’ traits, much like those seen in normal stem cells; these cells are known as CSCs and have the capacity to self-renew and differentiate into multiple lineages (previously described in Section 1.1.6) (Basu, *et al.*, 2018)

There is sufficient evidence to suggest that Wnt signalling activity involving β-catenin increases tumourigenic potential in breast CSCs; it has been demonstrated by numerous studies that Wnt signalling upregulation has increased breast tumour metastasis, and conversely, inhibiting Wnt signalling suppresses breast cancer metastasis in mice models (Jang, *et al.*, 2015; Chen, *et al.*, 2011). CSCs have become an attractive research area for potential targeting, and the Wnt signalling pathway may be the most fruitful option to attempt to improve clinical outcomes.

However, the mechanisms through which the Wnt pathway contributes to CSC function have not fully been elucidated, and it has been suggested that the most viable starting point would be Wnt target gene investigation, in genes such as *ASCL2, LGR5, MYC, CCND1,* and *CD44* (Kim, *et al.*, 2017). One such target gene, and acknowledged CSC marker is cluster of differentiation 44 (CD44); CD44 is a cell adhesion molecule induced by EMT and has been directly linked with the acquisition of CSC traits. Overexpression of CD44 is associated with advanced stages of breast cancer development and has been directly linked to enhancing Wnt signalling in tumour cells (Basu, *et al.*, 2018; Kim, *et al.*, 2017). Another example of a Wnt target gene is Achaete-scute Complex Like-2 (*ASCL2*), a transcription factor implicated in colorectal cancer. This gene has been shown to work as a transcriptional switch within the Wnt pathway, activating genes that are crucial for the maintenance of stem cell identity and persistence of CSCs (Kim, *et al.*, 2017).

1.2.3 Achaete-scute Complex Like-2 (ASCL2) in Cancer

Achaete-scute Complex Like-2 (*ASCL2*) (11p15.5) belongs to a conserved family of transcription factors containing a basic helix-loop-helix (bHLH) domain, which activates transcription by dimerising through this basic domain and binding to the E-box of target genes (Jubb, *et al.*, 2006). In the normal state, *ASCL2* is involved in the development and maintenance of trophoblasts in the placenta, neuronal precursor determination in both the central and peripheral nervous system, as well as controlling the fate of intestinal crypt stem cells (Hu, *et al.*, 2015; Tian, *et al.*, 2014). Expression of *ASCL2* has been found in the placenta and at the base of small and large intestinal crypts, but is generally low or even undetectable in other normal tissues (Tian, *et al.*, 2014; Zhongfeng, *et al.*, 2018).

ASCL2 is a key intestinal stem cell marker and a known downstream target of Wnt signalling (Figure 1.4), activated by *WiNTRLINC1* (van der Flier, *et al.*, 2009; Giakountis, *et al.*, 2016). The gene is thought to act as a Wnt dependant and responsive transcriptional switch, regulated by an autocrine loop, that defines stem cell (LGR5+) fate in the intestine by acting in cooperation with β-catenin and Tcf (Hu, *et al.*, 2015; Schuijers, *et al.*, 2015). However, recently *ASCL2* has been seen to be involved in tumour progression; due to the loss of imprinting at the 11p15.5 locus (where *ASCL2* resides) being a common occurrence in colorectal cancer (CRC), this prompted the investigation of *ASCL2* in CRC, leading to the findings that this gene is upregulated in colorectal tumours (Jubb, *et al.*, 2006). Since, a number of groups have shown that *ASCL2* overexpression is seen in colon and intestinal tumours (Giakountis, *et al.*, 2016; Jubb, *et al.*, 2006; Schuijers, *et al.*, 2015). Not only has *ASCL2* been found to be upregulated in these tumours, but its activity within the Wnt pathway appears to be disturbed; this disturbance is thought to lead to the overexpression of *ASCL2* in CRC.

As well as expression in primary CRC, Stange, *et al.*, (2010) found that *ASCL2* is a likely signature affecting CRC metastasis to the liver; *ASCL2* overexpression leads to changes in stem/progenitor cell hierarchy (Kim, *et al.*, 2017). Additionally, it is thought that this gene can potentially affect the behaviour of these metastatic tumours by altering the potential of stem and progenitor cells, subsequently causing self-renewal as opposed to differentiation (Tian, *et al.*, 2014). *ASCL2* overexpression has previously been found in a number of other cancers such as lung squamous cell (Hu, *et al.*, 2015), gastric (by induction of EMT) (Kwon, *et al.*, 2013; Zuo, *et al.*, 2018) and osteosarcoma (Liu, *et al.*, 2016), conferring a poor prognosis. A recent study by Juarez, *et al.*, (2018) found that Ivermectin, an antiparasitic agent being explored as an anticancer drug, interacts with the Wnt pathway and supresses *ASCL2*, and has the ability to target CSCs. It was also discovered that cells overexpressing *ASCL2* show a resistance to 5- fluorouracil in gastric cancer, a drug used to treat many solid tumours including breast cancer (Kwon, *et al.*, 2013; Kim, *et al.*, 2017).

In 2014, a study published by Conway, *et al.*, (2014) used DNA methylation profiling and clustering analysis to reveal a group of hypermethylated developmental genes, including *ASCL2*, in hormone receptor positive breast cancers; *ASCL2* was one of many genes showing large differential methylation, reduced expression, and a predictor of poor prognosis in patients. This study showcased evidence of an association between epigenetic profiles such as DNA methylation and breast tumour classification. Other than this, *ASCL2* had not been investigated specifically or thoroughly in breast cancer when the present study began. Since, an *in silico* meta-analysis of the *ASCL* gene family was published in 2017 (Wang, *et al.*, 2017), revealing that *ASCL2* demonstrated significantly increased expression in breast, stomach, head and neck, ovarian and testicular cancers, as well as exhibiting low expression in melanoma, sarcoma, prostate and neurological cancers. In terms of breast cancer specifically, a study by Xu, *et al.*, (2017) found that *ASCL2* was expressed highly in breast cancer cells compared to normal epithelial cells, and expression appeared to correlate with tumour size, growth, and metastasis; the study also suggested that *ASCL2* may be used as a marker to assess the risk of relapse in cancers. In 2018, Wang, *et al.*, (2018) used Gene Ontology (GO) functional enrichment analysis and identified *ASCL2* as one of many differentially expressed genes in BT474 breast cancer cells compared to MCF10A cells.

Despite the study by Xu, *et al.,* (2017) being the first to document the clinical relevance of *ASCL2* in breast tumours, experimental analysis was extremely scarce and relied on immunohistochemical staining alone, with small sample sizes; no elucidation of the function of *ASCL2* was explored. Although there have been other recent studies hinting at the involvement of *ASCL2* in breast cancer, none of these have led an in-depth investigation into the role of *ASCL2* in breast tumourigenesis, or integrated multiple layers of examination.

From the studies mentioned, it is clear that interest in *ASCL2* in breast cancer is peaking in the scientific community and research in this area is extremely prospective, however, functional investigation or further exploration of this gene has not been pursued by any of these studies to date. Still, these studies not only provide further rationale for the selection of *ASCL2* as a candidate gene in this study, but also provide evidence that this area of research is a current field of interest requiring knowledge contribution, and has given some new insight that can be built upon within this project.

Therefore, these factors pave the way for investigation into the function of the *ASCL2* gene in breast cancer, particularly within the Wnt signalling pathway. Wnt signalling is considered to be important in the regulation of CSC migration and self-renewal, and taking into account what is already known about the role of *ASCL2* in development and defining stem cell fate, this may be a link that needs exploring.

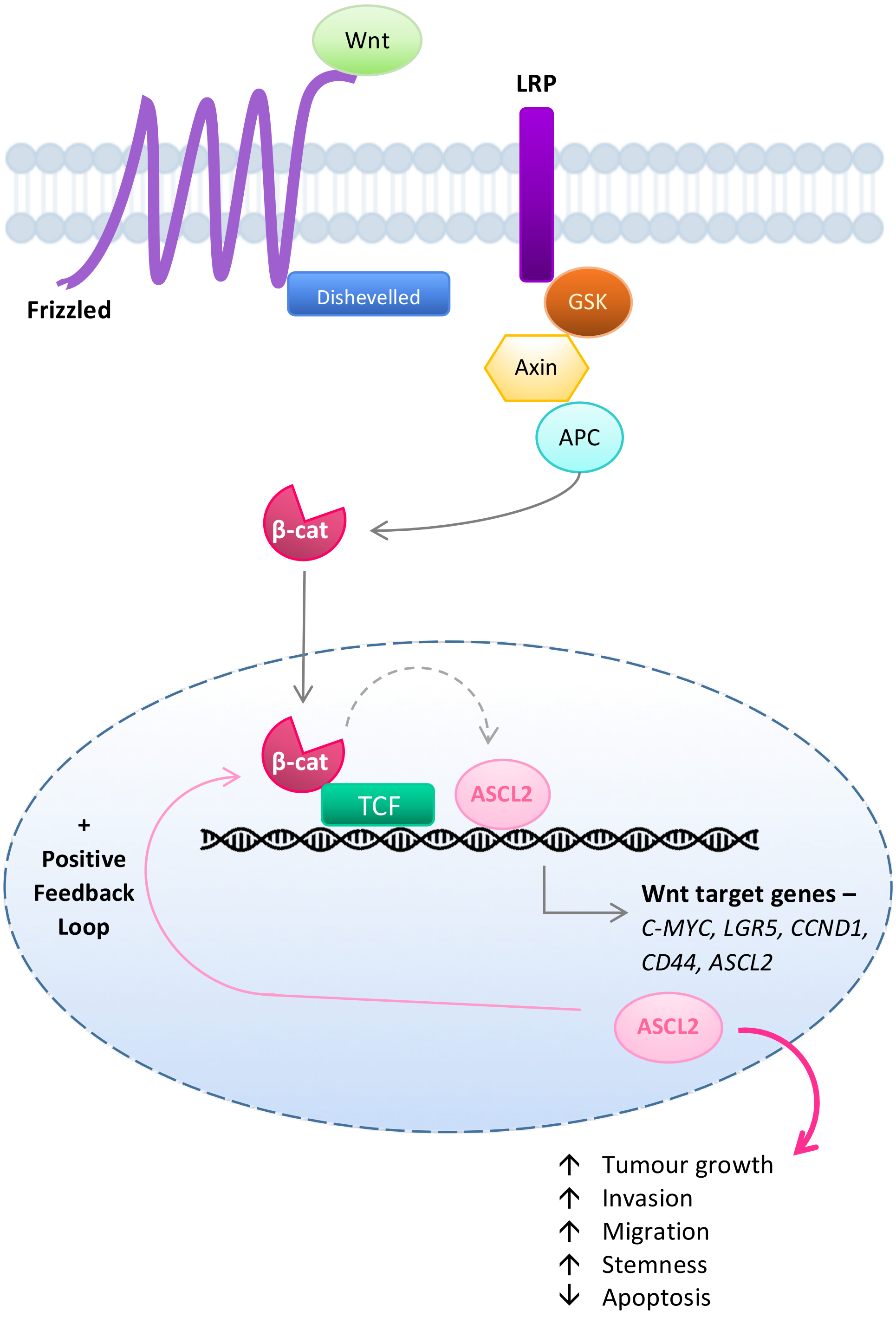


Figure 1.4. Proposed signalling mechanism by which *ASCL2* works within the Wnt signalling pathway to affect the expression of Wnt target genes and advance the growth of cancer. Diagram adapted from Schuijers, *et al.*, (2015) & de Sousa, *et al.*, (2011).

1.3. Cancer Genetics, Biomarkers & Gene Discovery

There has been great progress in identifying many germ-line mutations, such as *BRCA1/2*, which have given the ability to detect susceptibility, predict prognosis and dictate patient stratification. However, the success of these genes is dampened by the fact that hereditary mutations are only responsible for approximately 5% of breast cancers. Therefore, the remaining 95% are sporadic and instigated by an accumulation of somatic mutations (van der Groep, *et al.*, 2011).

Somatic mutations occur in all dividing cells due to exogenous (environmental factors such as radiation) or endogenous (faults in DNA replication) mutagens. These types of mutations are acquired and may be classed as a 'driver' or 'passenger'. Solid tumours typically can contain up to thousands of genetic aberrations and alterations, but only a handful of these are considered driver mutations (Tomasetti, *et al.*, 2015). Driver mutations allow cells a selective growth advantage and are considered positively selected in cancer cells; these alter critical cellular processes leading to the hallmarks of cancer (Gonzalez-Perez, 2016). In contrast, passenger mutations may arise within the cell, but do not give the cell any growth advantage (Greenman, *et al.*, 2007). A driver gene therefore, is a gene containing driver mutations (Tomasetti, *et al.*, 2015).

Vogelstein, *et al.*, (2013) estimated that an average tumour contains two to eight driver gene mutations. These driver mutations are thought to only provide a small growth advantage to cells, which eventually build up over many years and result in billions of additional cells. Hence, it follows that the number of these somatic mutations is correlated to age. In this sense, sequential somatic mutations occurring during tumourigenesis can be thought of as an 'evolutionary clock' (Vogelstein, *et al.*, 2013).

However, despite the exact number of driver gene mutations required for breast tumour initiation and progression being unknown, Tomasetti, *et al*., (2015) have shown that for the development of lung and colon adenocarcinomas, only 3 mutations are needed. This has important implications for driver gene identification highlighting that although there is unlikely to be one single gene responsible, there may only be a small handful which can be taken advantage of for targeting.

Although, hereditary mutations and mutations in driver genes are not the only initiators of cancer development. Epigenetic alterations are also crucial for the pathogenesis of breast cancer. Despite all cells holding essentially the same genetic information, the variation observed between cell types and cellular functions is a result of differences in gene expression. Epigenetics in its broadest and simplest form therefore describes the changes in gene expression and activity that are not encoded by DNA (Gibney, & Nolan, 2010; Byler, *et al.*, 2014). There are a number of ways in which gene expression is controlled via epigenetic mechanisms including DNA methylation, histone modification and microRNA expression (Gibney, & Nolan, 2010). Epigenetic alterations, such as dysregulated microRNAs, can therefore affect the expression of tumour suppressor or oncogenes and result in tumourigenic growth (Byler, *et al.*, 2014).

MicroRNAs are small non-coding RNAs and their involvement in the regulation of gene expression has been established in breast cancer, among other cancers, for some time. Iorio, *et al.*, (2005) revealed that a number of miRNAs were aberrantly expressed in breast cancer compared to normal tissue, namely mir-125b, mir-145, mir-21, and mir-155, with a number of miRNAs being associated with clinical parameters in patients; these include hormone receptor status, stage, vascular invasion and proliferation index. Since, it has been recognised that microRNAs also play a role in treatment resistance (Rodriguez-Barrueco, *et al.*, 2017).

1.3.1 The Importance of Biomarker & Gene Identification

Identifying key genes in breast cancer, as well as in other cancers, is pivotal in revealing crucial information regarding tumour biology, such as which pathways are disturbed during tumourigenesis. By identifying the genes responsible for driving and altering oncogenic signalling pathways, these can be further explored and may be used to gather information on individual tumours during diagnosis in order to enhance clinical decisions. Additionally genes within a pathway can be potentially targeted, or used to predict and tailor response to therapy (Gatza, *et al.*, 2014).

The important implication that driver genes can be targeted for therapeutic development is supported by a study by Rubio-Perez, *et al.*, (2015) highlighting that in 4000 tumour samples across 28 tumour types, only 6% were shown to be manageable using currently approved agents. This highlights a need for greater in-depth genetic characterisation in cancers, particularly for breast cancer patients where current therapies are ineffective; this is the case for many HER2+ cancers, where nearly half of patients exhibit resistance to trastuzumab (Esteva, *et al.*, 2010).

1.3.2 The Current Genetic Landscape of Breast Cancer

It has been established that there are on average approximately 57 somatic mutations per breast cancer case, but only a small number of tumours have overlap in driver mutations, and no tumour can be considered identical at the genome level (Desmedt, *et al.*, 2016; Ng, *et al.*, 2015). It is important to note that depending on laboratory methods, sample selection and data analysis, many studies identify different sets or signatures of somatic gene mutations (Ng, *et al.*, 2015). However, there are still a small set of potential driver genes that are recurrently identified across breast cancer studies, such as *ERBB2, TP53, MYC, PIK3CA, GATA3, CCND1, FGFR1* and *MAP3KI* (Desmedt, *et al.*, 2016). Studies have shown that ER-positive tumours have fewer mutations than ER-negative tumours, which primarily affect *PIK3CA*. Of all of the intrinsic subtypes, HER2+ has been shown to have the highest mutation rates, with the most frequently mutated gene in HER2+ and the basal-like subtype being *TP53* (Ng, *et al.*, 2015).

Presently, it appears that the term ‘driver gene’ can be interpreted in 2 different ways in the scientific literature. At its core, a ‘driver gene’ is used to define a mutation that gives a cell a selective growth advantage. Thus, a driver gene must have a driver mutation. However, some recognised driver genes do not possess a mutation and enhance tumourigenicity via changes in expression by epigenetic alterations. Although both mutated or over/under expressed genes can still drive the neoplastic process and can therefore be regarded as a driver gene, this terminology is vague. Vogelstein, *et al.*, (2013) suggest clarifying this by categorising driver genes as a ‘mut-driver’ or ‘epi-driver’. Epi-drivers can therefore be considered to be aberrantly expressed in tumours but not necessarily mutated (Vogelstein, *et al.*, 2013).

Gatza, *et al.*, (2014) used gene expression microarray data and a panel of gene expression signatures to examine patterns of pathway activity to identify specific DNA amplifications and genes within these that represent key drivers. This study identified 8 genes (*FGD5, METTL6, CPT1A, DTX3, MRPS23, EIF2S2, EIF6* and *SLC2A10*) amplified only in patients with proliferative luminal breast cancers, a subtype with few therapeutic options. Liu, *et al.*, (2015) have also identified candidate driver mutations in the luminal subtype, revealing mutations in *BRAF, GNAS, IDH1* and *KRAS*, by sequencing hotspot regions from cancer related genes.

A study by Lawrence, *et al.*, (2014) revealed 33 novel candidate genes related to the hallmarks of cancer. The authors suggest that a complete catalogue of cancer genes, which would be helpful for precision medicine, is still far from achievable, with the new number of new candidate genes continuously expanding, especially with increasing sample sizes. This would ultimately be useful for clinicians to select the optimum combination of therapies based on the disrupted pathways in each patients' specific tumour. Despite this study, and numerous others identifying large sets of novel candidate genes in breast cancer, few of these candidate genes are functionally investigated to determine their role on tumour growth and progression. It is important that for our knowledge of genes to be useful, follow up studies must fully validate them and explore their potential drugability (Lawrence, *et al.,* 2014).

A comprehensive genomic, transcriptomic and proteomic analysis integrated with clinical data, by Michaut, *et al.*, (2016) has confirmed that PI3K pathway mutations and *CDH1* inactivating mutations are most frequently altered in invasive lobular breast carcinoma. Other mutations in *HER2, MAP3K1*, and *MAP2K4* were revealed at low frequency. As can be seen from these studies, there is seldom complete agreement or overlap of identified driver genes across the different breast cancer subtypes; the variety of subtypes also make it difficult to obtain a generalised picture of the genes present in breast cancer. This further demonstrates the complexity of breast tumorigenesis and the challenge of identifying true driver genes, necessitating further investigation and characterisation.

Other studies have put focus on the immunoglobulin superfamily genes, such as *ALCAM, CXCR4, MUC18* and *L1CAM* (Li, *et al.*, 2016). This novel study investigated the superfamily by integrating different levels of data (genomic, gene expression, protein-protein interactions). Results indicated that the majority of these genes could be considered cancer drivers, or have links to drivers, thus show potential as breast cancer biomarkers.

In 2016, a landmark study used whole genome sequencing and comprehensive bioinformatics analyses to analyse 560 breast cancers (Nik-Zainal, *et al.*, 2016). Ninety-three potential driver genes (protein coding somatic mutations) were identified, along with 12 base substitution and 6 base rearrangement driver gene signatures. To date, this research has been the largest study to attempt to encompass the majority of somatic mutations in breast cancer, and has made efforts to confirm the accepted notion of each breast tumour's genetic profile being individual (Nik-Zainal, *et al.*, 2016). Despite this study gaining a general picture of the driving mutations in breast tumourigenesis, it is likely that subtle or uncommonly mutated genes still need classification. As well as this, further functional analysis and development of this list of 93 genes is needed to improve the clinical utility of this research. Smid, *et al.*, (2016) illustrates the importance of functional analysis of somatic mutations by undertaking a comprehensive genomic and pathway analysis in breast cancer. The group found that the type of genetic substitution has greater impact at triggering an immune response against a tumour, rather than the number of mutations, as was previously thought. Although smaller studies and research groups are often limited on resources to perform such large scale and varied analysis, the trend of integrating multiple platforms can still be feasible on a smaller scale, for example by using only two or three platforms, smaller samples sizes, or by data mining from free public databases rather than performing all laboratory analysis in-house.

The few driver gene mutations present in cancers in comparison to passenger mutations means it is difficult to investigate the function of all mutations identified by sequencing. In light of this, bioinformatics analysis tools have been developed to predict key genes and mutations, which can therefore be preferentially selected for functional analyses. There are two main approaches used in this instance, which either examine the mutation frequencies or aim to predict the functionality of the mutations (Pon, & Marra, 2015). Alternatively, systematic approaches can reveal groups of genes that are functionally related, or genes that are linked by a functional network or significantly enriched signalling pathway (Figure 1.4) (Gonzalez-Perez, 2016).

The molecular markers and gene expression signatures (discussed above) used to currently classify patients' breast cancers, and thus predict prognosis and treatment course, have been used in the private clinical laboratory for some time, for example, OncoType DX by Genomic Health, and the 'MammaPrint 70-gene prognostic signature' by Agendia (Cronin, *et al.*, 2007; Slodkowska, & Ross, 2009). Furthering knowledge of the molecular heterogeneity of breast cancer can expand these clinical tests for greater predictive and informative power which could be crucial in reducing cancer mortality (Vogelstein, *et al.*, 2013)

1.3.3 The Challenges of Gene Identification in Breast Cancer

The ongoing research in this field by academics and industrial laboratories worldwide emulates the challenge of gene identification. Despite sequencing and array technologies moving at a phenomenal pace, this is not matched by 'big data' handling techniques or user-friendly bioinformatics analyses. Furthermore, the number of bioinformatics algorithms, possible analysis pipelines and databases can make it difficult for researchers with little bioinformatics experience. This is further discussed in Section 1.5.

Aside from technical challenges, there are many factors that cause difficulties in gene identification. One important thing to consider is that somatic mutations rarely occur at greater than 10% prevalence, meaning that most genes have a much lower incidence and are mutated infrequently. This is due to the fact that there is such an enormous and assorted range of somatic mutations occurring in cancer cells, that the frequency of any identified driving mutation can be extremely low, even if they provide cells with a significant growth advantage (Tomasetti, *et al.*, 2015; Liu, & Hu, 2014). This can be seen across all breast cancers where only somatic mutations in *TP53, PIK3CA* and *GATA3* appear at >10% incidence (The Cancer Genome Atlas Network, 2012). The overwhelming amount of tumour suppressor or oncogenes, and possible epigenetic changes, manifests a huge number of possible genotypic outcomes per tumour. An additional complication is that driver genes can also contain mutations that are not driver mutations (Vogelstein, *et al.,* 2013).

Despite breast tumours originating from the same mammary tissues, the different subtypes can be considered as molecularly different diseases with differing gene expression profiles and therapeutic responses. It is now accepted that these subtypes do not exhibit identical sets of mutations or gene expression patterns; it is unlikely that each subtype can be represented by a single driving gene or biomarker, although the pathways affected may be similar. In this sense, it seems that no two breast tumours are genetically the same, and thus no two breast cancer cell lines are the same.

There is growing evidence suggesting that many primary breast tumours consist of several genetically distinct clones, rather than existing as a single entity. This inter and intra-tumour heterogeneity (Figure 1.5) has been demonstrated in approximately two-thirds of triple negative breast cancers, particularly in basal-like subtypes, and often means that potential driver somatic mutations are actually only seen in a minority of tumour cells (Ng, *et al.*, 2015). This can also affect secondary tumours, with the majority of metastatic lesions varying significantly in their genetic profile compared to primary breast tumours.

A single tumour consisting of different cell clones is thought to be a result of distinct CSC populations and tumour cell plasticity. There is now evidence that breast cancer cells have the ability to convert between different subtypes of the disease (Yeo, & Guan, 2017); a study by Jordan, *et al.*, (2016) revealed that circulating tumour cells demonstrated reversible HER2 expression plasticity. In light of this, it may be advantageous in the future to utilize multiple subtype-specific therapeutic agents together to lessen the chance of resistant populations from remaining (Yeo, & Guan, 2017).

Regarding inter- and intra-tumoural heterogeneity, there is currently a disconnect between scientific research and clinical practice (Yeo, & Guan, 2017). Progress within research has led to the identification of 10 distinct integrated clusters (discussed previously) to segregate breast tumours, however these, as well as PAM50 genomic testing (a 50 gene prognostic subtype classifier) (Parker, *et al.*, 2009) have not yet been employed in the clinical setting. Owing to intra-tumoural heterogeneity, diagnosis at the cellular level instead of the tumour as a single entity would be more beneficial (Yeo, & Guan, 2017). This is where single-cell technologies would be favoured in diagnosis, rather than an isolated biopsy of a section of the tumour which may not be reflective of the entire mass.

Additionally, the diversity of somatic alterations found in breast tumours can change over time as the tumour develops (Desmedt, *et al.*, 2016). This presents a challenge to researchers, as driving mutations found in single breast cancer samples may not be characteristic of the whole tumour. Although these complex factors pose challenges for gene identification in breast cancer, they may be useful for more long-term development and implementation of targeted medicine; this would ensure that future therapeutics would be based more upon the molecular biology of individual tumours (Ng, *et al.*, 2015).

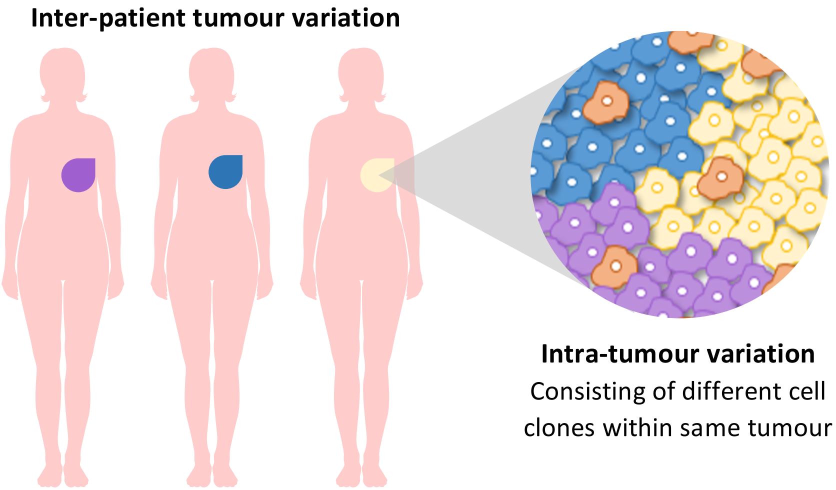

Inter-patient tumour variation
Intra-tumour variation
Consisting of different cell clones within same tumour

Figure 1.5. Inter- and intra- tumour heterogeneity, a prominent challenge in breast cancer research.

1.3.4 Personalised Medicine

When taking into account the complexity, variability and unpredictability of tumourigenesis, precision medicine has been the goal of cancer research for a number of years, and has been used for some time, e.g. ER and HER2 status. The aim of precision medicine is to provide a level of care for the patient that will yield the best clinical outcome and avoid adverse reactions, based on the patients' influential profile of genetic variants and factors. Each tumour will exhibit different genetic and epigenetic variations, which in turn is responsible for different manifestations, behaviour, prognosis, and response to therapeutic agents. This goal is becoming seemingly more attainable due to the advancements in technology and research discussed in this chapter, allowing a patient's particular tumour to be characterised at the clinic and then administered more specific and targeted drugs with greater therapeutic impact (Uzilov, *et al.*, 2016). As these targeted drugs have proved effective at treating patients with these biomarkers, efforts continue to find more of these links to be exploited, but, identification and drug development can be extremely complex (Uzilov, *et al.*, 2016).

There are many targeted drugs for a variety of cancers, including breast cancer, which have been FDA approved or are awaiting approval. Some of these rely on an increased expression of cell surface receptors (characterised by immunohistochemistry) whereas others rely on genetic biomarkers. An example is Lapatinib, a tyrosine kinase inhibitor for HER2 overexpression in breast tumours, which received full FDA approval in 2010 (Mcveigh, & George, 2017; National Cancer Institute, 2011); A more widely used example is Herceptin (Mcveigh, & George, 2017). However, it is important to note that the idea of precision medicine does not mean individual drugs custom-made for each patient. Instead, drugs and treatment options are given based on a patients' genetic profile and predicted response.

A number of large studies (including those previously mentioned in this chapter) have worked towards using Next Generation Sequencing (NGS) for genetic testing to build a well-defined landscape of possible mutations in breast cancer to increase the capacity of personalised care. Uzilov, *et al.*, (2016) have described an integrative genomic approach to aid in the clinical application of precision medicine. The study performed whole exome sequencing (WES), targeted sequencing, panel small nucleotide (SNP) microarray genotyping, and RNA sequencing on matched tumour and normal samples. The results demonstrated that using WES highlighted a greater and more complete ‘spectrum’ of alterations in comparison to the targeted panel sequencing. Whole exome sequencing also allowed investigation of variants involved in drug metabolism processes.

Overall, the integrated approach of this study gave greater clarity and ‘more actionable alterations than several commercially available targeted cancer panels’, as well as providing a clinically applicable workflow. Using WES and RNA sequencing in this clinical study gave a more inclusive genetic profile for patient samples, which could be used to make therapeutic decisions, and identified more rare somatic mutations in some patients that were initially missed by targeted panel sequencing.

There are many advantages to this type of investigation for patients. Primarily, it allows examination at the pathway level, of whether drivers or multiple components are altered consistently in the same pathways. Also, cancer panels only consist of already well established, characterised and common mutations; rarer mutations or those that are functionally documented but not yet considered as a driver may be missed, and may have held important clues for treatment determination. Additionally, this type of screening allows more accurate differentiation of germline and somatic mutations (and also between tumour and germline DNA) which can affect the patient and family. Making this distinction is important as it can highlight any possible alterations in DNA repair pathways that will determine chemotherapy response and thus dictate the dosage and drug toxicity administered. These responses cannot be seen with standard panel testing (Uzilov, *et al.*, 2016).

These types of studies show that NGS and integrated omic approaches are extremely effective and more reliable at finding ‘actionable’ genetic variants in cancer, and should where possible, be incorporated into clinical testing in comparison to techniques with much lower resolution and discovery power (Garraway, & Baselga, 2012). With these techniques being so readily available, it is now more important than ever to strive for continued identification and classification of genes and pathways in breast cancer, so that these can be used for diagnostic purposes and to tailor treatment approaches.

1.4. High-Throughput Technologies

1.4.1 Next-Generation Sequencing

The development of DNA sequencing by chain termination and fragmentation methods was first established by Sanger and colleagues in 1977 (Sanger, *et al.*, 1977), and is known today as first generation Sanger sequencing. This technique was time consuming and required radioactive material, however was automated in 1987 using capillary electrophoresis and fluorochromes (Liu, *et al.*, 2012). This became the predominant method of DNA sequencing, but despite being considered as the gold standard, still had many limitations. Firstly, the technique requires gels or polymers to separate the DNA fragments that are fluorescently labelled. The technique has low throughput so only a few samples can be analysed in parallel; sample preparation needs to be manual as automation of these steps is difficult. Cloning of the DNA fragments into bacteria is needed to produce larger sequences, sensitivity is also low for the detection of low-level mutations, even those considered to be clinically relevant, and assembly of whole genomes *de novo* is challenging (Fakruddin, & Chowdhury, 2012).

Finally in 2004, the first draft of the human genome was successfully sequenced and published in its entirety (International Human Genome Sequencing Consortium, 2004). From this point onwards there was a call for more rapid and cheaper technology that addressed the limitations associated with Sanger sequencing. In 2005, the pyrosequencing method by 454 Sequencing™ (Life Technologies, Roche) was the first next-generation sequencing (NGS) technology to be released. By 2010, the founder of the 454 developed and released the Ion Torrent Personal Genome Machine (PGM™) (Thermo Fisher, Life Technologies) which resulted in faster, cheaper and more user friendly sequencing, making the cost of sequencing more accessible to independent laboratories; the target of the $1000 genome was reached in 2014 (van Dijk, *et al.*, 2014).

In the past decade, or the ‘omic era’, massively parallel, high-throughput DNA sequencing platforms have become a mainstream preference used across academic and industrial laboratories and has transformed genomic and transcriptomic research. The advent of bench-top NGS platforms has improved time consumption during sequencing, making the technology much more translatable to clinical use (Hansen, & Bedard, 2013).

To utilise this potential offered by such significant shifts in the capability of technology, the UK's landmark 100,000 Genomes Project was initiated in 2013 by Genomics England. The project aimed to sequence 100,000 whole genomes from the UK National Health Service (NHS) patients by 2017. Ultimately, the goal of this large-scale and ambitious project was to transform clinical practice using genetic information for rare diseases and cancers, and to drive change such that WGS is established and adopted as part of routine assessment and care (Turnbull *et al.,* 2018).

1.4.2 Whole Exome Sequencing

Depending on the requirements of sequencing, it may be practical or cost-effective to carry out whole genome sequencing (WGS). Alternatively, a more distinct or targeted region of the genome may be required. In this instance, and is the case in the majority of clinical studies, it may be beneficial to use whole exome sequencing (WES) which only sequences the protein coding region (1-2%) of the genome, but is still able to identify a wealth of single nucleotide polymorphisms (SNPs) and small insertions and deletions (INDELS). Moreover, almost 85% of potential disease causing aberrations are incorporated into this minority (van Dijk, *et al.*, 2014).

One disadvantage of WES is that it doesn't investigate the impact that non-coding alleles might have on diseases. WES is sometimes criticised as sequences other than exons can also be very important in disease, e.g. non-coding RNA such as microRNA, and the controversially termed 'junk-DNA'. A particular challenge facing WES is defining the exome and which sequences are rightly protein coding in the human genome, as our knowledge of protein coding exons is currently unfinished. Although currently, WES is a very economic and prolific option for gaining a vast amount of information at a reasonable cost (Bamshad, *et al.*, 2011).

1.4.3 Microarrays

Prior to NGS for cancer genomics was the production and utilisation of DNA microarrays, which continue to be used for gene expression (mRNA) analyses, especially in cancer research to gather information on potential transcriptomic targets and markers. Gene expression microarrays offer a broad view of the entire transcriptional activity in a sample (Slonim, & Yanai, 2009). Other applications include assessment of gain or loss of genetic material, DNA aberration profiling to identify cancer causing genes, SNP arrays for germline mutation identification, the study of protein expression, and DNA methylation and microRNA expression analysis (Malone, & Oliver, 2011).

A microarray is the hybridisation (singularly or multiplex) of labelled samples of interest, alongside a complementary nucleic acid probe bound to a miniaturised silicon thin-film chip or glass slide (Figure 1.6). A microarray chip is made up of thousands of individual nucleic acid probes (these can be complementary DNA or known oligonucleotides) which match to a short piece of nucleic acid sequence from a human (or other known organism). Ultimately, all of the probes combined provide a genome-wide view of all coding regions (Malone, & Oliver, 2011). One such commercial array platform is manufactured by Affymetrix (www.affymetrix.com). Affymetrix GeneChips are the most commonly used microarray platforms for expression profiling, able to interrogate the whole genome for transcription activity (Auer, *et al.*, 2009).

This technology can be performed relatively cheaply (in comparison to NGS) in a massively parallel manner, yielding results for thousands of genes in one experiment. The Minimum Information About a Microarray Experiment (MIAME) standards have also been introduced ensuring consistent and reliable results across laboratories. However, microarrays are limited in that they tend to focus on more common variants and require knowledge of the sequence prior to analysis for primers to be designed, as opposed to RNA sequencing, that requires no *a priori* knowledge (Brazma, *et al.*, 2001; Buermans, & den Dunnen, 2014; Christie, 2005).

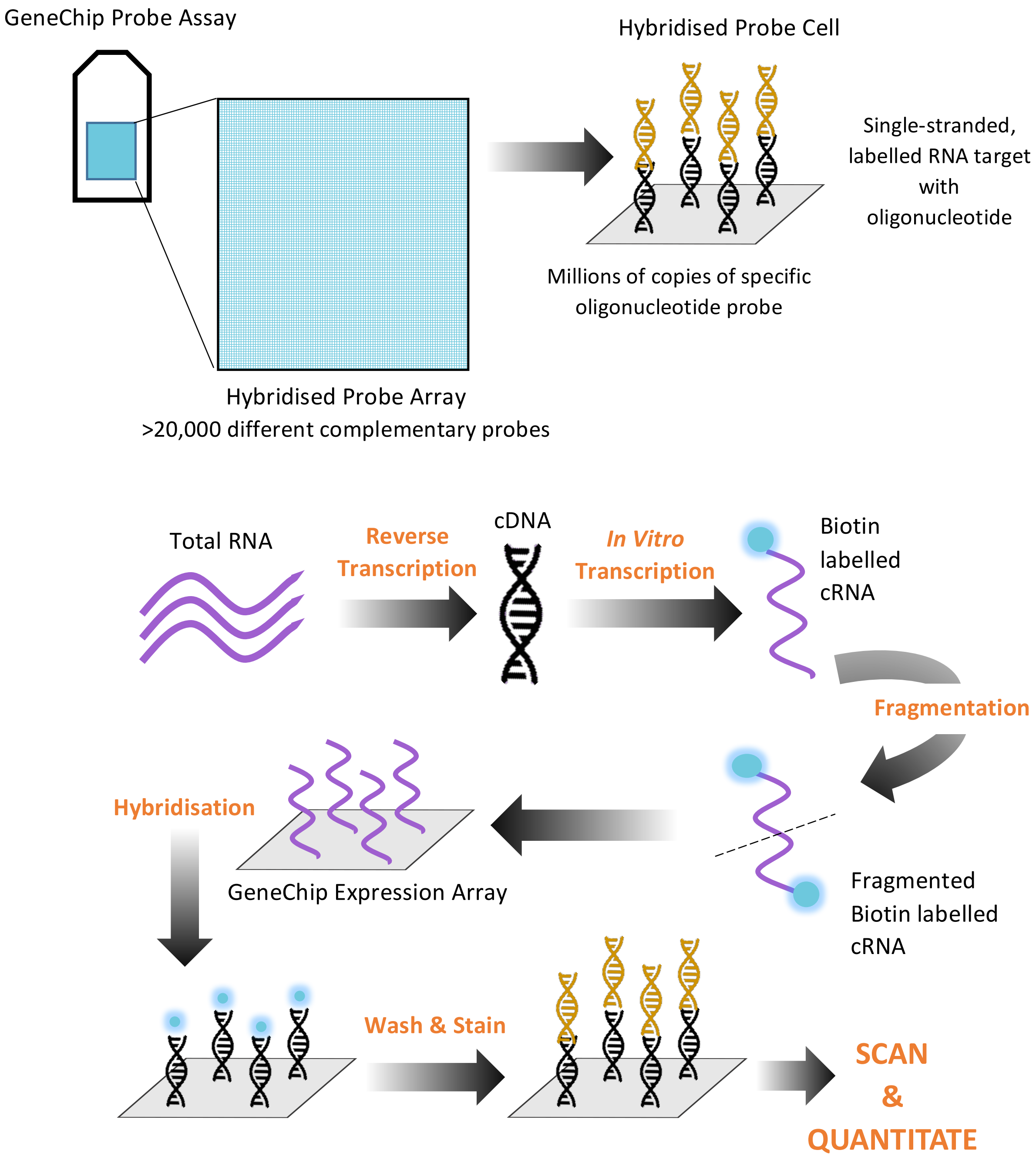

GeneChip Probe Assay
Hybridised Probe Cell
Single-stranded, labelled RNA target with oligonucleotide
Millions of copies of specific oligonucleotide probe
Hybridised Probe Array
>20,000 different complementary probes
Total RNA
Reverse Transcription
cDNA
In Vitro Transcription
Biotin labelled cRNA
Fragmentation
Hybridisation
GeneChip Expression Array
Fragmented Biotin labelled cRNA
Wash & Stain
SCAN
&
QUANTITATE

Figure 1.6. An Affymetrix gene chip, a popular commercial microarray, showing the process of hybridisation of labelled RNA probes to a gene chip expression array in order to quantitate RNA expression of selected targets (Figure adapted from Bumgarner, 2013, and Macgregor, & Squire, 2002).

1.4.4 Data Mining

There is an unprecedented amount of genomic and transcriptomic data being generated worldwide that is freely available for researchers to use; the act of extracting novel information from large datasets is known as data mining. There are many purposes of biological data mining, such as association analysis (to evaluate relationships in data), pathway and network analysis, gene prioritisation, function prediction, pharmacological predictions and toxicology, all of which provides researchers with great prospects for discovery (Gonzalez, *et al.*, 2016).

Since the advent of microarrays and NGS, the number and capacity of sequence databases have expanded. The Gene Expression Omnibus (GEO, NCBI) is the principal and most established public gene expression data repository, including microarray, genomic and proteomic experiments for a variety of organisms. The goal of GEO was to 'provide a robust, versatile database in which to efficiently store high-throughput functional genomic data', and currently holds 3848 separate datasets of nearly 2 million samples (Barrett, *et al.*, 2005).

Over the years, the explosion in biological data generation has resulted in the emergence of many other gene expression databases allowing researchers to successfully share genomic data or mine relevant datasets to advance their studies and thus, the global knowledge of cancer (Table 1.4). Data from multiple cell lines, tissues, clinical samples and matched normal samples are available. Although this international data sharing and mining is extremely useful for smaller laboratories lacking the resources for WGS or microarray analysis, there are also many drawbacks that researchers must be aware of prior to analysis.

Firstly, variability between sources is the most prominent difficulty with data mining – sample collection and processing, raw data processing and data conversion or formatting can all result in discrepancies. Regarding microarray databases, a variability can be seen between data using different microarray platforms, therefore researchers must take this into consideration if using multiple different databases. Also as there is likely to be a large number of genes corresponding to a relatively small sample size, this kind of data is at risk of being prone to false positives, hence there may be issues extracting relevant alterations within the data (Piatetsky-Shapiro, & Tamayo, 2003). In general, large data collections are intrinsically prone to errors, so this should be kept in mind by the researcher (Werner, *et al.*, 2014). However, despite these pitfalls, data mining is an excellent tool for validation purposes, can provide an excellent foundation for answering research questions or formulating hypotheses, and is an economical way of strengthening research conclusions.

Table 1.4. Examples of prominent cancer related open genomics databases (adapted from Yang, *et al.*, (2015b)

Database	Description	URL
CGHub	Sharing of 42 cancer types and normal controls	www.cghub.ucsc.edu/
COSMIC	Largest database of somatic mutations in cancer	http://cancer.sanger.ac.uk
cBioPortal	Multidimensional cancer genomic data, also supporting pathway exploration	http://www.cbioportal.org/public-portal/
Gene Expression Omnibus	Genomics data repository for array and sequenced based data	https://www.ncbi.nlm.nih.gov/geo/
UCSC Cancer Genomics Browser	Online analysis tool for cancer genomics and clinical data	https://genome-cancer.soe.ucsc.edu/

1.5. Systems Biology, Bioinformatics & Data Integration

Since high-throughput nucleotide sequencing and microarrays are much more accessible to researchers, scientists are now flooded with an unprecedented amount of omics data on a daily basis. These advances have put pressure on developing more efficient bioinformatics tools to handle the increasing volume of biological data, and the challenge of extracting relevant information.

In spite of developments in high-throughput technologies making omics data acquisition and generation relatively straightforward, the potential of this data to add to our knowledge of breast cancer is limited by the difficulty in interpretation. For example, just one run of WES can provide a researcher with a large amount (terabytes) of raw data; larger studies with multiple samples can generate a vast amount of data that can be extremely time-consuming to analyse. Hence, analysing these results and pinpointing the most pertinent information in cancer research can be a mammoth task (Pabinger, *et al.*, 2014).

1.5.1 A Bioinformatics Workflow

In order to obtain meaningful results from raw sequencing or microarray data, data management needs to be structured and the correct tools need to be used to match the goal of the research. There are a large number of bioinformatics analysis tools and programs (aside from commercial tools) now available for various purposes, or sections of workflows, with each potentially requiring different operating systems and data formats. Therefore, it is imperative to select analysis programs compatible with the data formats generated from each platform, and that are compatible with the operating systems preferred by the user. The appropriate analysis tools for the desired workflows should be carefully considered for the selected application, and time should be taken to meticulously formulate a suitable workflow prior to beginning analysis (Pabinger, *et al.*, 2014).

Regarding bioinformatics analytical pipelines and workflows, existing systems published in the literature or commercially available can be used by the researcher; various individual platforms may be used simultaneously and compared, or users can formulate their own algorithms which commonly involve building R scripts. Tokheim, *et al.*, (2016) examined various different ‘driver gene prediction methods’ and concluded that each method varied a large amount in its predictions; this poses problems for comparisons between tools and deciding which tools are most likely to give the a realistic picture. As there is no current set mode of analysis, ‘gold standard’ or leading pipeline, scientists can tailor NGS and microarray analysis to their own preferences or expertise (Pabinger, *et al.*, 2014). This is extremely convenient and allows researchers flexibility, but creates a huge amount of variation and disparities between similar studies – ultimately, this has led to a ‘snowball’ effect where unique lists upon lists of distinct candidate genes have been identified by groups around the world, but many genes have not been followed up.

1.5.2 Pathway Analysis

A frequently used approach for analysing genomic or transcriptomic data for the identification of genes in cancer is pathway analysis, discussed in more detail in Section 3.1. This allows an understanding of the functional biology underpinning a set of differentially expressed genes. This approach sorts and condenses a comprehensive gene list into smaller sets of related genes that fit into similar pathways, and reduces thousands of genes into hundreds of components with functional consequences. There are many established and user friendly web-based pathway repositories and resources such as Gene Ontology, Kyoto Encyclopaedia of Genes and Genomes (KEGG), Database for Annotation, Visualization and Integrated Discovery (DAVID) and PANTHER, which use ‘knowledge base driven pathway analysis’ (Khatri, *et al.*, 2012). Another popular method for understanding gene expression data is Gene Set Enrichment Analysis (GSEA). These methods all use similar principles for pathway analysis, by analysing gene sets based on their corresponding biological pathways. These examples and other popular bioinformatics tools are shown in Table 1.5.

Table 1.5. Examples of popular bioinformatics and pathway analysis and enrichment tools. Information collated from Omictools (2017).

Name	Description	Interface/ Programming	Level	Reference
Bioconductor	Open source, for high throughput omic data	R, command line interface	Advanced	Gentleman *et al.,* 2004
CHASM	Prediction of SNV contribution to tumour growth	C++, Python	Advanced	Wong *et al.,* 2011
DAVID	Functional interpretation of lists of genes	Web interface	Basic	Huang *et al.,* 2007
Gene Ontology (GO)	Classification of genes	Web interface	Basic	Ashburner *et al.,* 2000
GSEA	Determines statistically significant groups of related genes	Java, R, Command line interface	Advanced	Subramanian *et al.,* 2005
IMPaLA	Pathway analysis of transcriptomics by enrichment analysis	Web interface	Basic	Kamburov *et al.,* 2011
MADGiC	Prioritises somatic mutations based on frequency and functional impact	R, command line interface	Advanced	Korthauer and Kendziorski, 2015
MutsigCV	Analyses mutation lists to see which ones are mutated more than expected	Command line interface	Advanced	Lawrence *et al.,* 2013
OncodriveFM	Driver gene identification by functional impact	Perl, command line interface	Advanced	Gonzalez-Perez and Lopez-Bigas, 2012
PANTHER	Protein functional classification	Web interface	Basic	Thomas *et al.,* 2003
Pathway Commons 2	Pathway queries and annotation	Web portal	Basic	Cerami *et al.,* 2011

1.5.3 Using a Multi-Platform Approach for Gene Identification

The advent and progression of sequencing and array technologies over the past few decades has revealed a vast amount of information regarding breast cancer, and provides researchers with an unmatched ability to continuously identify genetic alterations driving the oncogenic process. Gene expression and DNA microarrays have steered the way for understanding the heterogeneity of breast oncogenesis, suggesting that the behaviour of an individual's cancer is based on the tumour's genetic profile and pattern of gene expression (Perou, & Børresen-Dale, 2011). In addition, NGS technology has superseded traditional Sanger sequencing, and is evolving rapidly into widespread use across research and clinical laboratories for cancer surveillance, allowing researchers to sequence whole genomes in parallel.

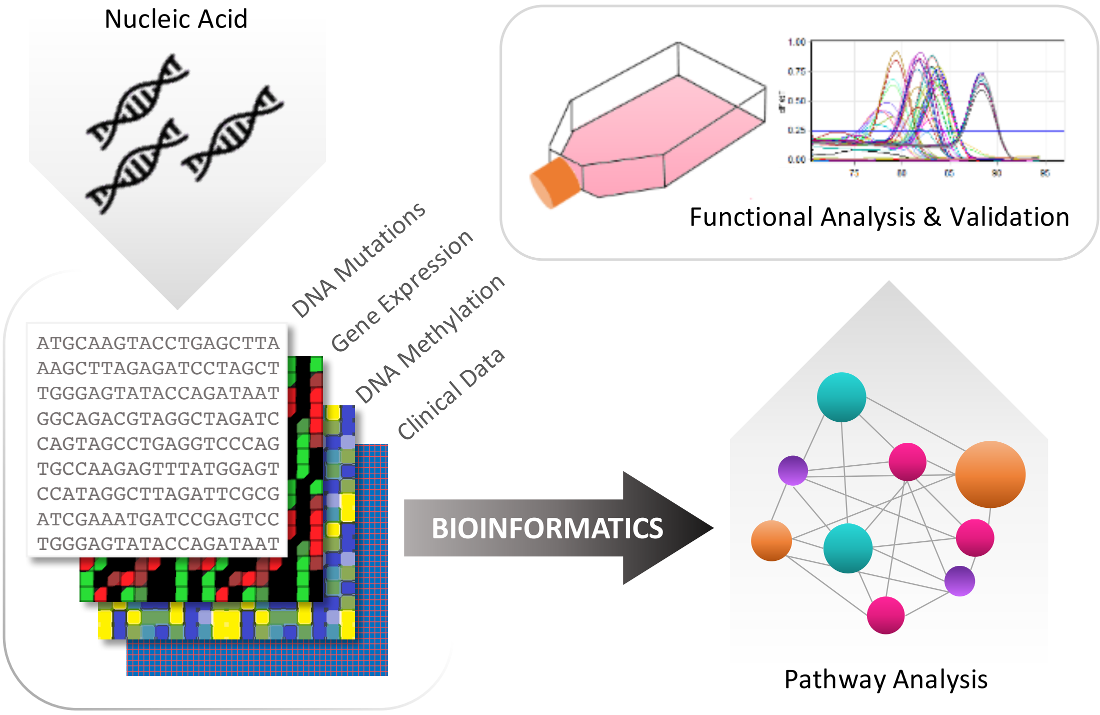


Figure 1.7. An example of a typical workflow for gene identification using an integrated and multi-platform data approach. Bioinformatics analysis can be extremely varied between studies and is dependent on available resources and the expertise of the researcher. Image based on information from Mo *et al.*, (2013) and Suo *et al.*, (2015).

With these technologies being increasingly used for gene investigation, alongside the notion that data integration can reveal more information than singular analysis alone, a multi-level approach for integrating different types of omic data or pathway analysis has risen to the forefront of research, with a large number of studies coupling gene expression and genomic data (Suo, *et al.*, 2015) (Figure 1.7). An extensive and comprehensive example of a multi-level approach is the use of five data types by The Cancer Genome Atlas Network (2012). This study used genomic DNA copy number arrays, exome sequencing, DNA methylation, mRNA arrays, microRNA sequencing and reverse phase protein arrays, and integrated data across these platforms to analyse primary breast cancers. This integrated analysis provided confirmation of previously known somatic mutations, as well as novel subtype associated mutations, *GATA3, PIK3CA* and *MAP3KI* in the Luminal A subtype (The Cancer Genome Atlas Network, 2012). This type of analysis appears to be the most productive in gaining an in-depth view of underlying tumour biology.

1.5.4. Future Directions in 'Big Data' and 'Omics' technologies

Third generation sequencing (TGS) is now emerging with further improvements to NGS; the main factors distinguishing TGS from NGS is the lack of PCR steps during sample preparation which will decrease time consumption and reduce errors arising, and the measurement of signal in real time which can be helpful for structural variance predication. These newer technologies, such as Nanopore and PacBio, which sequences based on an electric current, aim to increase read length and turnover (Liu, *et al.*, 2012).

This greater turnover will continue to produce large quantities of data, therefore systems biology and bioinformatics will play a vital role in the search for cancer marker genes. In the past, the ability to interpret, analyse and integrate data has greatly fallen behind the ability to generate high quality sequence data. As therapeutic resistance and low response rates are prominent causes of therapy failure, systems biology approaches can aid in selecting which patients are most likely to benefit from specific treatments, thus enhance efficacy and potentially reduce emerging resistance, as patients in the future will not be wrongly or over treated (Werner, *et al.*, 2014).

In the clinical setting, molecular characterisation of solid tumour samples to inform clinical decisions is well established (Harbeck, *et al.*, 2014; Russnes, *et al.*, 2017). However, accessing samples from patient's tumours can be extremely invasive. More recently, a more easily obtainable source of tumour genetic material is via circulating tumour cells (CTCs) which can shed into the vasculature and subsequently the bloodstream from many primary tumours, for example breast tumours (Yee, *et al.*, 2016). These cells can be collected by a traditional blood test, and sequential or multiple samples can be collected over time. However, sequencing the genetic material of CTCs in blood using NGS can be problematic due to low numbers and impure samples (Yee, *et al.*, 2016).

To address these problems, whole genome amplification can be used to amplify the amount of CTC generic material prior to sequencing from just a few or single cells. Studies are continuing to improve purification of CTCs from blood for use of NGS in the clinical setting (Yee, *et al.*, 2016). Other alternatives investigate the circulation of free nucleic acids and exosomes (Friel, *et al.*, 2010). In order for the exceptional capacity and potential for these novel technologies to be utilised in the clinic, genes need to be established and characterised to aid diagnosis, prognosis prediction, and eventually direct targeted and personalised breast cancer treatment.

1.6. Gene Investigation in the Laboratory: Harnessing Bioinformatics for In Vitro Investigation

Although bioinformatics investigation, systems biology methods, and ‘big data’ analysis have soared in popularity over the last few years, these methods are far from superseding traditional wet lab investigation. In fact, now it is imperative to incorporate *in silico* investigation with laboratory experiments to gain a clearer and more accurate picture of biological processes; researchers must question whether conclusions drawn from computational investigation can be replicated in cell lines, tissues or even patients.

For gene investigation in cancer, many studies rely on computational biology to generate lists of candidate genes from cell line or tissue data, that may be mutated, differentially expressed, or function within specific biological pathways. From here, researchers may pursue a number of avenues, however arguably the most valuable would be to functionally validate and investigate these genes using conventional and reliable *in vitro* and *in vivo* laboratory techniques. Thus, the knowledge of said candidate genes would be more successful in being developed in the future as diagnostic, prognostic, or therapeutic markers. In this study, *in vitro* validation has taken place using cell lines, and gene investigation was carried out using RNA interference, therefore these methods will be the focus of this section.

1.6.1 Cell Lines as Models

Research based on immortalised cell lines has been established since the 1950s, when HeLa cervical cancer cells were first cloned (Puck, & Marcus, 1955). Now, gene exploration using cancer cell lines is still common practice. In fact, an extensive amount of foundational and novel knowledge on cancer is the result of biological research using cell lines.

Despite this, cancer cell lines are continuously being scrutinised about whether they are representative of the tumours from which they were propagated, and the translatability of such research findings to the clinic. Regarding breast cancer cells, it’s important to be aware that cell lines may not always mirror the inter- and intra-tumour heterogeneity seen in patient tumours, lack the complexity of the tumour microenvironment, and the stability of gene expression in cell lines can be unpredictable (Choi, *et al.*, 2014); studies have demonstrated transcriptomic drift with prolonged cell culture (Gillet, *et al.*, 2013; Ross, & Perou, 2001). Other issues such as cross-contamination and misidentification of cells should also be kept in mind, and therefore, to maintain good practice laboratories should perform regular cell line authentication (Gillet, *et al.*, 2013).

Nevertheless, cancer cell lines are considered to be an acceptable experimental model for tumours and a basis for screening the efficacy and testing of new therapeutics, as well as testing new hypotheses and novel research. If handled correctly, chosen with consideration, and good practice is maintained, overall, cell lines do well to reflect the behaviour of tumours for initial studies, if their limitations are kept in mind (Katt, *et al.*, 2016).

In relation to gene investigation, cell lines are a prime model for examining novel genes as they require little ethical permission for gene manipulation (for example, gene knockdown or knockout studies). They are relatively diverse, cheap and easy to acquire (multiple cell lines can be compared at low cost), and experiments can be performed flexibly and in high-throughput if needs be (Katt, *et al.*, 2016).

1.6.2 RNA Interference

RNA interference (RNAi) is an innate regulatory biological process in cells that results in sequence-specific gene silencing (Gavrilov, & Saltzman, 2012). The silencing of target genes is mediated by non-coding small interfering RNAs (siRNAs) and is shown in Figure 1.8.

In summary, long double-stranded RNA (dsRNA) molecules are cut and separated into 21-25 nucleotide siRNAs by the ribonuclease enzyme Dicer. RNA-binding protein TRBP together with Dicer, loads this siRNA duplex onto the Argonaute protein (AGO2), to create the RNA-induced silencing complex (RISC). Argonaute selects the siRNA guide strand, cleaves and removes the passenger strand. The guide strand remains tethered to AGO2 and couples with complementary mRNA targets which are long enough to be divided, sliced and released. Once this process is complete, the RISC is recycled and uses the same guide strand to repeat this cycle multiple times (Gavrilov, & Saltzman, 2012)

Since the discovery of RNAi was published in 1998, its research potential for gene suppression, manipulation and regulation has been extremely valuable, especially in the field of oncology research (Fire, *et al.*, 1998; Gavrilov, & Saltzman, 2012). Now, using synthetic siRNAs and hijacking this innate RNAi pathway for artificial gene knockdown is almost common practice in research laboratories (Gavrilov, & Saltzman, 2012). This allows researchers to study the effects of genes on cancer growth once expression is switched off, using a number of assays or methods that focus on factors such as proliferation, apoptosis and migration.

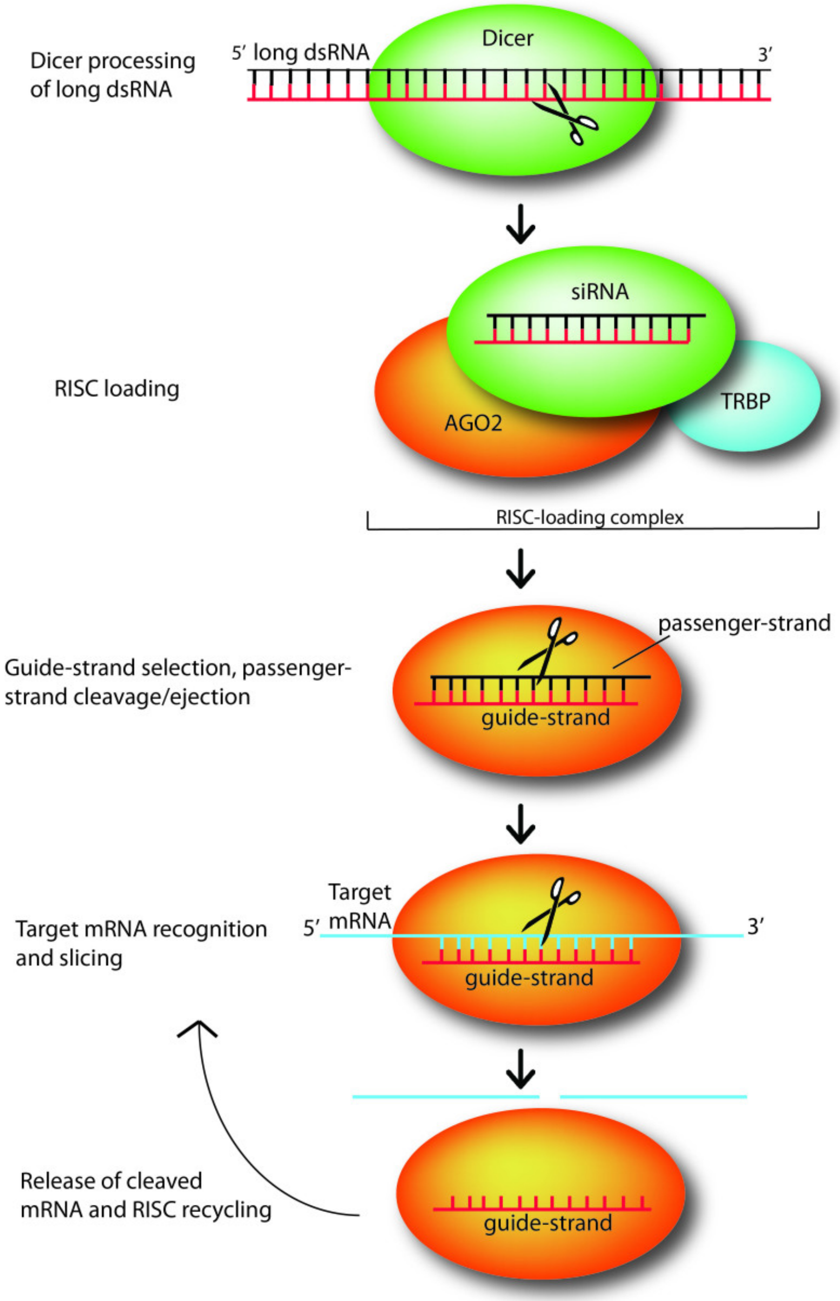

Dicer processing of long dsRNA
5′ long dsRNA
Dicer
3′
RISC loading
siRNA
AGO2
TRBP
RISC-loading complex
Guide-strand selection, passenger-strand cleavage/ejection
passenger-strand
guide-strand
Target mRNA recognition and slicing
Target mRNA
5′
3′
guide-strand
Release of cleaved mRNA and RISC recycling
guide-strand

Figure 1.8. The mechanism of gene silencing by RNA interference (Gavrilov, & Saltzman, 2012). Reproduced under CC BY-NC terms.

1.7. Overall Conclusions & Aims of this Study

There is an increasing need for greater characterisation of genes across the distinct breast cancer subtypes to understand their mechanistic role in tumourigenesis within their respective pathways. These may be used in the future to aid in the implementation of precision medicine, such as for therapeutic exploration or biomarker discovery, or to develop an extensive catalogue of genes for which tailored treatment can be built around, addressing the current inefficacy of existing breast cancer therapies.

A multi-method approach for gene identification is a highly strategic, prolific and lucrative method for enhancing our knowledge of the molecular foundation underpinning breast carcinogenesis, whilst identifying attractive potential targets. However, in order to exploit the information generated from sequencing studies, the future must focus on addressing the challenges associated with gene identification and large-scale omic data, to develop more robust and user-friendly bioinformatics pipelines for processing.

The candidate gene of interest in this study, *ASCL2*, is a transcription factor known to be involved in precursor determination during embryonic and nervous system development. This gene has been previously associated with tumour progression in colon cancers, proposed as a target of Wnt signalling influencing the fate of intestinal stem cells; in a number of other cancers, *ASCL2* has also been related to poor prognosis. However, despite considerable research into the role of *ASCL2* in colon cancer, as well as other cancers, its role in breast cancer is yet to be defined.

The work presented in this thesis aimed to:

1. Integrate multiple bioinformatics pathway analysis tools to select a novel oncogenic candidate gene (*ASCL2*) from transcriptomic data, for further investigation in breast cancer.

2. Showcase and implement a simple yet integrated *in silico* analysis pipeline, to aid both expert and non-expert researchers with gene identification from transcriptomic data.

3. Investigate the role of the *ASCL2* candidate gene in breast cancer cell lines, using RNAi and multiple functional assays, to improve understanding of its function and shed light on its role in breast cancer tumourigenesis.

4. Investigate the relationship between *ASCL2* gene expression and clinicopathologic features in patient breast tumours, evaluating the suitability of *ASCL2* as a possible prognostic indicator in breast cancer.

Chapter II

Methods

2.1 Transcriptomics & Bioinformatics

2.1.1 Gene Prioritisation Using 'Extreme Variation' Analysis

Publically available Affymetrix microarray (U133 Plus 2.0 Chip) data (mRNA gene expression profiles) for the human breast cancer cell lines, MCF7, T47D, BT474, MDA-MB-231, and the non-tumourigenic epithelial cell line, MCF10A, were downloaded as raw .CEL files from the online repositories Array Express (www.ebi.ac.uk/arrayexpress) and Gene Expression Omnibus (GEO) (www.ncbi.nlm.nih.gov/geo). All information relating to raw data were recorded, such as dates, publication and experimental design; Array Express and GEO accession numbers for datasets, publication ID, citations and replicates can be found in Tables 2.1 and 2.2.

An extreme variation filtering analysis (a gene prioritisation algorithm written in R by Dr Rifat Hamoudi) (Hamoudi *et al.,* manuscript in preparation) was performed across the five cell lines. Briefly, raw data was subjected to normalisation using the GCRMA and MAS5 algorithms, based on the noise in the data. For MAS5, the gene would pass filtering if its value was more than 200 across 3 or more cell lines. For GCRMA, the gene would pass the filtering if its coefficient of variation was more than 50%. The extreme variation filtering analysis was applied to the normalised cell line data, identifying 915 differentially expressed gene probes.

The work in this Section (2.1.1) was carried out by Dr Rifat Hamoudi and Dr Nadège Presneau, (with James Whitehead), and formed the basis for candidate gene selection and pathway analysis in this study.

Table 2.1. Details of the gene expression datasets for the breast cancer cell lines downloaded from online repositories.

Subtype	Cell Line Replica	GEO Accession Number Array Express Number	PubMed ID	Citation/Company
Non-tumourigenic	MCF10A	GSE34211 E-GEOD-34211	PMID: 22222631 PMID: 24107449	Pfizer Hook, *et al.*, 2012 Pavlicek, *et al.*, 2013
	MCF10A	GSE12790	PMID: 19567590 PMID: 21673316	Hoeflich, *et al.*, 2009
	MCF10A	GSE29327 E-GEOD-10890	PMID: 21673316	Stinson, *et al.*, 2011
Luminal A	MCF-7	GSE18912	PMID: 21220496	Hou, *et al.*, 2011
	MCF-7	GSE41445	PMID: 23894636	Bayer Pharma AG Bayer, *et al.*, 2013
	MCF-7	GSE40057 GSE40059	PMID: 23497265	Luo, *et al.*, 2013
Luminal A	T47D	GSE41445	PMID: 23894636	Bayer Pharma AG Bayer, *et al.*, 2013
	T47D	GSE40057 GSE40059	PMID: 23497265	Luo, *et al.*, 2013
	T47D	GSE34211 E-GEOD-34211	PMID: 22222631 PMID: 24107449	Pfizer Hook, *et al.*, 2012 Pavlicek, *et al.*, 2013
Luminal B	BT474	GSE12790	PMID: 19567590 PMID: 21673316	Hoeflich, *et al.*, 2009
	BT474	GSE57083 E-GEOD-57083	Citation missing	Astra Zeneca, 2014
	BT474	E-MTAB-37	PMID:26107615	Li, *et al.*, 2015
Triple Negative	MDA-MB231	GSE41445	PMID: 23894636	Bayer Pharma AG Bayer, *et al.*, 2013
	MDA-MB231	GSE40057 GSE40059	PMID: 23497265	Luo, *et al.*, 2013

Table 2.2. Details of the gene expression files downloaded from online repositories for breast cancer cell lines. These were used for the extreme variation analysis.

Cell Line	Files Used for Extreme Variation Analysis
MCF10A n=10	MCF10A_221004-4.CEL MCF10A_GSM320243.CEL MCF10A_GSM320244.CEL MCF10A_GSM320245.CEL MCF10A_GSM320246.CEL MCF10A_GSM320247.CEL MCF10A_GSM724633.CEL MCF10A_GSM724634.CEL MCF10A_GSM724635.CEL MCF10A_GSM844584_1_PFIZER.CEL
MCF-7 n=9	MCF-7_GSM1017487_mRNA_12a_081206.CEL MCF-7_GSM1017488_mRNA_12b_081206.CEL MCF-7_GSM1017489_mRNA_12c_111206.CEL MCF-7_GSM468593.cel MCF-7_GSM468594.cel MCF-7_GSM468595.cel MCF-7_GSM468596.cel MCF-7_GSM468597.cel MCF-7_GSM984498_MCF7_HG-U133_Plus_2_.CEL
T47D n=7	T47D_GSM1017511_mRNA_20a_200907.CEL T47D_GSM1017512_mRNA_20b_200907.CEL T47D_GSM1017513_mRNA_20c_210907.CEL T47D_GSM844714_1_Good_NCI50_WYETH.CEL T47D_GSM844715_2_Good_BREAST_WYETH.CEL T47D_GSM844716_2_Good_NCI50_WYETH.CEL T47D_GSM984496_T47D_HG-U133_Plus_2_.CEL
BT474 n=7	BT474_brst_SS117188_HG-U133_Plus_2_HCHP-85191_.CEL BT474_brst_SS117189_HG-U133_Plus_2_HCHP-85192_.CEL BT474_brst_SS117190_HG-U133_Plus_2_HCHP-85193_.CEL BT474_GSM1374408_Bx051b_035_HG2.CEL BT474_GSM1374409_ap071105.CEL BT474_GSM1374410_EA08079_80494.CEL BT474_GSM320596.CEL
MDA-MB231 n=4	MDA_MB_231_GSM1017490_mRNA_13a_111206.CEL MDA_MB_231_GSM1017491_mRNA_13b_111206.CEL MDA_MB_231_GSM1017492_mRNA_13c_111206.CEL MDA_MB_231_GSM984500_MDA-MB-231_HG-U133_Plus_2_.CEL

2.1.2 Pathway Enrichment Analysis

Extreme variation data was subjected to pathway and ontology enrichment analysis using multiple tools (discussed below). Extensive literature reviews were undertaken to determine the most suitable methods for this. The purpose of this was to identify in terms of gene expression which pathways and processes were most active within breast cancer cell lines, to identify candidate genes for further functional analysis, as well as to cross-compare the results of the various pathway tools.

2.1.2.1 DAVID, GO & PANTHER

Database for Annotation, Visualisation and Integrated Discovery (DAVID) functional annotation and gene functional classification (Huang *et al.,* 2007), Gene Ontology (GO) enrichment analysis (Ashburner *et al.,* 2000), and PANTHER gene list analysis (Thomas *et al.,* 2003), were used in succession for analysis of extreme variation genes across breast cell lines. For all analyses, a threshold of $p<0.05$ was used to represent statistically significant data, and for consistency, the 'GO terms biological process' annotation set was used in each analysis.

DAVID (Version 6.8) functional annotation clustering was selected and a high classification stringency was used. Genes were identified against 'Affymetrix_3Prime_IVT_ID'. Of the extreme variation list, 650 gene IDs were recognised. The 'GOTERM_BP_ALL' annotation category was selected. Statistically significant clustered enrichment scores were ranked.

For GO (release 2019-02-01) and PANTHER (version 14.0, 2018-12-03) analysis, Affymetrix IDs from the extreme variation list were converted to Ensembl IDs (www.ensembl.org), of these, 634 genes were mapped. Two overrepresentation tests were carried out and combined. The 'GO biological process complete' overrepresentation test, was assessed using Fisher's exact test, and the Bonferroni correction for multiple testing. The 'PANTHER GO-slim biological process' overrepresentation test was also used with Fisher's exact test, and corrected using the false discovery rate. For both of these tests, statistically significant GO sets were ranked based on their fold enrichment scores.

2.1.2.2 Gene Set Enrichment Analysis (GSEA)

Gene Set Enrichment Analysis (GSEA) by the Broad Institute was used to determine the classes of genes and biological pathways over-represented within the list of extreme variation genes (Subramanian, *et al.*, 2005). This represented a more in-depth and cell line specific level of pathway enrichment analysis, highlighting the differences in gene expression between breast cancer subtypes.

The GSEA software (Release 2.0) was downloaded via the Broad Institute from the Massachusetts Institute of Technology (Subramanian *et al.,* 2005). GSEA was performed comparing tumour vs non-tumourigenic (MCF10A) cell lines against the MSigDB C5 Gene Ontology (GO) gene set collection (c5.all.v6.1). This collection consisted of 5,917 gene sets divided into three categories - ‘Biological Process’, ‘Cellular Component’ and ‘Molecular Function’. The normalised enrichment score (NES) was used as a means to quantify the scale of enrichment, and the false discovery rate (FDR) was used to measure statistical significance.

2.1.3 Candidate Gene Selection

For DAVID and combined GO and PANTHER analysis, the genes present in the top 20 enriched annotation clusters and the genes present in the top 20 GO annotations were extracted respectively. For GSEA analysis, the genes from the top 5 significant annotation terms were taken for each cell line. Hence, gene lists were created for each analysis tool (Appendix 1). The gene lists were compared and candidate genes were selected based on their commonality between enriched pathway lists for each analysis. This identified 10 candidate genes. To further condense this list for the selection of a single gene, a literature review was conducted for each gene; criteria was relevance to other cancers, potential oncogenic (favoured above tumour suppressor) function, and possible miRNA regulation. Genes previously well characterised in breast cancer were excluded.

2.2 Validation of ASCL2 in Patient Tumours via the METABRIC Study

Microarray data obtained from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) study (Pereira, *et al.*, 2016) was accessed through the cBioPortal online web application (Cerami, *et al.*, 2012; Gao *et al.,* 2013). Primary fresh-frozen breast cancer specimens were clinically annotated (samples assigned to the PAM50 intrinsic subtypes) and obtained from tumour banks in the United Kingdom and Canada. For transcriptional profiling, RNA was isolated from samples and hybridised to the Illumina HT-12 (v3) platform (Curtis, *et al.*, 2012). Normalised expression level Z-score data of ±2 was used as a threshold to classify clinical breast cancer cases into 3 groups according to the expression of *ASCL2*: upregulated (>+2), downregulated (<-2) and unaltered (-2 to +2).

For the METABRIC study, 2509 patient samples were downloaded and analysed. The parameters measured were ER status, HER2 status, PR status, PAM50 subtype, age, overall survival, stage, grade and integrative cluster (proposed by Dawson *et al.,* 2013). Owing to missing clinical information or expression level data, the total number of valid samples varied for each parameter - e.g. out of 2509 patient samples, *ASCL2* expression was measured in 1904 samples.

Data was downloaded and patient samples were matched with clinical data using Microsoft Excel. In some cases, RStudio, version 3.4.3 (RStudio, Inc., Boston, USA) was used during the formatting process. Using SPSS for Windows, version 24 (SPSS, Inc., Chicago, USA), data was then categorised, numbered, and missing values were defined. The differences between molecular parameters were compared between *ASCL2* expression groups using SPSS generating descriptive statistics and graphs.

Statistical differences in the distribution of *ASCL2* expression between receptor status, subtype, stage, grade, and integrative cluster were analysed using the Pearson Chi-Square (^{X2}) test. Subsequently, to model the impact of *ASCL2* expression on HER2 receptor status, a logistic regression analysis was performed. Comparison between means was analysed using a one-way ANOVA (with Tukey post hoc multiple comparisons test) for patient age of onset and overall survival. Estimates of overall survival were also generated using Kaplan-Meier curves, and statistical significance was based on Mantel-Cox log-rank tests. To examine the prognostic significance of *ASCL2*, univariate and multivariate analyses were performed using the Cox proportional hazards regression models, to estimate Hazard Ratios (HR) and 95% confidence intervals (CI) for associations with overall survival. Gene expression was set as a linear variable, and adjusted for other known prognostic factors; ER status, HER2 status and PR status (all positive vs negative), PAM50 subtype (5 levels), age at diagnosis (≤60 vs >60 years), stage (4 levels), grade (3 levels) and integrative cluster (10 levels) as categorical covariates. Statistical significance was set at <0.05.

2.3 Cell Lines & Culture

The breast cancer cell lines (Table 2.3) were obtained from the American Type Culture Collection (ATCC), provided by Dr Nadège Presneau, Dr Miriam Dwek (University of Westminster) and Professor Marilena Loizidou (University College London). Cells had been previously tested for mycoplasma and genotypes had been verified by short tandem repeat (STR) profiling.

Table 2.3. Details of the breast cancer cell lines used for molecular investigation in this study, including details of the complete growth medium used as recommended by the literature.

	Immunohistochemical Markers			
Cell Line	**ER**	**PR**	**HER2**	**Complete Media**
MCF10a	-	-	-	Mammary Epithelial Cell Grown Medium (MEBM) 2ml bovine pituitary extract 0.5ml human epidermal growth factor 0.5ml insulin 0.5ml hydrocortisone 0.5ml gentamicin-amphotericin 10% FBS 1% L-glutamine 1% Pen/Strep
MCF7	+	+	-	Dulbecco's Modified Eagle's Medium (DMEM) 10% FBS 1% L-glutamine 1% Pen/Strep
T47D	+	+	-	Roswell Park Memorial Institute (RPMI) 1640 10% FBS 1% L-glutamine 1% Pen/Strep
SKBR3	-	-	+	Roswell Park Memorial Institute (RPMI) 1640 10% FBS 1% L-glutamine 1% Pen/Strep
BT474	+	+	+	Dulbecco's Modified Eagle's Medium (DMEM) 10% FBS 1% L-glutamine 1% Pen/Strep
MDA-MB-231	-	-	-	Dulbecco's Modified Eagle's Medium (DMEM) 10% FBS 1% L-glutamine 1% Pen/Strep

Cell lines were maintained in complete media (Table 2.3) as recommended by the American Type Culture Collection (ATCC) or as recommended in the literature, supplemented with 10% v/v fetal bovine serum (FBS), 1% v/v l-glutamine and 1% v/v penicillin streptomycin (unless otherwise stated in the following methods) (all purchased from Lonza BioWhittaker, Switzerland).

2.3.1 Routine Cell Maintenance

Cells were maintained as an adherent monolayer in the appropriate medium, at 37ºC in 5% CO_2 to control for changes in pH. Media and other reagents were warmed to 37ºC prior to use. Cells were subcultured (passaged) at 70-80% confluence. Medium was removed from flasks, and cells were washed with 5 mL phosphate buffered saline (PBS) to remove any residual media. To detach cells, 1 mL 2.5% Trypsin 10X (Gibco, Thermo Fisher Scientific, United Kingdom) was added to the flask, and incubated for 5 minutes at 37ºC. To ensure cells had detached, cells were observed under a light microscope. To prevent cells from further toxicity, Trypsin was deactivated by adding 5mL of complete media, and this mix was added to a 15 mL conical tube. Cell suspensions were centrifuged at 189 x g for 3 minutes, and the supernatant was discarded. Cell pellets were resuspended in media depending on the split ratio and added to 75 cm^2 flasks (usually at a density of ~2.1 x 10^6). Flasks were topped up to a total of 10 mL media and incubated.

All cells were passaged 2-3 times prior to experimental use, allowing cells to recover from thawing; passaging was carried out as required. Cells were checked multiple times weekly to ensure media pH was kept consistent (as indicated by a colour change in the media), and cells were ‘healthy’. Where possible, experiments were conducted with cells that had been passaged up to a maximum of 10 times to minimise genetic drift or genotypic changes that may have affected experimental results.

2.3.2 Cell Storage & Seeding from Frozen

Frozen stocks at low passages were regularly collected and stored in liquid nitrogen until needed.

To freeze a stock of cells, cells were prepared as in Section 2.3.1, however, cell pellets were instead resuspended in 1 mL freezing media (maintained cold), made up of 10% v/v dimethyl sulfoxide (DMSO) (Sigma-Aldrich, United Kingdom) plus 90% v/v FBS, and added to a 1.5 mL cryovial. These were placed in a freezing container containing 100% isopropyl alcohol and stored in a -80ºC freezer for 2 days. Cryovials were moved to liquid nitrogen for long term storage.

To seed cells from frozen, cell stocks were removed from liquid nitrogen and thawed in a water bath at 37ºC. This was transferred to a 15 mL conical tube and topped up with 10 mL media. To remove the freezing solution containing DMSO (which is toxic to cells), cell suspensions were centrifuged at 189 x g for 3 minutes, and the supernatant discarded. The cell pellet was then resuspended in 10 mL media and transferred to a 75 cm^2 flask. Cells were observed daily, and media was changed accordingly.

2.3.3 Cell Counting

Where a defined seeding density was required, a haemocytometer was used to count cells. In this case, cells were prepared as in Section 2.3.1, however, cell pellets were resuspended in 1 mL media and mixed thoroughly. To count, 40 µl of cells was mixed with 40 µl Trypan blue (Lonza BioWhittaker, Switzerland) (1:1 dilution) – as only dead cells take up the blue dye, these were excluded to ensure that only viable cells are counted. Cells were counted using a haemocytometer to determine the volume of cells needed for a set concentration. The number of cells seeded were dependant on the size of the flask or well used.

2.4 Primer Design

Once a gene was selected for analysis, primer sequences were designed and the corresponding Affymetric ID probeset was used to find the target sequence. Primer 3 (Untergasser, *et al.*, 2012) was then used with specified parameters – product size=100-150 bp, primer size optimum=22 bp, primer GC% optimum = 50%. Once primers had been selected, Blat and In-Silico PCR (genome.ucsc.edu), as well as NCBI BLAST (blast.ncbi.nlm.nih.gov/Blast.cgi) were used to check specificity. Finally, SNP checker (secure.ngrl.org.uk/SNPCheck/snpcheck.htm) was used to check if any validated SNPs lay in the regions of either primer.

Other primer sequences were identified in the literature, and were checked for the above parameters using the aforementioned tools prior to use. Oligonucleotide primers were purchased from Eurofins Genomics and resuspended at 100 pmol/μl in accordance with manufacturer instructions. Primer sequences used in this study are shown in Table 2.4.

Table 2.4. Primer sequences for PCR and RT-qPCR.

Gene Symbol	Forward (5' – 3')	Reverse (5' – 3')	Reference
ASCL2	CGT GAA GCT GGT GAA CTT GG	GGA TGT ACT CCA CGG CTG AG	Tian, *et al.*, 2014
BIRC5	CTG GCA GCC CTT TCT CAA GGA CC	CCA AGT CTG GCT CGT TCT CA	
CD44	CCA TCC CAG ACG AAG ACA GT	CCA GAG GTT GTG TTT GCT CC	
CMYC	GTC TCC ACA CAT CAG CAC AAC T	GTT CGC CTC TTG ACA TTC TCC T	Zhu, *et al.*, 2012
CTNNB1	AAA ATG GCA GTG CGT TTA G	TTT GAA GGC AGT CTG TCG TA	
HPRT	GCT ATA AAT TCT TTG CTG ACC TGC TG	AAT TAC TTT TAT GTC CCC TGT TGA CTG G	
LGR5	GAG GAT CTG GTG AGC CTG AGA A	CAT AAG TGA TGC TGG AGC TGG TAA	Zhu, *et al.*, 2012
RPII	GCA CCA CGT CCA ATG ACA T	GTG CGG CTG CTT CCA TAA	Radonić, *et al.*, 2004

2.5 RNA Extraction & Quality Assessment

2.5.1 miRNAeasy Mini Kit, Qiagen

Total RNA was extracted from cell lines grown in 75cm^2 flasks using the miRNAeasy Mini Kit (Qiagen, United Kingdom) following manufacturer's instructions. Working under a Class 2 cabinet, cells were washed with PBS, QIAzol lysis reagent was added to flask, and cells were scraped from the surface of the flask. The cell suspension was collected, 140 µl chloroform was added, and cells were incubated at room temperature for 2 minutes. The sample was centrifuged for 15 minutes at 12,754 x g, 4ºC. The upper aqueous phase was carefully added to a new collection tube, and 525 µl of 100% ethanol was added. This was added to a mini column and centrifuged at 12,281 x g for 15 seconds at room temperature. DNA digest was completed by washing the column membrane with buffer, and adding a DNase solution to incubate for 15 minutes (room temperature). Buffers were added and the sample was centrifuged as per the manufacturer's protocol. The membrane was dried by centrifuging a final time at 12,281 x g for 1 minute. Nuclease-free water was added to the column and the RNA was eluted from the column during the final centrifugation step. RNA was stored at -20ºC.

2.5.2 Microprep Kit, Zymo

To isolate RNA from 6-well plates after siRNA transfection, the Microprep kit (Zymo, Cambridge Bioscience, United Kingdom) was used following manufacturer's instructions. Working under a Class 2 cabinet, media was removed from wells and cells were washed with PBS. Qiazol lysis reagent, 100 µl, was added, and cells were scraped from the surface of the flask using a P200 tip. The cell suspension was collected, vortexed, and 100 µl 100% ethanol was added and mixed. This mix was transferred to a column and centrifuged for 30 seconds at 12,281 x g (all centrifugation steps using this kit was carried out at room temperature). To perform a DNase digest, the column was washed and centrifuged with 400 µl wash buffer. The DNase mix consisted of 5 µl DNase and 35 µl digest buffer, which was added to the column and incubated at room temperature for 15 minutes. Buffers were added and the column was centrifuged in accordance with manufacturer's protocols – 400 µl Directzol and 700 µl RNA wash buffer respectively. The column was transferred to an RNAse free tube and RNA was eluted by centrifuging the column with 15 µl nuclease-free water.

RNA concentration (ng/µl) was quantified using the NanoDrop Spectrophotometer (software ND-2000). For quality control purposes, the absorbance ratios of 260/280 nm (~1.8-2.0) and 260/230 nm (~1.8-2.0) were used to assess purity.

2.6 cDNA Synthesis

Single-stranded cDNA was synthesised from RNA using the High-Capacity RNA-to-cDNA™ Kit (Applied Biosystems, Thermo Fisher, USA). The reverse transcription (RT) reaction mix was prepared using 2X RT Buffer, 20X RT Enzyme mix, nuclease-free H_2O and 500 ng of RNA, to a total of 20 µl. A RT negative control sample was also made containing all components except the 20X RT Enzyme. Samples were incubated at 37ºC for 60 minutes, and 95ºC for 5 minutes, then stored at -20ºC.

2.7 Polymerase Chain Reaction (PCR) & Agarose Gel Electrophoresis

PCR experiments were performed using the AmpliTaq Gold PCR Mastermix kit according to manufacturer's protocols (Applied Biosystems, Thermo Fisher, USA). PCR was used for the purposes of checking primer specificity, qualitatively assessing cDNA sample integrity, and checking reference genes. For each candidate gene, a mastermix was prepared including 10x Buffer II, 10 mM dNTPs, $MgCl_2$, 10 µM Forward and Reverse primer, AmpliTaq Gold, nuclease free H_2O, and diluted cDNA. Samples were loaded onto a thermo cycler (Techne Prime). The following cycles were run: initial denaturation of 95ºC for 5 minutes, denaturation, annealing and extension at 95ºC for 10 seconds, 56ºC for 20 seconds, 72ºC for 30 seconds respectively (35 cycles of each), and a final extension of 72ºC for 5 minutes.

Upon completion, PCR products were separated on an agarose (Fisher Scientific, United Kingdom) gel (1.4%) via electrophoresis (100V, 400mA, 45 minutes), and visualised using SYBR green dye (Lonza BioWhittaker, Switzerland). These steps were carried out on cDNA, to test all primers, prior to RT-qPCR to ensure cDNA was of sufficient quality.

2.8 PCR Purification & Sanger Sequencing

To validate the specificity of the amplified PCR product and primers for *ASCL2*, PCR purification was carried out using the QIAquick PCR Purification Kit (Qiagen, United Kingdom), following manufacturer's protocols. In brief, Buffer PB was added to the PCR sample (in a ratio of 5:1). The sample was added to a QIAquick spin column and centrifuged at 12,281 x g for 1 minute. 0.75 mL Buffer PE was added, and the column was centrifuged. Any flow through was discarded. The column was then transferred to a 1.5 mL Eppendorf tube. Finally, to elute the DNA sample, 50 µl elution buffer was added directly to the QIAquick membrane and this was left to stand for 1 minute before being centrifuged. Purified PCR fragments were packaged and sent to GATC Biotech for confirmative LightRun Sanger sequencing.

Upon receipt of results, GATC viewer (www.gatc-biotech.com) and FinchTV chromatogram software (digitalworldbiology.com/FinchTV) were used to analyse the chromatogram and sequence of *ASCL2*, and thus confirm the identity of the amplified gene.

2.9 Quantitative Reverse Transcription-PCR (RT-qPCR)

Quantitative reverse transcription-PCR (RT-qPCR) was used to check gene expression of candidate genes across tumour and normal cell lines. Reactions were carried out in a Rotor-Gene cycler using the Rotor-Gene SYBR Green PCR kit (Qiagen, United Kingdom) following manufacturer's instructions. The cycling conditions were an initial single hold cycle of 95ºC for 10 minutes (denaturation), 40 cycles of 95ºC for 10 seconds (primer annealing) and 60ºC for 45 seconds (primer extension). A final melt stage was performed at the end of the run to generate a melt curve and check reaction specificity (Figure 2.1), heating from 55ºC to 95ºC at a rate of 1ºC every 5 seconds.

Each cDNA reaction was performed in duplicate, alongside a negative cDNA sample, and a negative non-template control for each pair of primers. All samples were also repeated with the reference gene, RNA polymerase II (*RPII*), chosen due to its stable and equal expression across tissues (Radonić, *et al.*, 2004). Melt curves (Figure 2.1) were inspected to ensure reliability of data, and Ct values were generated from amplification curves (Figure 2.1). Relative quantification of gene expression in cell lines was performed using the calculations below. For statistical analysis, a one-way non-parametric ANOVA was used to test for differences between experimental samples (GraphPad Prism 7 software).

∆CT = mean gene Ct - mean housekeeper Ct

∆∆CT = ∆CT sample of interest - ∆CT control sample

Fold change (log) difference = 2^(-∆∆CT)

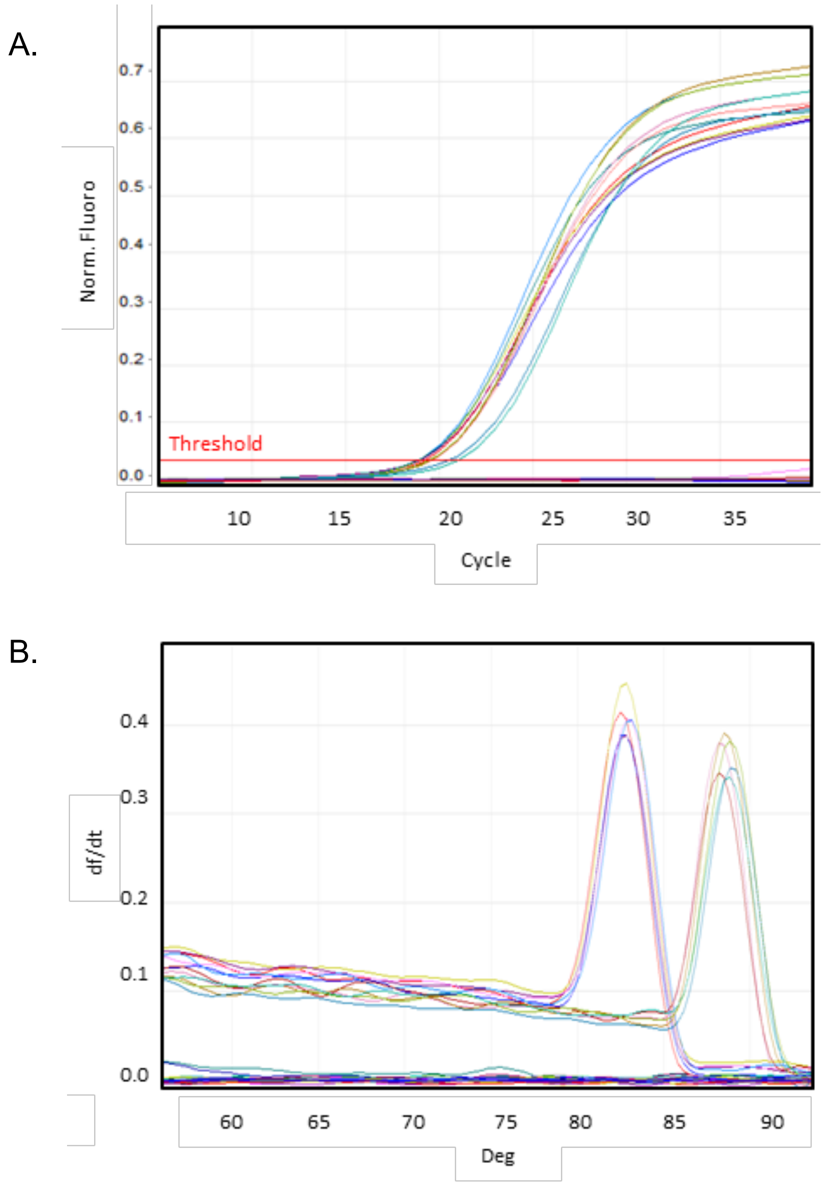


Figure 2.1. RT-qPCR amplification curves and melt curves. A. The typical shape of an amplification curve. The threshold level was set manually at the point where the curve trends upwards, which was the point at which Ct values were calculated. The middle portion of the curve, from cycle 20-25 represents the exponential phase in the reaction. Around cycle 30, the lines begin to plateau, indicating a slowing of the reaction limited by reagents. B. The typical shape of amplicon peaks seen in a melt curve, exemplifying specific gene products, rather than primer dimers or non-specific binding i.e. primers are amplifying specific gene products – in this diagram, 2 peaks are seen representing the gene of interest and a reference gene.

2.10 Short interfering RNA (Dicer siRNA) Knockdown of ASCL2

siRNAs were obtained from IDT using the TriFECTa Dicer-Substrate RNAi kit (Integrated DNA Technologies, Belgium), containing a positive control gene duplex Hypoxanthine-guanine phosphoribosyltransferase (*HPRT*), a non-targeting negative control duplex (NC), a fluorescently labelled transfection control duplex (TYE 563), and three *ASCL2* target specific duplexes, which were subsequently pooled for experiments (siRNA sequences in Table 2.5). Both DharmaFECT (Dharmacon, GE, United Kingdom) and Lipofectamine RNAiMAX (Thermo Fisher, United Kingdom) transfection reagents were used to deliver siRNA targets.

Table 2.5. Three target-specific Dicer-Substrate siRNA (dsiRNA) duplexes, supplied by the TriFECTa kit, for siRNA transfection and knockdown of *ASCL2.* Gene sequence and position of targets are shown in Appendix 3.

siRNA Duplex	Forward (5’ – 3’)	Reverse (5’ – 3’)
ASCL2 13.1	GGGUUAUCUAUACAUUUAAAAACCA	CUCCCAAUAGAUAUGUAAAUUUUUGGU
ASCL2 13.2	GCACCAACACUUGGAGAUUUUUCCG	CCCGUGGUUGUGAACCUCUAAAAAGGC
ASCL2 13.3	GAGCGUGAACUUUAUAAAUAAAUCA	CCCUCGCACUUGAAAUAUUUAUUUAGU

2.10.1 Dicer siRNA Preparation

Lyophilised control duplexes (1 nmole) and target specific duplexes (2 nmole) were briefly centrifuged and resuspended with RNase-free Duplex Buffer (100mM KAc/30 mM HEPES pH 7.5) to a final concentration of 20 µM. DsiRNAs were vortexed and mixed thoroughly. These were diluted to create working solutions at 2 µM. Cells were treated with a final concentration of 10 nM dsiRNA per well.

2.10.2 Knockdown of ASCL2 in Cells

Cells were seeded into plates (24-well, 0.05 x 10^6 cells, or 6-well, 0.3 x 10^6 cells) the day before transfection using antibiotic free media. At approximately 24 hours after seeding, growth medium was removed from each well, and Opti-MEM reduced serum medium (Gibco, Thermo Fisher Scientific, United Kingdom) was added. A transfection mixture was prepared containing Opti-MEM, Lipofectamine (or DharmaFECT) and the corresponding siRNA duplex, and incubated for 15 minutes at room temperature. For each siRNA, a final concentration of 10nM was added to cells. At 24 hours post transfection, cell transfected with siRNA-TYE 563 were visualised under a fluorescence microscope (to check transfection was successful before proceeding). Cells were then harvested and RNA was extracted as in Section 2.5.2.

The cell lines, MCF7, T47D and SKBR3 were transfected and used for further analysis.

2.10.3 Measurement of Knockdown Efficiency

Gene expression and percentage knockdown was assessed by RT-qPCR. Percentage knockdown was calculated by normalising samples to the reference gene, *RPII* (ΔCt), and the non-targeting negative control siRNA sample (NC) (ΔΔCt), then the $(1\text{-}2^{-\Delta\Delta Ct})$*100 equation was used (Section 2.9) (Haimes, & Kelley, 2010). In this study, a knockdown of 70% was preferable, however at least 60% was deemed successful and appropriate for further analysis, resulting in an average of approximately 70% across replicates.

2.10.4 Knockdown Optimisation

To begin with, optimisation was planned with a positive control gene (*HPRT*) in order to establish the correct conditions for siRNA knockdown. Once successful knockdown of *HPRT* was observed via fluorescence microscopy and RT-qPCR, knockdown of the *ASCL2* gene could proceed. However, optimisation presented many challenges, and numerous different experimental and troubleshooting conditions were tested (Appendix 2). An in-depth summary of optimisation and troubleshooting conditions are described in Appendix 2. A timeline of steps proceeding siRNA transfection can be seen in Table 2.7 (at the end of this Chapter).

2.11 Immunostaining of Cells

MCF7 cells were cultured and transfected as per Section 2.10.2. Once approximately 80% confluence was reached (at 48h post transfection) cells were washed with PBS and fixed in 10% v/v formalin in PBS for 30 minutes. Formalin was removed, cells were washed twice with PBS, and cells were stored in PBS at 4°C before analysis.

Prior to staining, PBS was removed from cells, and cells were incubated with blocking solution (5% w/v BSA in PBS) for 30 minutes gently rocking at room temperature. Cells were washed twice in PBS. Primary antibody diluted in 5% w/v BSA in PBS (Table 2.6) was added to cells and incubated gently rocking overnight at 4°C. Cells were then washed multiple times with PBS prior to being incubated with secondary fluorescently labelled antibody (Table 2.6) for 1 hour, gently rocking at room temperature (protected from light to prevent quenching). Cells were washed multiple times in PBS, and incubated with NucRed™ Live 647 (Invitrogen, Thermo Fisher Scientific, United Kingdom) for 15 minutes at room temperature.

Cells were imaged using the GFP and RFP fluorescence channels at 40X magnification using the EVOS FL Auto 2 Cell Imaging System (Thermo Fisher Scientific, funded by The Guy Foundation), and processed using ImageJ (imagej.nih.gov/ij/).

Table 2.6. Antibodies used for western blot experiments.

Antibody	Dilution	Supplier
Primary		
Rabbit Anti-ASCL2, ab107046	1:500	Abcam
Secondary		
Goat anti-rabbit IgG H&L (Alexa Fluor 488), ab150077	1:500	Abcam

2.12 Functional Investigation

Table 2.7 summarises the timeline of steps for validation and functional investigation experiments subsequent to siRNA knockdown.

2.12.1 Alamar Blue Cell Viability Assay

AlamarBlue® cell viability reagent (Invitrogen, Thermo Fisher Scientific, United Kingdom) was used for the assessment of cell viability and proliferation subsequent to siRNA knockdown. Analysis was carried out using untreated cells, non-targeting negative control cells (siRNA-NC) and siRNA-ASCL2 cells. A 'no-cell' control containing media only was also included.

Cells were seeded in 24-well plates and transfected with siRNA. AlamarBlue® reagent was added to growth medium 24h post-transfection, at a concentration of 10% v/v. Cells were then incubated at 37ºC (5% CO_2) for 2 hours, then 50 µl of the Alamar Blue-medium mixture was transferred to a 96-well plate. Absorbance was measured at 570 nm and 600 nm using the SPECTROstar Nano microplate reader (BMG Labtech, United Kingdom).

To assess proliferation, absorbance values were substituted into the following equations (Bio-Rad Laboratories, Inc., 2016):

$$\text{\% difference in reduction of alamarBlue® } = \frac{117216 \bullet A_1 - 80586 \bullet A_2}{117216 \bullet P_2 - 80586 \bullet P_1} \times 100\%$$

$$\text{\% different in reduction compared to untreated } = \frac{117216 \bullet A_1 - 80586 \bullet A_2}{117216 \bullet P_1 - 80586 \bullet P_2} \times 100\%$$

Where,

117216 and 80586 = molar extinction coefficients of oxidised AB at 600nm and 570nm wavelengths respectively.

A_1 and A_2 = absorbance values in experimental samples at 570nm and 600nm respectively.

P_1 and P_2 = absorbance values of untreated samples at 570nm and 600nm respectively.

Three biological replicates were carried out for the cell lines MCF7 and T47D, each of which included 3 technical replicates. For SKBR3, cells followed the same pattern as the previous cell lines, and experiments were concluded after n=1 (inclusive of 3 technical replicates). After calculating percentage difference of Alamar Blue reduction in Excel, values were transferred to GraphPad Prism.

2.12.2 Trypan Blue Assay

The Trypan blue viability assay was used as a means to estimate cell proliferation; the cells that exclude the blue dye can be considered as viable. Cells were seeded in a 24 well plate at a density of 0.05 x 10^6 per well. After 24 hours, cells were transfected with siRNA and incubated for a further 24 hours. The media was removed, cells were washed with PBS and then trypsinised. Trypan blue was added to wells at a 1:1 ratio; the cell suspension was collected and counted using a haemocytometer under a microscope (x10 objective). The four corner squares of the haemocytometer were counted and % viability was calculated. For each cell line, at least three biological replicates (each including 3 technical replicates) were carried out.

2.12.3 Caspase-Glo® 3/7 Assay

The Caspase-Glo® 3/7 Assay (Promega, United Kingdom) was used to assess caspase activity in cells, therefore corresponding to apoptosis, after siRNA knockdown. Analysis was carried out using untreated cells, siRNA-NC and siRNA-ASCL2 cells. A ‘blank’ control containing media only was also included.

Cells were seeded in opaque white 96-well plates (0.02 x 10^4 cells per well) and transfected with siRNA. Caspase reagent was added to growth medium at a ratio of 1:1 and mixed well, 24 hours post transfection. Cells were incubated at room temperature, away from light, for 30 minutes. The luminescent signal was measured using the SPECTROstar Nano microplate reader (BMG Labtech, United Kingdom). For each cell line, three biological replicates were carried out, each of which included 2 technical replicates.

2.12.4 Wound-Healing Scratch Assay

To assess the effect of *ASCL2* gene knockdown on cell migration, a wound-healing scratch assay was performed. This assay was designed based on Yue, *et al.*, (2010). Cells were seeded in 6-well plates to reach a high confluence on the day of transfection. Immediately after transfection (0h), a P200 pipette tip was used to make a scratch in the cell monolayer, across the width of the well. Wells were visualised and imaged manually at 0h, 24h and 48h, using a bright-field microscope and a DCM310-A digital microscope camera. ScopePhoto image software by ScopeTek Ltd (China) was used to capture images.

When imaging cells, a cross was marked on a plastic sheet placed over the plate lid, to ensure the same area was visualised at each time interval. The previous image was also referred to at 24h and 48h to ensure imaging was as precise as possible.

Images were analysed in Image J. Firstly, the scale was set for images, where pixels were converted to µm. Using the polygon selection tool, the cells at the edges of the scratch were traced using straight lines. The area of the scratch was then measured. This was repeated 3 times for each image (3 measurements per 3 time intervals per 3 well treatments), and an average was calculated. Average scratch area was then used to calculate the percentage closure between 0h, 24h and 48h for each experimental sample. The difference in percentage closure between negative controls and *ASCL2* knockdown samples was also calculated.

Experiments were carried out in multiple biological replicates (MCF7, n=6, T47D, n=2, SKBR3=3). Scratch area values, and percentage closure values were transferred to GraphPad Prism, where unpaired t-tests were used to analyse the difference between experimental groups (siRNA-NC vs siRNA-ASCL2).

2.12.5 Effect of ASCL2 Knockdown on Wnt-target Gene Expression

The effect of *ASCL2* knockdown on the Wnt signalling pathway was investigated by assessing changes in the gene expression of Wnt-related genes. These genes were selected based on their involvement in Wnt signalling, their varying roles in cancer development, as well as their relation/interaction with *ASCL2*.

Cells were grown in 6-well plates and subjected to transfection (Section 2.10.2), RNA was extracted from cells (Section 2.5.2), cDNA was synthesised (Section 2.6), and gene expression was measured by RT-qPCR (Section 2.9).

Samples were measured in technical duplicates, and biological triplicates for each gene and cell line combination. Samples were normalised to a reference gene (*RPII*) (ΔCt) and then against the non-targetting negative control (siRNA-NC) transfected sample (ΔΔCt) – the change in the expression of genes after *ASCL2* knockdown was compared to that of genes after non-targeting siRNA knockdown.

The genes measured were as follows: *CD44, CCND1 BIRC5 (SURV), CTNNB1, LGR5 and C-MYC*. Primer sequences can be found in Table 2.4.

2.12.6 Statistical Analysis

Unless otherwise stated, statistical analysis of experimental data was performed using the GraphPad Prism 7 software. Experimental data is represented as mean ± SEM, and statistical significance was determined using unpaired t-tests.

2.12.7 Summary of siRNA Knockdown & Functional Investigation

Table 2.7. Timeline of chronological steps for siRNA knockdown and functional investigation experiments.

	Day 0	**Day 1 (0h)**		**Day 2 (24h)**		**Day 3 (48h)**
Knockdown Validation	Seed Cells	Transfect		Fluorescence Check	Extract RNA and make cDNA	RT-qPCR
Alamar Blue				Treat 2 hrs	Measure fluorescence	
Immunostaining						Wash, Fix, Stain, Image
Trypan Blue				Treat & count cells		
Caspase				Treat 30 mins	Measure luminescence	
Wound Healing			Image	Image		Image

2.13 Summary of Study Design & Methods

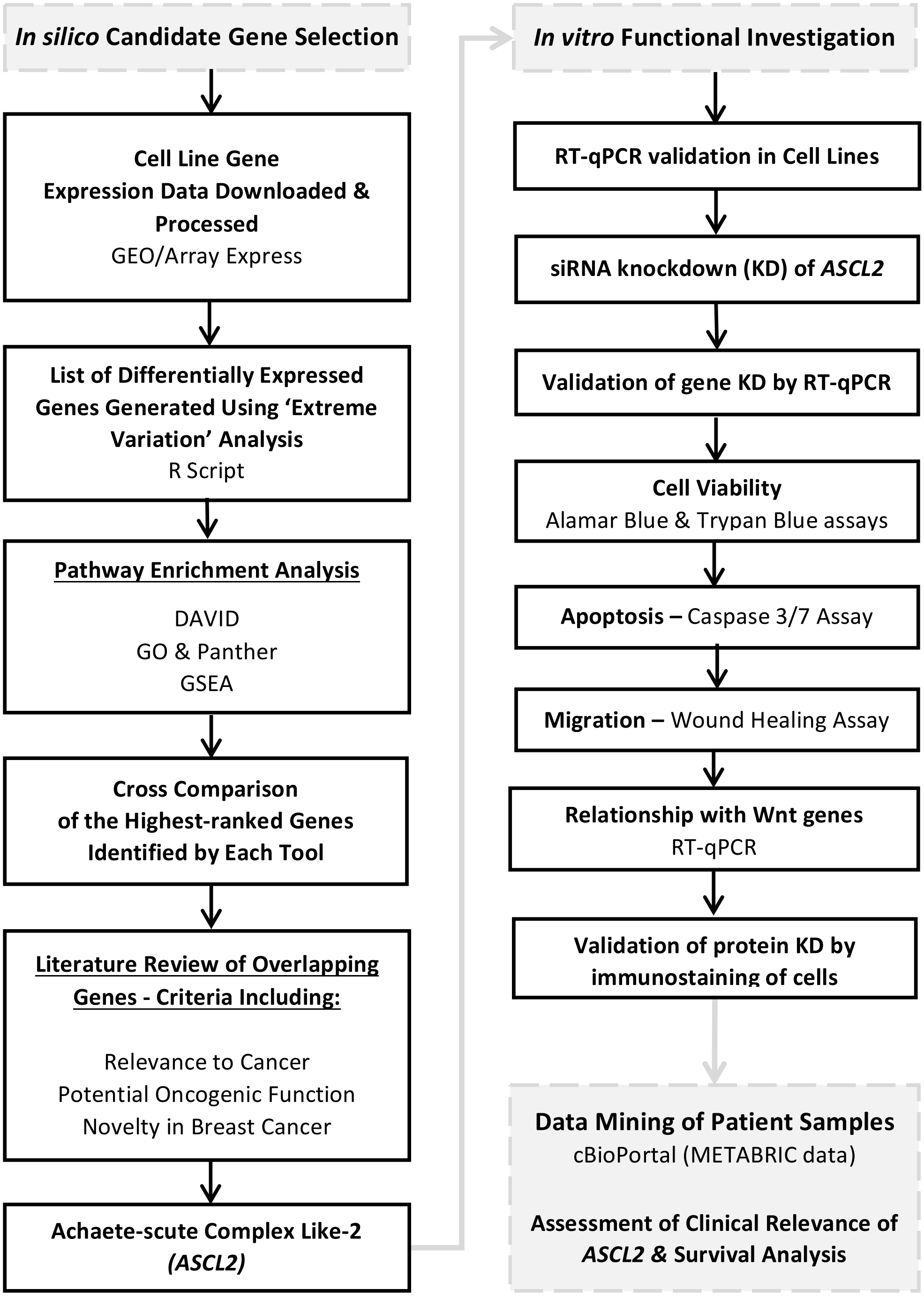

In silico Candidate Gene Selection
Cell Line Gene Expression Data Downloaded & Processed
GEO/Array Express
List of Differentially Expressed Genes Generated Using 'Extreme Variation' Analysis
R Script
Pathway Enrichment Analysis
DAVID
GO & Panther
GSEA
Cross Comparison of the Highest-ranked Genes Identified by Each Tool
Literature Review of Overlapping Genes - Criteria Including:
Relevance to Cancer
Potential Oncogenic Function
Novelty in Breast Cancer
Achaete-scute Complex Like-2 *(ASCL2)*
In vitro Functional Investigation
RT-qPCR validation in Cell Lines
siRNA knockdown (KD) of *ASCL2*
Validation of gene KD by RT-qPCR
Cell Viability
Alamar Blue & Trypan Blue assays
Apoptosis – Caspase 3/7 Assay
Migration – Wound Healing Assay
Relationship with Wnt genes
RT-qPCR
Validation of protein KD by immunostaining of cells
Data Mining of Patient Samples
cBioPortal (METABRIC data)
Assessment of Clinical Relevance of *ASCL2* & Survival Analysis

Figure 2.2. Flow chart depicting the *in silico* and *in vitro* workflow implemented in this study, to select and investigate the role of *ASCL2*.

Chapter III

Identification of ASCL2 as a candidate gene in breast cancer, using pathway enrichment analysis

3.1 Introduction

The clinical heterogeneity observed in breast cancer is mirrored in its complex genetic landscape. The search for novel genetic markers and greater molecular characterisation is necessary to understand breast carcinogenesis and the signalling pathways that are at the core of tumour development. However, research potential for the development of novel markers and therapeutic agents has been limited in the past, for example, due to the lack of specificity of some genes, the low incidence of some genetic changes, or the lack of in-depth knowledge of gene function (discussed in Section 1.3.3). The pursuit of oncogenes is of prominent research interest now more than ever in the era of precision medicine, to better understand the intricacies of each breast cancer subtype and in the hope of developing more targeted therapeutics in the future.

With the emergence of high-throughput techniques within transcriptomics, researchers are now able to rank and group large volumes of genetic data. Pathway analysis has become a prolific and powerful means of efficiently handling gene expression data to advance hypothesis generation and gene discovery (Mathur, *et al.*, 2018). Though the phrase 'pathway analysis' is somewhat broad, in the context of this study it refers to enrichment-based analysis, involving the condensing and aggregation of large transcriptomic datasets into functional groups of related genes termed gene sets; gene sets share a common biological function or property, defined by a reference knowledge base. Knowledge bases, such as Gene Ontology (GO) or MSigDB, are databases comprising molecular information relating to regulation, interaction and phenotypic associations (Khatri, *et al.*, 2012; Mathur, *et al.*, 2018). Large gene lists can be systematically mapped to related GO terms, emphasising the most statistically over-represented terms and highlighting the most relevant biological factors underpinning the samples of interest. In this way, large gene lists are sorted and reduced so biologically meaningful information can be extracted (Huang, *et al.*, 2007).

However, as noted in Section 1.5, selecting the optimal tools from the growing number available, and extracting biologically meaningful results has become increasingly demanding, especially for the evaluation of interconnecting cancer pathways. This process is very complex, involving many steps, and represents an area of cancer research that resides between traditional wet-lab biology and systems biology. To bridge this gap, many web-based tools have emerged to provide sophisticated *in silico* analysis with greater accessibility (Zhang, *et al.*, 2018).

The pathway tools selected in this study, DAVID, GO and PANTHER, and GSEA, (described in Section 1.5.2, Table 1.5) are well established and have been cited in numerous publications (DAVID alone has been cited over 33,000 times according to PubMed). These tools cover two types of pathway analysis methods. Overrepresentation analysis (ORA) (DAVID, GO, PANTHER) results in a list of relevant pathways based on the hypothesis that the proportion of differentially expressed genes, is greater than expected. Functional class scoring (FCS) (GSEA) works on the premise that small yet coordinated expression changes may have significant effects on a pathway, in the same way that large expression changes do (García-Compos, *et al.,* 2015).

The objective of this chapter was to formulate a user-friendly pathway analysis pipeline, utilising popular and convenient *in silico* methods, to identify a novel candidate gene that may play a role in breast tumourigenesis. To execute this, gene expression profiles for multiple breast cancer cell lines were obtained from public databases and were subjected to an extreme variation analysis (a gene prioritisation and filtering algorithm) to identify a list of differentially expressed genes across cell lines (as described in Section 2.1.1) (Hamoudi *et al.,* manuscript in preparation). The pathway analysis tools, DAVID, GO and PANTHER, and GSEA were used to biologically group the most significantly enriched genes in cell lines, to sort and reduce gene expression data into gene sets related to biological processes. Subsequently, the tools were cross-compared, and an in-depth literature review was undertaken, to identify a novel candidate gene for further investigation. Finally, the expression of this gene was validated in breast cancer cell lines using RT-qPCR.

3.2 Results

3.2.1 Pathway Enrichment Analysis Using DAVID, GO & PANTHER

Gene expression data from five breast cancer cell lines (BT474, MCF7, MDA-MB-231, T47D, and MCF10A, representing the non-tumourigenic subtype) were downloaded (described in Section 2.1.1). This data underwent extreme variation analysis to condense expression data to 915 gene probes, with varying expression (log2) across cell lines (Section 2.1.1) (Hamoudi *et al.,* manuscript in preparation).

Extreme variation data was subjected to multiple pathway enrichment analyses, using the tools DAVID, GO and PANTHER, and GSEA. The purpose of this was to efficiently sort gene expression data into biologically meaningful over-represented terms and classify genes into functions to ultimately guide and select a candidate gene for further investigation. From the extreme variation gene list, 650 Affymetrix IDs and 634 Ensembl IDs were mapped in DAVID and GO/PANTHER analysis respectively. This pipeline is illustrated in Figure 3.1.

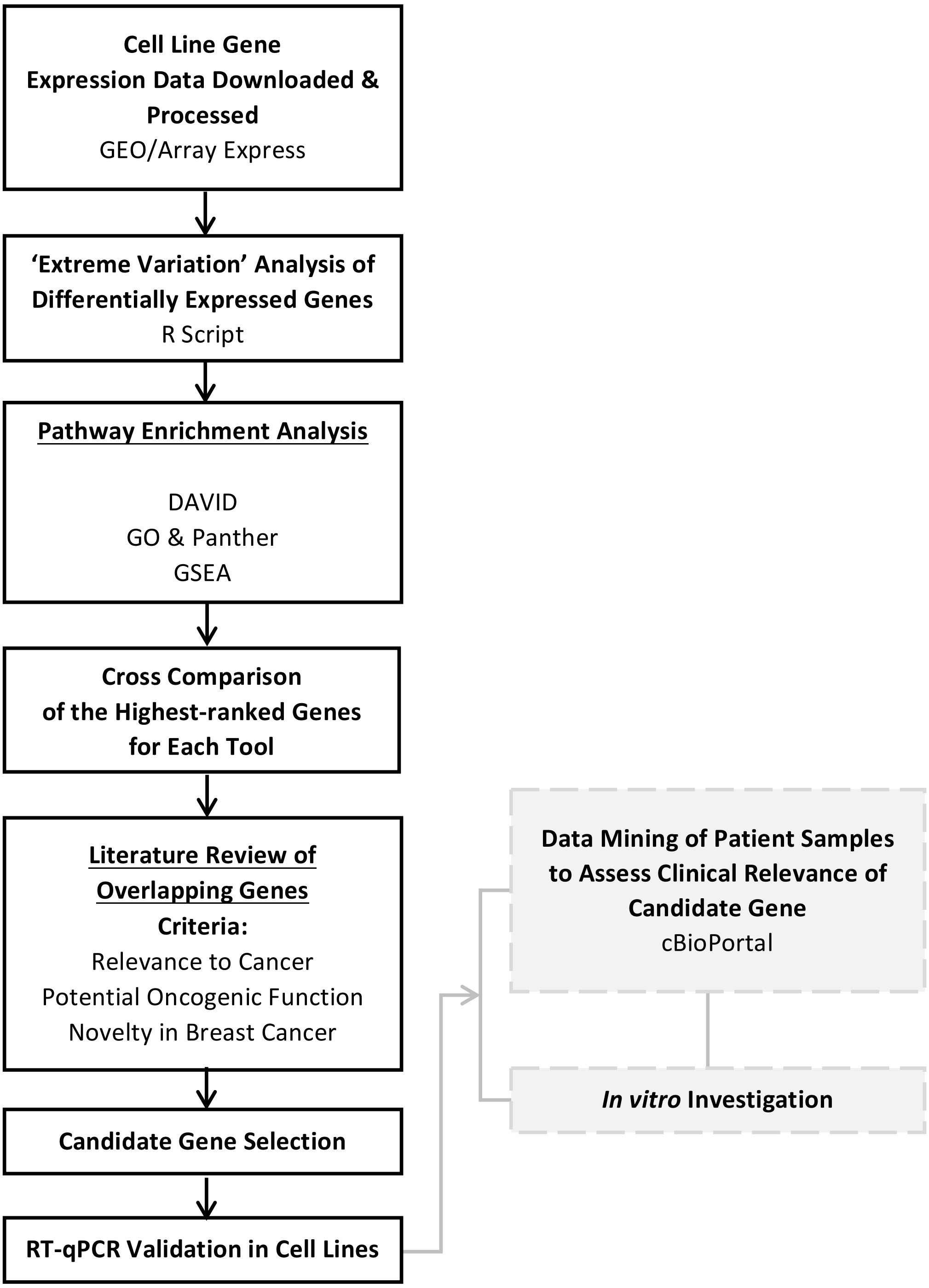

Cell Line Gene Expression Data Downloaded & Processed
GEO/Array Express
'Extreme Variation' Analysis of Differentially Expressed Genes
R Script
Pathway Enrichment Analysis
DAVID
GO & Panther
GSEA
Cross Comparison of the Highest-ranked Genes for Each Tool
Literature Review of Overlapping Genes
Criteria:
Relevance to Cancer
Potential Oncogenic Function
Novelty in Breast Cancer
Candidate Gene Selection
RT-qPCR Validation in Cell Lines
Data Mining of Patient Samples to Assess Clinical Relevance of Candidate Gene
cBioPortal
In vitro Investigation

Figure 3.1. Pathway enrichment analysis pipeline for transcriptomic data: workflow of pathway/ontology enrichment analyses and tools used to identify a candidate gene(s) for further analysis. Enriched ontologies were selected based on commonality between analyses. Data mining of patient samples and *in vitro* investigation will be followed up in subsequent chapters of this thesis.

Figure 3.2 shows the top 20 (of 348) pathway clusters from DAVID analysis exhibiting significant enrichment in breast cancer; among these, regulation of peptidases, developmental processes and cell migration/motility had the largest enrichment scores. Terms relating to migration and motility, including EMT, reoccurred numerous times in this top 20 list, as well as developmental ontologies such as reproductive development processes, and ontologies relating to apoptosis were also recurrently enriched.

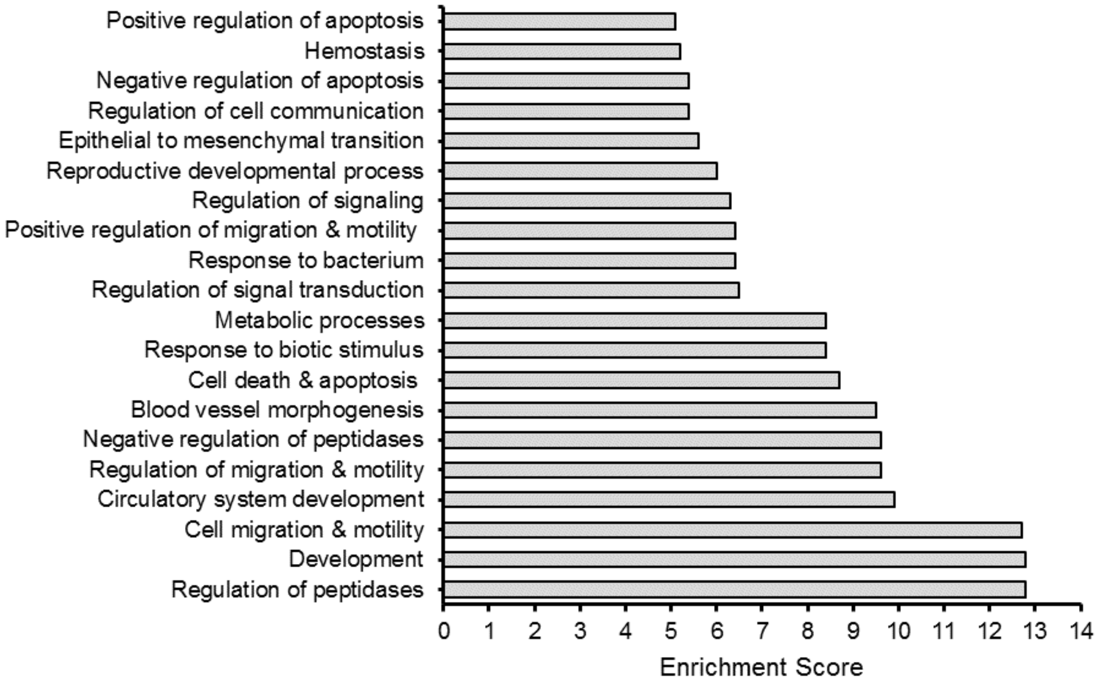


Figure 3.2. DAVID functional annotation clustering analysis using a high classification stringency, highlighting the top 20 (of 348) enriched annotated GO and pathway clusters, ranked by enrichment score. All top 20 clusters were considered statistically significant, *P<0.001*. Of 915 extreme variation gene IDs, 650 gene IDs were mapped.

For GO and PANTHER analysis, two overrepresentation tests were carried out (GO biological process complete and PANTHER GO-slim biological process analyses) shown respectively in Table 3.1 and 3.2. The latter test uses a smaller set of GO terms representing those that have been specifically curated and decided to be most informative of function and evolutionarily conserved (Mi, *et al.*, 2019).

The Top 20 significantly enriched GO pathways highlighted in Table 3.1, showed that enriched biological process ontologies were mainly relating to gland (prostate) development and morphogenesis (sitting at the top of the analysis on the basis of fold enrichment) EMT, hemostasis, wound healing and embryonic skeletal development. PANTHER analysis, shown in Table 3.2, displays significant enrichment in ontology terms associated with nervous system development and neurogenesis, cell development, cell death and cell adhesion.

Although different pathway analysis tools were used, built upon differing algorithms, they were all configured to use the same GO biological process annotation set for consistency. Overall, at a superficial level, the three tools discussed so far exhibited a good agreement and identified similar patterns of enriched ontology terms in the dataset; terms relating to cell development, migration/motility, EMT, cell death/apoptosis, and metabolism were consistently enriched in cell line data. It is also of note that the enriched ontologies apparent in the cell line expression data aligned well with many cancer hallmarks such as sustained angiogenesis, evasion of apoptosis, abnormal metabolic pathways and invasion and metastasis (Hanahan, & Weinberg, 2011). In light of this, although 650 Affymetrix IDs and 634 Ensembl IDs mapped to DAVID and GO/PANTHER respectively (compared to the full list of 915 IDs), a broad spectrum of GO terms were highlighted in correspondence with oncogenic features, representing that this analysis was sufficiently inclusive and wide-ranging, encompassing important pathways with biological significance.

Table 3.1. Gene ontology analysis of extreme variation genes showing the top 20 (of 248) enriched biological process ontologies. Of 915 gene IDs, 634 were mapped to the human reference list. Results classified as significant using Fisher's Exact statistical test and corrected using the Bonferroni correction for multiple testing (*p<0.05*) were ranked on the basis of fold enrichment score.

GO biological process complete	# Human Reference	# Ext Var	Expected	+/-	Fold Enrichment	P value
hemidesmosome assembly (GO:0031581)	12	7	0.36	+	19.47	5.54E-03
prostate gland morphogenesis (GO:0060512)	24	10	0.72	+	13.91	2.47E-04
prostate gland epithelium morphogenesis (GO:0060740)	22	8	0.66	+	12.14	1.42E-02
prostate gland development (GO:0030850)	40	11	1.2	+	9.18	1.67E-03
regulation of collagen metabolic process (GO:0010712)	42	10	1.26	+	7.95	1.86E-02
gland morphogenesis (GO:0022612)	99	22	2.97	+	7.42	5.05E-08
positive regulation of epithelial to mesenchymal transition (GO:0010718)	46	10	1.38	+	7.26	3.76E-02
cornification (GO:0070268)	112	22	3.36	+	6.56	4.13E-07
regulation of blood coagulation (GO:0030193)	77	14	2.31	+	6.07	3.19E-03
regulation of hemostasis (GO:1900046)	78	14	2.34	+	5.99	3.66E-03
regulation of coagulation (GO:0050818)	81	14	2.43	+	5.77	5.48E-03
regulation of epithelial to mesenchymal transition (GO:0010717)	82	14	2.46	+	5.7	6.25E-03
positive regulation of ossification (GO:0045778)	83	13	2.49	+	5.23	3.64E-02
regulation of wound healing (GO:0061041)	130	20	3.89	+	5.14	1.30E-04
negative regulation of epithelial cell proliferation (GO:0050680)	124	18	3.71	+	4.85	1.49E-03
digestive system development (GO:0055123)	140	19	4.19	+	4.53	1.75E-03
regulation of response to wounding (GO:1903034)	155	21	4.64	+	4.52	4.12E-04
digestive tract development (GO:0048565)	128	17	3.83	+	4.43	9.98E-03
extracellular matrix organization (GO:0030198)	333	42	9.98	+	4.21	6.91E-10
embryonic skeletal system development (GO:0048706)	127	16	3.8	+	4.21	3.80E-02

Table 3.2. PANTHER analysis of extreme variation genes, showing the top 20 list of biological process ontologies. Of 915 gene IDs, 634 were mapped to the human reference list. Results classified as significant using Fisher's Exact statistical test and corrected using the false discovery rate (*p<0.05*) were ranked on the basis of fold enrichment score.

PANTHER GO-Slim Biological Process	# Human Reference	# Ext Var	Expected	Fold Enrichment	P value	FDR
smooth muscle contraction (GO:0006939)	9	4	0.27	14.84	4.11E-04	4.59E-02
collagen fibril organization (GO:0030199)	13	5	0.39	12.84	1.28E-04	2.29E-02
regulation of neurogenesis (GO:0050767)	33	7	0.99	7.08	1.38E-04	2.24E-02
regulation of nervous system development (GO:0051960)	40	7	1.2	5.84	3.90E-04	4.64E-02
cell development (GO:0048468)	92	14	2.76	5.08	2.42E-06	7.20E-04
cellular protein metabolic process (GO:0044267)	110	16	3.3	4.86	8.00E-07	3.58E-04
multi-organism process (GO:0051704)	70	9	2.1	4.29	4.73E-04	4.02E-02
regulation of multicellular organismal development (GO:2000026)	71	9	2.13	4.23	5.19E-04	4.03E-02
cell proliferation (GO:0008283)	88	11	2.64	4.17	1.43E-04	2.14E-02
cell differentiation (GO:0030154)	350	35	10.49	3.34	2.97E-09	5.30E-06
cell death (GO:0008219)	174	16	5.21	3.07	1.46E-04	2.00E-02
regulation of cell death (GO:0010941)	214	17	6.41	2.65	4.53E-04	4.05E-02
protein metabolic process (GO:0019538)	441	34	13.21	2.57	2.21E-06	7.88E-04
cellular developmental process (GO:0048869)	506	39	15.16	2.57	3.41E-07	2.03E-04
anatomical structure development (GO:0048856)	389	27	11.65	2.32	1.53E-04	1.95E-02
cell projection organization (GO:0030030)	354	24	10.61	2.26	4.82E-04	3.91E-02
cell adhesion (GO:0007155)	380	25	11.38	2.2	4.46E-04	4.43E-02
biological adhesion (GO:0022610)	380	25	11.38	2.2	4.46E-04	4.19E-02
developmental process (GO:0032502)	1035	68	31.01	2.19	5.58E-09	4.99E-06
system process (GO:0003008)	493	32	14.77	2.17	9.71E-05	1.93E-02

3.2.2 Pathway Enrichment Analysis Using GSEA

GSEA was performed to examine subtype-cell line specific differences in gene expression and to shed further light on pathway functions prior to selection of an oncogenic candidate gene. Analysis was completed to compare gene set enrichment in the extreme variation data of tumour cell lines vs non-tumourigenic cells (MCF10A). The normalised enrichment score (NES) and false discovery rate (FDR) were used to assess enrichment.

GSEA (Table 3.3) revealed that MCF7 Luminal A tumours exhibited significant upregulation of genes associated with nervous system development; it is now recognised that the nervous system has a large role in cancer development, and metastasis, with a primary site of breast cancer metastasis being the brain (Kuol, *et al.*, 2018). The top GO gene set enriched in MCF7 cells compared to MCF10A cells, 'Regulation of Nervous System Development' (Table 3.3), was highlighted as 1 of 7 gene sets significant at FDR <25% out of 476 upregulated gene sets. Figure 3.3 illustrates the corresponding enrichment plot, exhibiting the highest normalised enrichment score in MCF7 cells. The enriched gene probes within this gene set are shown in Table 3.3; a number of these are already established as oncogenic genes in breast cancer, such as *FOXA1, SOX2, BMP5* and *BMP7*, and *ID4*. In concordance with GSEA analysis, regulation of nervous system development was also ranked highly in PANTHER analysis, Table 3.2.

TNBC tumours (MDA-MB-231) displayed gene sets significantly enriched in GTPase mediated signal transduction, while T47D-Luminal A and BT474-Luminal B showed no significantly enriched gene sets according to the false discovery rate values (Table 3.3).

Table 3.3. GSEA investigation showing the top 5 GO gene set results for each cell line ranked normalised enrichment score (NES). Gene probes are shown corresponding to their rank. Only gene probes from gene sets with false discovery rate (FDR) <25% were considered significant **Indicates significance at FDR <25% *** *ASCL2*, candidate gene selected in this study, significantly enriched within the regulation of nervous system ontology.

Cell Line/Subtype	Gene set	Gene Probes	NES	Nominal p-value	FDR q-value
MCF7 / Luminal A	GO_REGULATION_OF_NERVOUS_SYSTEM_DEVELOPMENT	*ASCL1, CXCL12, SOX3, BMP5, SYT1, DSCAM, OLFM1, FOXA1, SOX2, PCP4, ASCL2***, ID4, KLK6, BMP7, COL3A1, PREX1, EPHA7, NRCAM, PACSIN1, RET, DLX1, TBC1D30, SBF2*	2.28	<0.001	0.099**
	GO_REGULATION_OF_NEURON_DIFFERENTIATION		2.24	<0.001	0.070**
	GO_REGULATION_OF_NEURON_PROJECTION_DEVELOPMENT		2.21	<0.001	0.063**
	GO_REGULATION_OF_CELL_PROJECTION_ORGANIZATION		2.17	<0.001	0.078**
	GO_ENSHEATHMENT_OF_NEURONS		2.14	<0.001	0.089**
T47D / Luminal A	GO_NEURON_FATE_COMMITMENT	*HOXC10, FOXA1, DLX1, PON3, ACOX2, RASGEF1A, PREX1, APOE, PLCE1, LPAR1, PLEKHG4B, HLA-DMA, CD74, AZGP1*	1.87	<0.05	1.00
	GO_ORGANIC_ACID_CATABOLIC_PROCESS		1.85	<0.05	1.00
	GO_REGULATION_OF_SMALL_GTPASE_MEDIATED_SIGNAL_TRANSDUCTION		1.85	<0.05	1.00
	GO_REGULATION_OF_RAS_PROTEIN_SIGNAL_TRANSDUCTION		1.84	<0.05	1.00
	GO_MHC_PROTEIN_COMPLEX		1.71	<0.05	1.00
BT474 / Luminal B	GO_SIGNAL_RELEASE	*FAM3B, LIN7A, NRXN3, TBX3, SYTL5, SYTL2, GAL, SYNRG, CLTA, SYT1, HNMT, HGD, PON3, BCAT1, ACOX2, PREX1, PLEKHG4B, MCF2L. ERBB2, RET, SBF2, GFRA1, RASGEF1A, GSTM3, TBC1D30, SHANK2, MARCKS*	2.09	<0.001	0.704
	GO_ORGANIC_ACID_CATABOLIC_PROCESS		1.92	<0.05	1.00
	GO_RAS_GUANYL_NUCLEOTIDE_EXCHANGE_FACTOR_ACTIVITY		1.92	<0.05	1.00
	GO_CILIARY_PART		1.84	<0.05	1.00
	GO_TRANSPORT_VESICLE_MEMBRANE		1.81	<0.05	1.00
MDA-MB-231 / TNBC	GO_REGULATION_OF_RAS_PROTEIN_SIGNAL_TRANSDUCTION	*NRG1, LPAR1, ABCA1, TGFB2, PLCE1*	1.99	<0.001	0.116**
	GO_REGULATION_OF_SMALL_GTPASE_MEDIATED_SIGNAL_TRANSDUCTION		1.96	<0.001	0.085**
	GO_POSITIVE_REGULATION_OF_SMALL_GTPASE_MEDIATED_SIGNAL_TRANSDUCTION		1.88	<0.001	0.194**
	GO_REGULATION_OF_RHO_PROTEIN_SIGNAL_TRANSDUCTION		1.88	<0.001	0.153**
	GO_ENDOSOME		1.74	<0.001	0.703

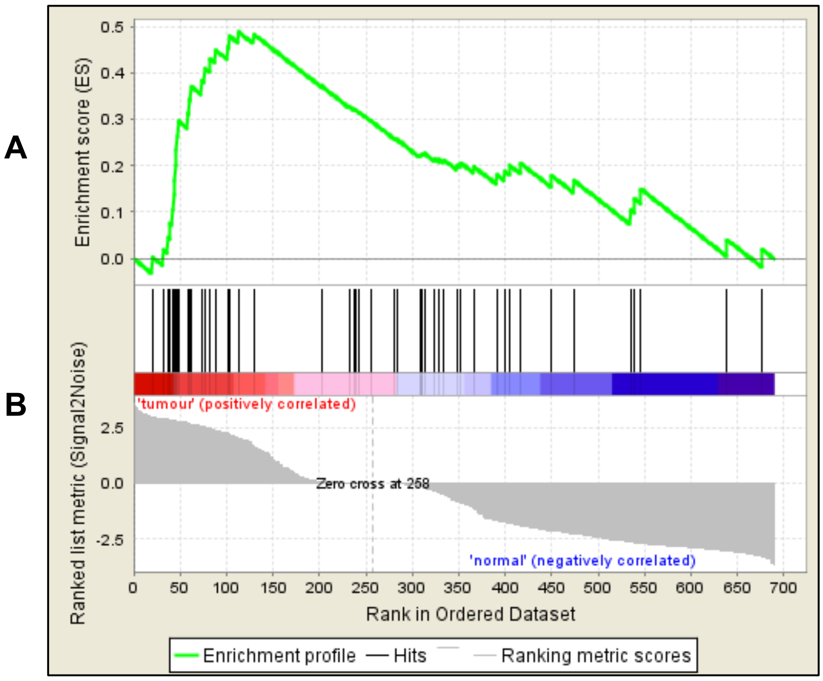


Figure 3.3. GSEA enrichment analysis showing an enrichment plot for the ‘regulation of nervous system development’ ontology. Panel A represents the enrichment score of genes (hits) ordered by a spectrum of correlation from MCF7 cells (tumour, red) to MCF10A cells (normal, blue). Panel B shows the ranked list of genes between tumour and normal phenotypes. The ‘regulation of nervous system’ development was the top most significantly enriched GO gene set in the MCF7 cell line.

3.2.3 Cross-comparison of Tools and Selection of ASCL2

As the GO biological processes most active within transcriptomic cell line data had been highlighted, the results of this analysis was further interrogated. The gene probes present within each of the top enriched GO terms highlighted by the distinct analysis tools (presented in Figure 3.2, Table 3.1, 3.2 and 3.3) were extracted and collated.

The Venn diagram in Figure 3.4 illustrates the number of gene probes identified by pathway analysis for each tool, as well as the number of genes commonly identified by each tool. Despite the gene sets and enriched ontology terms ranked by each tool exhibiting a good agreement overall, there appeared to be discrepancies between the tools at the gene probe level. As a result, direct comparison indicated there was no identical crossover of gene probes (other than the fact that all GSEA gene probes were also present in the high number of DAVID gene probes).

In spite of this, the large number of gene probes identified by DAVID analysis compared to the intermediate and small number of genes highlighted by combined GO and PANTHER analysis, and GSEA respectively, allowed a large number of gene probes (915) to be reduced to a group of 10 genes (Figure 3.4). Thus, candidate genes were chosen on the basis of their appearance in enriched lists and clusters common to all pathway tools used in this study.

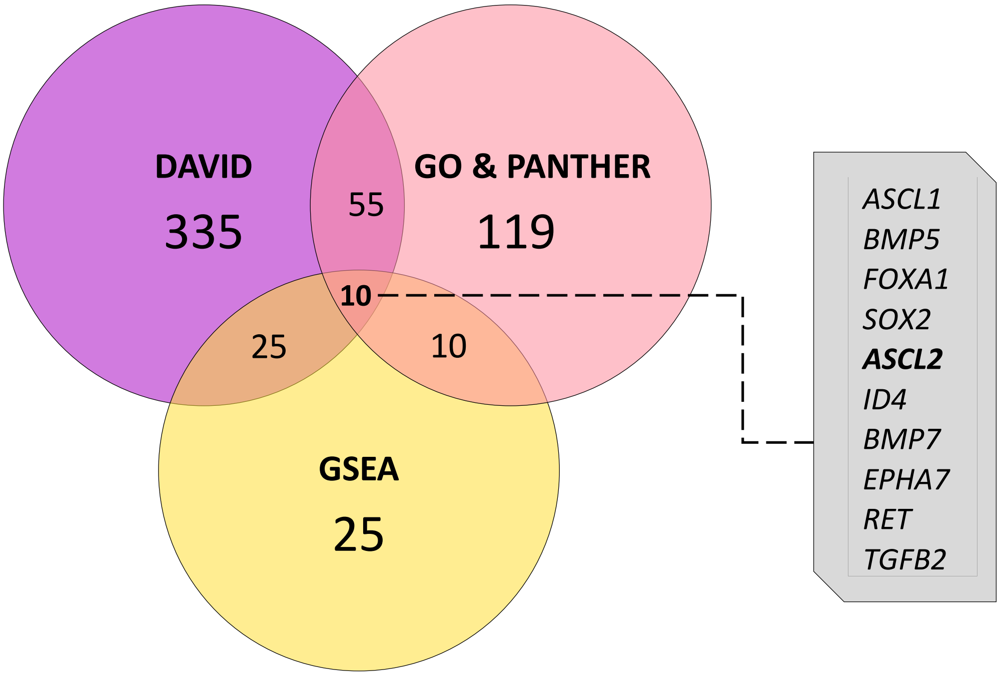


Figure 3.4. Venn diagram depicting the number of candidate genes identified by each pathway analysis tool, and the number of genes overlapping between tools. Overall, the genes present in the cross-section of all analyses (shown in the grey box) were considered as the final list of candidate genes.

An extensive literature review was subsequently undertaken as described in Section 2.1.3 and Figure 3.1. Genes with an already established link to breast cancer were omitted leaving *ASCL2*, which was yet to be defined in breast cancer. Previous reports highlighted the role of *ASCL2* in colon cancer, more specifically as a Wnt-target gene, affirming that this gene had oncogenic potential (Giakountis, *et al.*, 2016; Jubb, *et al.*, 2006; Schuijers, *et al.*, 2015). According to pathway analysis and in line with the literature, this gene was identified to be enriched within developmental processes and nervous system development (Table 3.4). Interestingly, *ASCL1*, a member of achaete-scute complex-like family, was also identified in pathway analysis (Figure 3.4); this gene similarly controls the development of the nervous system in early embryonic stages and has been previously identified as a potential oncogene in lung cancer (Augustyn *et al.*, 2014; Borromeo *et al.,* 2016; Wang *et al.,* 2017). These results, alongside a systematic review by Wang *et al.,* (2017), suggest that the *ASCL* family may be an interesting family of genes to focus on in future studies. However, for the purposes of this study, *ASCL2* will be of central focus; a detailed literature review and evidence supporting the rationale behind the selection of *ASCL2* can be found in Section 1.2.3.

With regards to the GSEA data previously presented (Section 3.2.2), Figure 3.3 illustrates an enrichment plot for the ‘regulation of nervous system development’ ontology. A strong enrichment for the nervous system gene signature was observed whereby *ASCL2* (0.27 enrichment score) was positively correlated with the MCF7 cell line. Within this ranked gene list, *ASCL2* was ranked 11 of 20 genes that were positively correlated with the MCF7 tumour phenotype, and of 52 genes overall. Of the entire ranked gene list of all gene sets identified in MCF7 cells, *ASCL2* ranked 49 of 692 genes with an enrichment score of 2.79.

Table 3.4. A summary of the gene ontology terms associated with *ASCL2*, identified in pathway analysis. This information highlights that *ASCL2* is a developmental gene, involved in the regulation of the nervous system and cell development.

Tool	GO Number & Name
DAVID	GO:0048856 anatomical structure development GO:0032502 developmental process GO:0048731 system development GO:0007275 multicellular organism development GO:0044767 single-organism developmental process
GO/PANTHER	GO:0050767 regulation of neurogenesis GO:0051960 regulation of nervous system development GO:0048468 cell development GO:2000026 regulation of multicellular organismal development GO:0030154 cell differentiation
GSEA	GO:0051960 regulation of nervous system development

3.2.4 Extreme Variation Analysis of ASCL2

After cross-comparison between multiple pathway enrichment analyses, process of elimination and consideration of novelty, the gene most prominently lacking investigation in breast cancer was *ASCL2* (Table 3.4). For an indication of the differential expression of *ASCL2* across breast cancer cell lines, the initial extreme variation analysis was referred back to.

As can be seen from the extreme variation analysis in Figure 3.5, *ASCL2* was significantly expressed most highly in the luminal subtypes, BT474 and MCF7 (9.16 and 8.84 respectively). A lower expression was seen in T47D, with lowest expression observed in MDA-MB-231 and MCF10A. Here, the expression in MCF10A (2.51) was regarded as 'normal' or baseline. The lowest expression seen in the cancer cell lines was in the triple negative subtype. This data is consistent with that of the GSEA analysis whereby *ASCL2* was seen to be significantly enriched in the MCF7 cell line, thus these findings suggest that *ASCL2* expression may be associated with the characteristics of MCF7 cells.

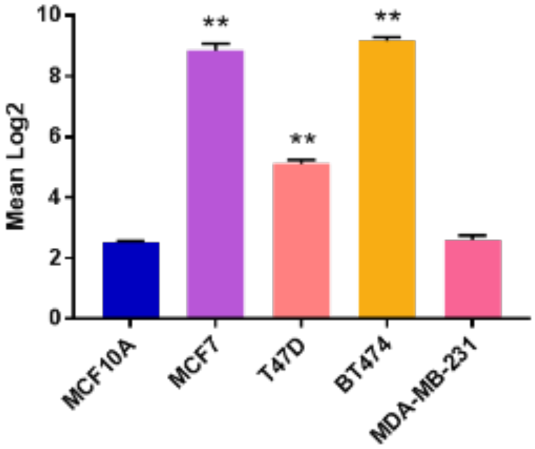


Figure 3.5. Extreme variation analysis of *ASCL2*, n<4 per cell line, represented as a bar chart and scatter plot. Transcriptomic data highlights statistically significant differential expression of breast tumour cell lines compared to a normal-like control, MCF10A (mean ± SEM, ***p<0.0001*). *ASCL2* is expressed most in BT474 and MCF7 respectively, and highly overexpressed compared to non-tumourigenic breast cells.

3.2.5 RT-qPCR Validation of ASCL2 Expression in Breast Cancer Cell Lines

The extreme variation analysis highlighted that *ASCL2* was expressed most highly in BT474, MCF7 and T47D cells respectively, with pathway enrichment analysis specifying the involvement of *ASCL2* in MCF7 cells. The next step was to validate these findings in breast cancer cell lines, ensuring that *in silico* data was concordant with cell lines models prior to continuation of the project.

Prior to quantitative measurement of gene expression in cell lines, a standard PCR was carried out in MCF10A, MCF7, T47D, BT474, SKBR3 and MDA-MB-231 cells, and samples were run on an agarose gel by standard electrophoresis. This was for quality control purposes to improve reliability of data, optimise experimental conditions, ensure sample integrity prior to proceeding and to ensure PCR products and primers were entirely specific for the *ASCL2* gene (Appendix 3). The PCR products were visualised and sequenced; Sanger sequencing (Appendix 3) validated the gene sequence and the identity of the *ASCL2* gene was confirmed. Results from PCR analysis indicated that *ASCL2* was expressed only in MCF7 cells (Appendix 3), mostly consistent with *in silico* data; though, no band was visualised in BT474 or T47D cells indicating no expression. However, due to the lack of sensitivity and qualitative nature of this technique, RT-qPCR was also carried out.

For quantitative evaluation of *ASCL2* gene expression across the cell lines, RT-qPCR, was undertaken. Figure 3.6 shows the gene expression across cell lines relative to the reference gene, RNA polymerase II (*RPII*), and the non-tumourigenic cell line, MCF10A; *ASCL2* was expressed most highly in MCF7 cells, concordant with *in silico* and previous data from PCR analysis. In experimental data, T47D and SKBR3 cell lines also showed expression of *ASCL2*, whereas MDA-MB-231 was seen to show little to no *ASCL2* expression; with the exception of SKBR3, which was not analysed in the previous chapter, T47D and MDA-MB-231 cells also matched transcriptomic data in Chapter 3. Although, contrary to RT-qPCR data, the expression of BT474 appeared inconsistent between *in silico* and *in vitro* data, as extreme variation data highlighted the greatest gene expression in this cell line. Descriptive data is presented in Appendix 4.

Due to unforeseen circumstances towards the end of this project, replacement MCF7 cell line stocks were obtained from a collaborator (Prof Marilena Loizidou); *ASCL2* expression matched previous MCF7 cells further validating the expression of *ASCL2* in MCF7 Luminal A breast cancer cells (data not presented).

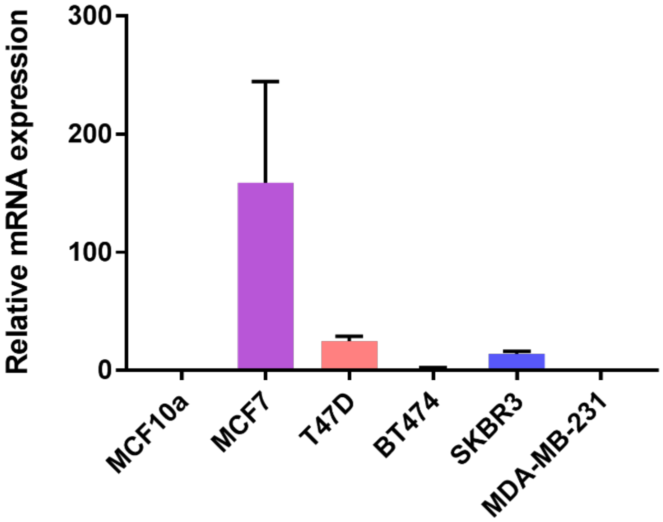


Figure 3.6. RT-qPCR analysis: Quantification of relative gene expression highlights varied expression of *ASCL2* across breast cancer cell lines, particularly in MCF7, T47D and SKBR3 cell lines, compared to 'normal-like' MCF10A cell line (mean ± SEM, n=2, *p<0.05*). MCF7 exhibited the highest gene expression. The variance in error bar size was likely due to biological variation between cell passages.

3.3 Discussion

The work presented in this chapter used multiple computational methods, to highlight the most over-represented functional groups of genes from the expression profiles of breast cancer cell lines. As a result, *ASCL2,* a novel candidate gene was identified for further investigation in breast cancer, and an original pathway analysis pipeline was developed (Figure 3.1). Pathway analysis was completed at the beginning of the study, and later reviewed at the end of the project.

A strength of using gene prioritisation (the extreme variation algorithm, by Hamoudi *et al.,* manuscript in preparation) combined with a layered pathway analysis approach, is that it captures critical aspects of tumour biology, which analysing mutation or gene expression data alone lacks. Such comparative analysis harnessing the power of multiple methods have been demonstrated to outperform the use of single tools, providing more biologically meaningful results (Alhamdoosh, *et al.*, 2017). At its core, cancer is a disease of dysfunctional pathways, therefore it is imperative that the growing amount of omics data is used to understand how genetic disturbances cooperatively impact normal pathway function (Frost, & Amos, 2018). This type of investigation not only raises awareness of likely genetic candidates in breast cancer, but also provides clues as to how these genes function to direct research hypotheses and allow efficient research execution. In this case, the gene *ASCL2* was selected for further investigation, and, extreme variation and GSEA data directed the use of the MCF7 cell line for primary laboratory investigation, which is useful as a starting point for experimental design and analysis.

For consistency, analysis tools were selected to use the GO classification and annotation database under the ‘biological process’ domain, therefore comparison between tools was as reliable as possible. It was also ensured that the tools selected in this study encompassed both ORA and FCS methods. These were deliberately used in combination to ensure the analysis pipeline was robust, and as an attempt to circumvent prominent pitfalls. Although these conventional tools have been well recognised in the literature, they were not without their limitations (Khatri, *et al.*, 2012).

Firstly, ORA treats each gene as an equal, overlooking any expression values and concentrating on the quantity of genes only. In doing this, ORA assumes the independence of genes and ignores any interactions, co-expression, or downstream effects; this results in the reduction of highly complex biological interactions in cancer, to a simplistic configuration. Therefore, these methods do not account for the dependence among pathways (Khatri, *et al.*, 2012). This type of analysis also disregards any genes deemed insignificant based on arbitrary thresholds, meaning that false negatives may arise and some information may be neglected. In this sense, the stringency of statistical analysis can be seen as a challenging equilibrium (Khatri, *et al.*, 2012).

Secondly, though FCS analysis can be considered as an improvement on ORA, FCS still considers pathways independently of one another, therefore omitting the biological crosstalk of pathways, and the fact that genes can function in multiple pathways - these only consider pathways as overly simplified and independent groups. This is challenging to bypass as defining gene sets based on GO terms is hierarchical by nature (Khatri, *et al.*, 2012). For example, GSEA assessment presented no significant gene sets in T47D and BT474 cell lines; however, this may not be completely reflective of the true biological situation – the lack of significant gene sets may be a result of more subtle expression. Ultimately, cancer is tremendously complex and the development of certain tumour subtypes may be the result of multiple gene sets working at lower levels but in synergy, rather than one overarching process. Although valuable, these types of analyses also ignore important biological information such as the positions of genes in pathways, the direction of interaction, and type of interactions with other genes, not to mention crosstalk with other pathways. Therefore, multiple stages of pathway analysis have been used as a foundation for *in vitro* work in this study, which intends to further investigate the *ASCL2* candidate gene in a laboratory environment (using cell lines) to shed light on its role in breast tumourigenesis.

Additionally, large variation in the candidate gene lists produced by each analysis tool was observed (Figure 3.4, Appendix 1); of 915 differentially expressed gene probes, only 10 genes were found to intersect between tools. These discrepancies may be because of innate differences in the tools themselves. Each tool uses unique mathematical methods and relies on different mathematical assumptions, meaning that results often lack concordance due to intrinsic complexities, and therefore small differences in approaches can lead to large discrepancies between outputs. This is exacerbated by the fact that there is no ubiquitous algorithmic method for ontology enrichment (Piccolo, & Frampton, 2016). Additionally, many knowledge-based computational biology tools rely on frequent updates; in this study, the tools used were regularly updated to align with constantly growing and newly published data, however, individual tools may be updated at different times. For example, the latest DAVID update was of 2016, in contrast to the current release of GSEA in 2018, and GO and PANTHER in 2019. As carried out in this work, Piccolo, & Frampton, (2016) suggest combining approaches to enhance reproducibility. In many reports, a single tool may be used for pathway analysis, whereas this study gained added value by comparing multiple tools (Alhamdoosh, *et al.*, 2017).

Although the ranking of genes could be considered as rudimentary in DAVID, GO and PANTHER analyses, GSEA was more intricate. In this work, the extreme variation gene list was input into DAVID, GO and PANTHER analysis tools as a whole, in contrast to the GSEA method which compared the gene expression of the gene list of each breast cancer cell line vs MCF10A cells individually. Similarly to DAVID, GO and PANTHER, GSEA was designed as a means of identifying groups of genes that are over-represented. However, GSEA takes into consideration the association of gene expression with a particular phenotype (in this case, a tumour); this has allowed researchers to better understand biological processes by observing the functional portraits of gene sets in tumours compared to normal controls (Subramanian, *et al.*, 2005).

For DAVID, PANTHER and GSEA analysis tools, the false discovery rate (FDR) was used for the assessment of significance rather than the nominal p value (the exception to this was GO analysis, where the Bonferroni test for multiple testing was used to ensure significance <0.05). This is because the nominal p value estimates statistical significance based on the enrichment scores of each gene set in isolation. In contrast, the FDR takes into account the size of the gene set and multiple hypothesis testing, and adjusts for these.

However, unlike the generally accepted 0.05 value of significance used in DAVID, GO and PANTHER tools, the FDR in GSEA uses 0.25 instead. The FDR is an estimation of validity – the probability that an enriched gene set is a false positive. Using a generous significance cut off of 0.25 has been designed to account for typical inconsistencies seen in expression datasets, and to avoid potentially noteworthy results from being disregarded. The Broad Institute GSEA user guidelines suggest this is practical for exploratory gene discovery as a basis for the further validation of candidate genes, and that these results (FDR<25%) are expected to be the most interesting for further research (Broad Institute, 2019; Mootha, *et al.*, 2003; Subramanian, *et al.*, 2005). However, in spite of this more lenient cut off, the GSEA method still produced the smallest number of candidate genes.

Aside from analytical limitations, the technical limitation of incomplete gene probe mapping and recognition is also a general and common issue associated with the use of pathway analysis tools (Khatri, *et al.*, 2012). The gene identifiers input into each pathway tool were not entirely recognised or mapped to a reference list, or the gene identifiers valid for each tool differed, for example, in DAVID analysis 650 Affymetrix IDs were mapped to the human reference list. This meant that the full list of 915 Affymetrix IDs generated from the extreme variation analysis were not analysed. However, this was due to the fact that some genes were mapped to multiple Affymetrix IDs, representing multiple transcript isoforms or different regions of the same gene (for example, *ASCL1* mapped to 209987_s_at and 209988_s_at probe identifiers). As well as this, some probe IDs could not be mapped to a HUGO gene symbol, nevertheless, it could be argued that the inability to map these genes to a reference list limits the genes' experimental possibilities; hence, those IDs that weren't captured within GO terms were unlikely to compromise the biological insight of this study, and it was decided that no genes of interest were lost of masked in this process. Despite this drawback of pathway analysis tools, a large sample of genes were recognised and analysed, and consistent patterns between the analyses could be demonstrated. Also, most candidate genes identified in Figure 3.4 do show heavy involvement in tumourigenesis when browsing the literature, so this issue was not thought to affect the integrity of the study.

Overall, the limitations discussed in this chapter were outweighed by the advantages held by these widely used pathway analysis tools. These tools were an excellent starting point as they are freely available, and therefore accessible to all researchers regardless of their financial means. These tools are relatively user-friendly (they don't require software installation, advanced coding or bioinformatics experience which could limit application), and due to their mainstream use and web-application, they can be used on any operating system with lots of resources (such as user-guidelines and tutorials) present to ensure correct usage. They also do well to reduce the complexity of large volumes of data whilst preserving explanatory power (Khatri, *et al.*, 2012), and their usage has also lead to the recognition of novel research possibilities due to unexpected associations between biological functions (Mathur, *et al.*, 2018). Therefore, these tools provide an informative framework from which to build an analysis model for the pursuit of candidate genes in breast cancer.

As well as the identification of a novel candidate gene for further exploration, an efficient, strategic and comprehensive pathway analysis pipeline was utilised (Figure 3.1). Many methods have been designed for the analysis of large transcriptomic datasets, but researching and choosing the correct tool or combination of tools can be time-consuming. This model for candidate gene selection utilised multiple tools to gain a thorough and consistent picture of the pathways and biological processes responsible for driving breast tumourigenesis, and therefore hopes to have generated a representative list of candidate genes (Figure 3.4). Despite the shortfalls discussed surrounding this type of analysis, these have been addressed by using multiple analyses to 'correct' for any intrinsic biases or inaccuracies in tools; hence genes seen in the cross-over between tools were considered to be most likely involved in the breast cancer cell lines examined. In light of this, the pipeline used in this study is considered to be advantageous as it combines numerous tools in parallel whilst maintaining simplicity to bridge the gap between computational scientists and biological scientists. As systems biology approaches for pathway analysis is an active and continuously growing area of research, this pipeline could be utilised as a foundation by researchers of any level of expertise, which could be adapted by the user to evolve with the field.

From a biological standpoint, findings from DAVID, GO and PANTHER analysis indicate enrichment in ontologies involving cell development, migration and motility, regulation of apoptosis and the EMT; which are considered to be trademarks of the Wnt signalling pathway. Adding weight to this was the GSEA results, exhibiting the association of MCF7 cells with genes relating to nervous system development; Wnt signalling, a highly conserved morphogenic signalling pathway, has also been recognised as crucial for nervous system formation, development and maintenance, as well as neural plasticity (Freese, *et al.*, 2010; Ille, & Sommer, 2005). This corresponds with the selection of *ASCL2,* as this gene has known involvement in the central nervous system (Liu, *et al.*, 2016), development of the neuroectoderm (Simionato, *et al.*, 2008) and has been recognised as a target of the Wnt pathway in colon and gastric cancer studies (Schuijers, *et al.*, 2015; Zhu, *et al.*, 2012; Tian, *et al.*, 2014; Basu, *et al.*, 2018). However, its role has not yet been established in breast cancer. Although this is only considered to be foundational evidence, the data presented in this chapter substantiates the hypothesis that via Wnt signalling, *ASCL2* may be a prominent force influencing Luminal A tumours with a similar molecular profile to MCF7 cells. It may also be hypothesised from this analysis that tumour cell migration may be heavily involved.

Overall, this chapter demonstrates that a thorough investigation of transcriptomic data for candidate gene selection need not be overly complicated, and therefore the pipeline (Figure 3.1) used in this study may be utilised by other researchers in the future. Given knowledge from the literature regarding *ASCL2* in cancer discussed in Section 1.2.3, and the ontologies and functions alluded to in pathway analysis, it is reasonable to suggest that *ASCL2* is likely to exercise its effect within the Wnt signalling pathway in breast cancer. The dysfunction of the Wnt pathway has a renowned effect on invasion and metastasis in cancer, a primary cause of cancer related deaths, therefore, novel targets of the Wnt/β-catenin pathway are igniting interest in breast cancer specifically, to benefit from newly emerging anti-metastatic drugs (Liang *et al.*, 2016).

To develop the *in silico* work presented in this chapter it was important to validate these findings *in vitro* in well-established cell lines. Currently, the use of RT-qPCR is the gold standard for confirming microarray data, to ensure that the same results can be observed using multiple techniques; this biological replication in a different set of samples provides greater power and confidence of conclusions. Therefore, to validate the microarray gene expression changes seen in *ASCL2*, RT-qPCR was employed (Ding, *et al.*, 2007; Horgan, & Kenny, 2011; Morey, *et al.*, 2006). Results demonstrated that the differential gene expression patterns of *ASCL2* in breast cancer cell lines mostly mirrored the transcriptomic data; in accordance with microarray data, the expression of this gene was significantly different between tumour and non-tumourigenic cells. *ASCL2* was differentially expressed across the breast cancer cell lines that were chosen to reflect the different breast cancer subtypes. *ASCL2* was overexpressed most highly in the MCF7 Luminal A cell line, as well as T47D (Luminal A) and SKBR3 (HER2+) cell lines; as analysis in the previous chapter lacked representation of the HER2 subtype, the SKBR3 cell line was added to RT-qPCR investigation here. To further validate this data, the RNA expression of *ASCL2* was checked using the Human Protein Atlas (Pontén, *et al.*, 2008); of 65 cell lines present in the database, T47D, MCF7 and SKBR3 cells ranked fifth, seventh and thirteenth respectively in terms of *ASCL2* gene expression, behind colorectal, placenta and myeloid cell lines, as expected from the literature.

In the next Chapter, the role of *ASCL2* will be investigated to improve understanding of its function and role in breast tumourigenesis. This will be actioned using MCF7 cells primarily (other cell lines may be used for comparison) with conventional methods to assess gene function.

Chapter IV

Investigating the potential role of ASCL2 in breast cancer

4.1. Introduction

The identification of molecular markers over the years has directed detection and treatment development, to combat the heterogeneity of breast cancer and thus improve mortality rates (Braune, *et al.*, 2018). Therefore, establishing robust and specific markers, or key genetic features, is crucial for the development of therapeutics or strategic management of the disease.

The Wnt signalling pathway has emerged as a major driver of breast tumour development, and its connection to cancer stem cells (CSCs) has received much attention. This is largely because the main challenges associated with cancer, such as metastasis, relapse, and therapy resistance, can be attributed to CSCs (Kazi, *et al.*, 2016). The deregulation of the Wnt pathway and overexpression of the genes present in this cascade, plays a key role in these factors, therefore therapeutic blockade of this pathway could be exploited in the future (Kazi, *et al.*, 2016). Although there are a large number of components and target genes present within the Wnt pathway expected to play a role in breast tumour development, there has been a lack of exploration in the past.

Evidence has demonstrated that levels of Wnt signalling varies between breast cancer subtypes and that the pathway is highly expressed in breast populations enriched with CSCs (Lamb, *et al.*, 2013). Additionally, deliberately activating the Wnt pathway in breast cell lines increases cell motility and migratory potential (Jang, *et al.*, 2015). Despite this, abnormal activation of this pathway in breast cancer is not completely understood, and the exact molecular biology remains unclear.

In Section 1.2.3, Achaete-scute Complex Like-2 (*ASCL2)* was broadly reviewed, and it was noted that this gene, a determinant of neuroblast fate, has been established as a Wnt target gene implicated in colorectal cancer (Zhu, *et al.*, 2012). Published studies in colon cancer have revealed that *ASCL2* functions as a transcriptional switch in the Wnt pathway, and accumulating evidence points to *ASCL2* as an interesting gene requiring further exploration.

Although *ASCL2* has been investigated in tumourigenesis (lung squamous cell, gastric, osteosarcoma), the majority of work has focused on colon cancer (Zhu, *et al.*, 2012; Zuo, *et al.*, 2018; Kwon, *et al.*, 2013; Basu, *et al.*, 2018). Though some studies have emerged more recently suggesting the expression of *ASCL2* in breast cancer (Conway, *et al.*, 2014; Wang, *et al.*, 2017; Xu, *et al.,* 2017), these studies have been rather incomprehensive, and within the scope of literature searches carried out for this thesis, have not been ongoing. Despite what is known about this gene in colon cancer, further research in this area is highly prospective and may yield lucrative results.

Thus, *ASCL2*, essential for the maintenance of intestinal stem cells and linked with the CSC-phenotype in other cancers, is a prime candidate for further investigation (van der Flier, *et al.*, 2009). Furthermore, as this gene is a transcription factor and known to be an important developmental gene (for example, in embryogenesis), this suggests that its biological function may have been underrated in the past (Guillemot, *et al.*, 1994; Schuijers, *et al.*, 2015); transcription factors account for approximately 20% of all oncogenes currently known, they can be attractive ‘drugable’ targets in cancer, and may be responsible for aberrant activation of other key regulatory genes (Lambert, *et al.*, 2018).

Given that poor survival of breast cancer patients is predominantly due metastasis and relapse, investigating genes, such as *ASCL2*, which could potentially be fundamental to these processes in tumorigenesis, are of significant interest. As discussed throughout this thesis, *ASCL2* is a target of the Wnt-signalling pathway, and linked with an oncogenic CSC-phenotype typically associated with metastasis, relapse, and therapy resistance in cancer. Thus, by this reasoning, it could be considered that expression levels of *ASCL2* may be increased in aggressive breast cancers or those enriched with CSCs, such as HER2+ or TNBC tumours. However, Xu *et al.,* (2017) found that high levels of ASCL2 were related to high tumour reoccurrence rate, yet found no correlation of ASCL2 expression between the subtypes of breast cancer, implying that the gene may play a central role in breast tumour development and progression, as opposed to a specific role, thereby affecting all or most subtypes. On a cellular level, it was hypothesised that overexpression of *ASCL2* may contribute to increased cellular migration and an aggressive phenotype in breast cancer cells.

As it stands, there is a scarcity of research conducted on this gene specifically relating to the function and clinical implications of this gene in breast cancer, so thus far, general assumptions have been drawn and applied from work on other cancers and biological processes, and built upon to form the infrastructure of this study. Given what is known about this gene (discussed in Section 1.2.3), current evidence suggests that exploring the role of *ASCL2* in breast tumourigenesis may yield interesting findings.

In light of this, the aim of this chapter was to investigate the expression and role of *ASCL2* in breast cancer cells. To assess the relationship between *ASCL2* expression and breast tumourigenesis *in vitro*, cells were subjected to siRNA transfection to knockdown expression of the gene; cellular processes commonly disrupted in cancer, such as proliferation, apoptosis, migration and possible association with stemness were then investigated.

4.2 Results

4.2.1 Selection of Breast Cancer Cell Lines for Functional Analysis

In Chapter 3, extreme variation analysis highlighted that *ASCL2* was expressed most highly in BT474, MCF7 and T47D cells respectively, with pathway enrichment analysis specifying the involvement of *ASCL2* in MCF7 cells. To validate this, *ASCL2* expression was quantified in six cell lines, demonstrating that *ASCL2* was expressed most highly in MCF7 cells, and also T47D and SKBR3 cell lines.

The combination of *in silico* and *in vitro* data was considered and thus the MCF7 (Luminal A) cell line was the focus of further experiments; as T47D demonstrated expression of *ASCL2*, this cell line was also chosen for further investigation and comparison as another Luminal A cell line. Considering all of the evidence, as well as the lack of HER2+ representation in the extreme variation datasets, SKBR3 cells were chosen for further investigation in additional to Luminal A cell lines.

4.2.2 Transfection Validation and Knockdown of ASCL2 in Breast Cancer Cell Lines

In order to investigate the possible impact of *ASCL2* overexpression seen in tumour cells (MCF7, T47D, SKBR3) compared to non-tumourigenic cells (MCF10A), as well as the involvement of *ASCL2* in the Wnt signalling pathway, the *ASCL2* gene was temporarily silenced using siRNA. It was hypothesised that temporarily silencing the gene expression of *ASCL2* may result in slowed tumour growth or function.

To validate genetic knockdown prior to data collection, control cells were transfected with TYE 563-labelled siRNA. These were visualised under a fluorescence microscope 24h post-transfection. The presence of a fluorescent signal in approximately 70% of nuclei was observed qualitatively, allowing quantitative measurement of knockdown efficiency to continue (Figure 4.1).

Knockdown of *ASCL2* was confirmed and quantified compared to non-targeting control siRNA (NC) samples using RT-qPCR. The relative mRNA expression of *ASCL2* in cell lines are shown in Figure 4.2. Detailed descriptive statistics are presented in Appendix 4. All cell lines achieved an average of approximately 70% knockdown of *ASCL2* (MCF7, 67%, T47D, 70%, SKBR3, 78%), which was used as a threshold to represent any true changes in tumour biology via functional laboratory investigation (Yang, *et al.*, 2011). However, transfection efficiency was seen to vary between biological replicates.

Functional exploration of *ASCL2* in breast cancer proceeded, and evaluated several parameters subsequent to gene silencing - cell viability, apoptosis, migration, and Wnt-target gene relationships.

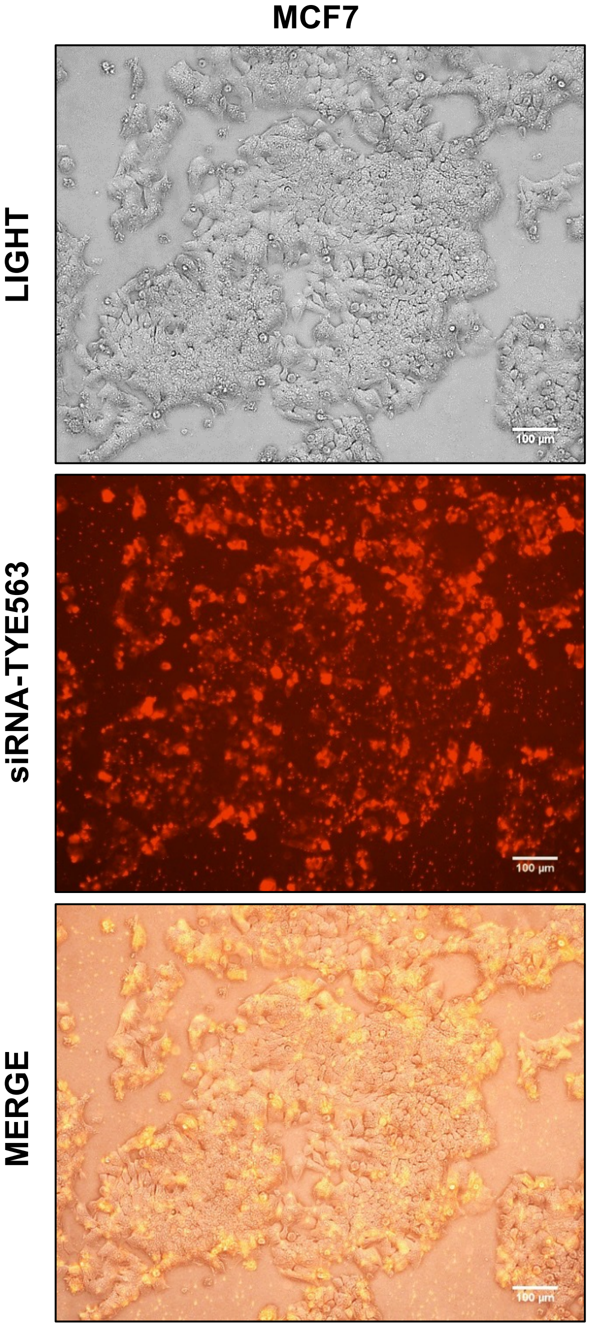


Figure 4.1. Cell lines transfected with TYE 563-labelled siRNA and visualised 24 hours post-transfection to qualitatively validate transfection success, with bright-field (light) and fluorescence, 4X objective. Once uptake and sufficient gene silencing was confirmed, RT-qPCR was completed. MCF7 cells pictured as a representative image.

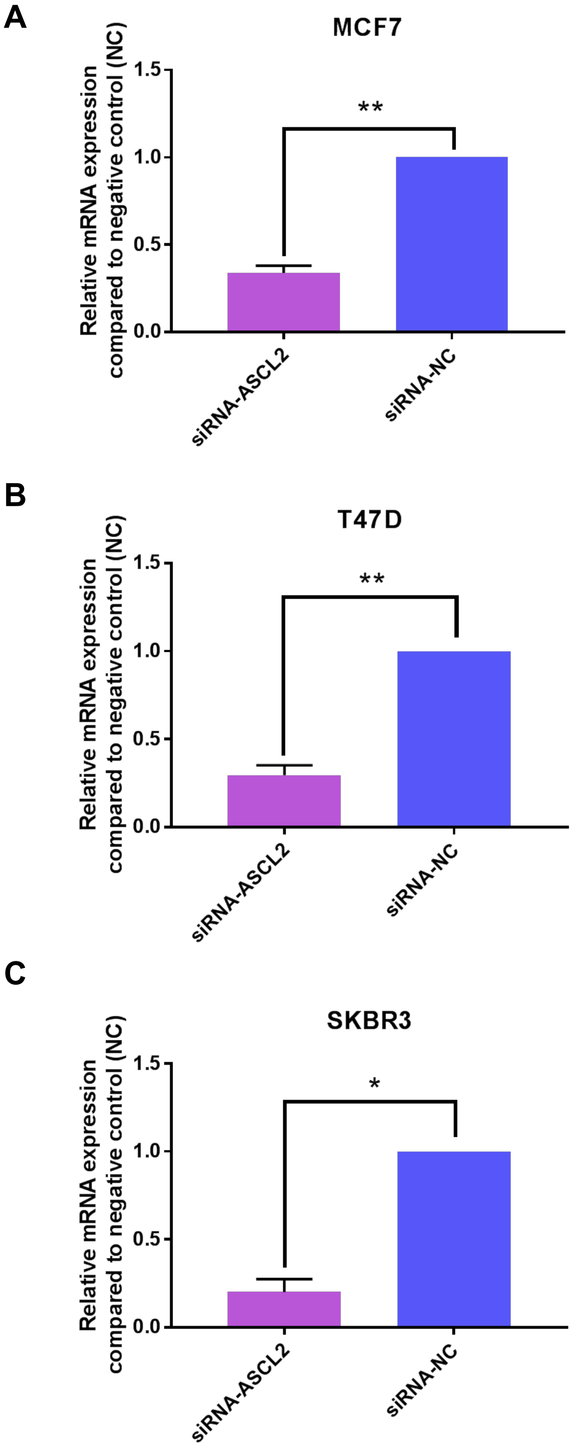


Figure 4.2. Validation of knockdown efficiency after 24h: mean expression (±SEM) of *ASCL2* in experimental samples relative to expression in negative control samples (relative fold change is shown), as measured by RT-qPCR after knockdown (MCF7, n=5, T47D, n=4, SKBR3, n=3, **p<0.0005, **p<0.0001*).

4.2.3 The Effect of ASCL2 on Cell Viability

Cell viability was measured using two common methods – the Alamar Blue assay and the Trypan Blue Exclusion Test. The Alamar Blue assay (a fluorometric method) was used as an indicator of cell health and metabolic activity by analysing the percentage difference in reduction of Alamar Blue in experimental samples compared to untreated controls. However, by nature of estimating cell viability based on metabolic activity, results may appear equivocal due to potential metabolic reprogramming in cancer cells (discussed further in Section 4.3). Thus, to validate these findings, the Trypan Blue assay, a dye exclusion method, was also used based on the principle that viable cells had intact membranes. Incubation times were optimised for these experiments.

Overall, there was no change in percentage reduction of Alamar Blue after *ASCL2* knockdown in all 3 cell lines, and to validate this, no change was observed in the percentage viability after Trypan Blue staining in MCF7 and SKBR3 cells (Figure 4.3 and 4.4). Unpaired t-tests showed no statistically significant differences in cell viability between *ASCL2*-silenced cells and negative controls for either method.

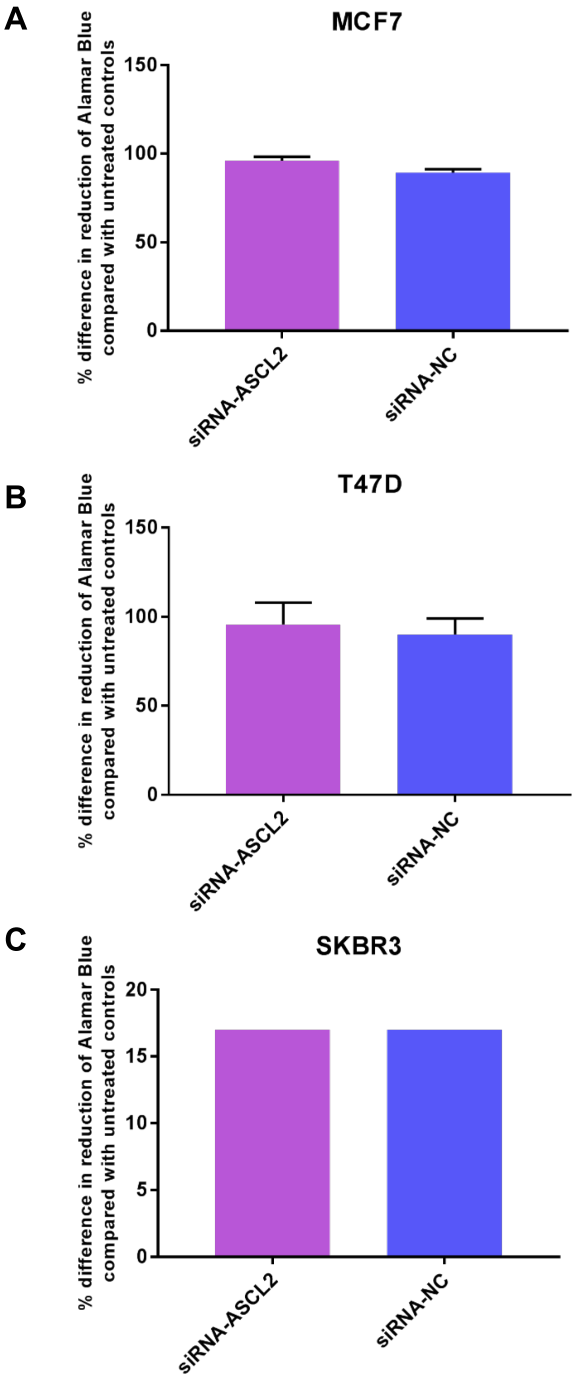


Figure 4.3. Alamar blue assay: percentage difference in reduction of Alamar Blue compared to untreated controls (mean ± SEM), across MCF7 (n=3), T47D (n=3) and SKBR3 (n=1) respectively. No changes were observed.

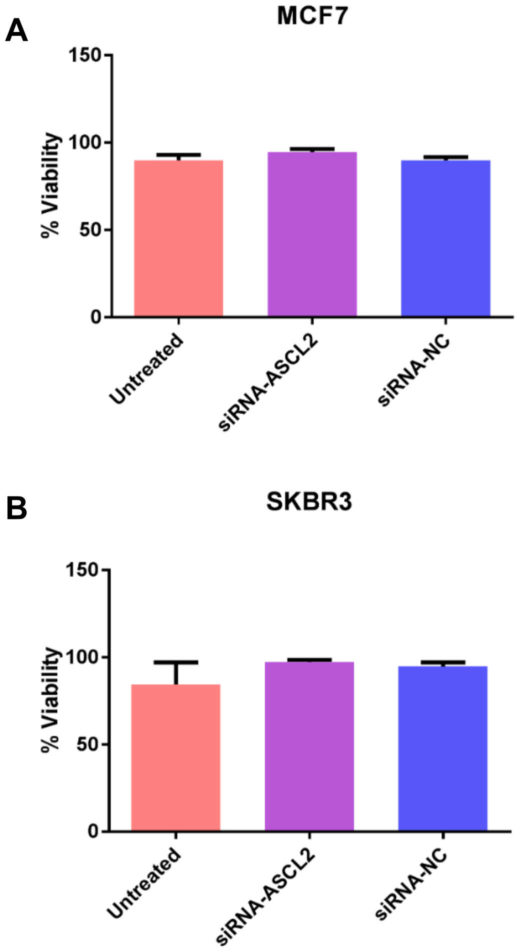


Figure 4.4. Trypan blue cell viability assay: percent viability of cells (mean ± SEM). No change in viability was observed. (MCF7, n=4, SKBR3, n=3).

4.2.4 The Effect of ASCL2 on Apoptosis

The Caspase-Glo 3/7 assay was used to determine apoptosis activity in cells. Apoptosis activity was judged based on caspase 3 and 7 activities in cells, as these are crucial in cell death induction and the promotion of apoptosis (Chen, *et al.*, 2016). In both MCF7 and SKBR3 cells, apoptosis was marginally elevated in siRNA-ASCL2 treated cells compared to untreated and siRNA-NC treated cells (Figure 4.5). However, since the role of *ASCL2* could not be definitively measured or statistically established between *ASCL2* silenced and negative control samples, there was insufficient evidence to confirm a role of *ASCL2* in apoptosis.

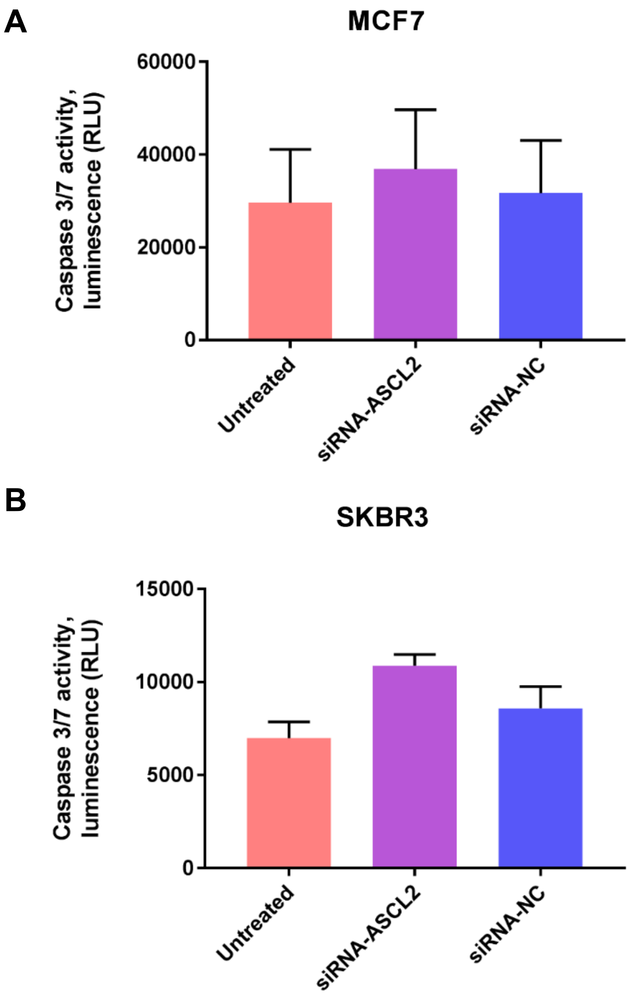


Figure 4.5. Caspase 3/7 assay: Measurement of caspase 3/7 activity representative of apoptosis shows a minor increase in cells with *ASCL2* knockdown (mean ± SEM, MCF7 n=4, SKBR3 n=3, $^*p<0.05$). The large error bar size and variation between replicates is likely due to variable transfection efficiency.

4.2.5 The Effect of ASCL2 on Wound Healing & Migration

Previous reports in colon and gastric cancers have shown that downregulation of *ASCL2* has reduced cellular invasion and migration *in vitro*, therefore suggesting that *ASCL2* plays a role in promoting the migration of tumour cells (Jubb, *et al.*, 2006; Tian, *et al.*, 2014; Zhu, *et al.*, 2012; Zuo, *et al.*, 2018). However, this link has not been explored in breast cancer.

Wound closure was measured over 48h in experimental (siRNA-ASCL2) and negative control (siRNA-NC) samples. The area of the wound (µm) was measured at 0h, 24h and 48h, and percentage closure was calculated between time points for each sample. Measurements and descriptive statistics are summarised in Appendix 4.

As can be seen from the microphotographs in Figure 4.6 and the data in Figure 4.7, there was a clear visual and numerical trend between the rates of wound closure in MCF7 cells where *ASCL2* expression had been silenced, compared to non-targeting siRNA control (NC) cells. Figure 4.6, highlights the observable differences in wound closure between MCF7 samples; it can be seen that cells migrated closed together, and the wound gap decreased more in control cells (untreated and siRNA-NC transfected) compared to siRNA-ASCL2 cells. To better observe complete closure of wounds, analysis would have benefited from data collection at a time point of 72 hours, however, this was limited by the transient nature of siRNA transfection. Figure 4.7 depicts the trend that wound closure was slowed in MCF7 cells after *ASCL2* silencing with a mean difference of 14.9% after 24h and 13.8% after 48h between experimental conditions (Figure 4.7 C). Figure 4.7 A and B both illustrate that the area of the scratch decreases more steadily in MCF7 cells after *ASCL2* silencing.

After 48 hours, no statistical significance was observed when comparing scratch area or percentage closure between *ASCL2* and NC transfected samples (Figure 4.7). These findings may be because, despite consistent trends, there was a large variation between biological replicates in this assay, which in turn had an effect on the determination of statistical significance. This was due to the semi-quantitative nature of manual measurements, and will be discussed further in Section 4.3. Nevertheless, this data was still sufficient for further inquiry that *ASCL2* may play a role in migration in MCF7/Luminal A breast cancer.

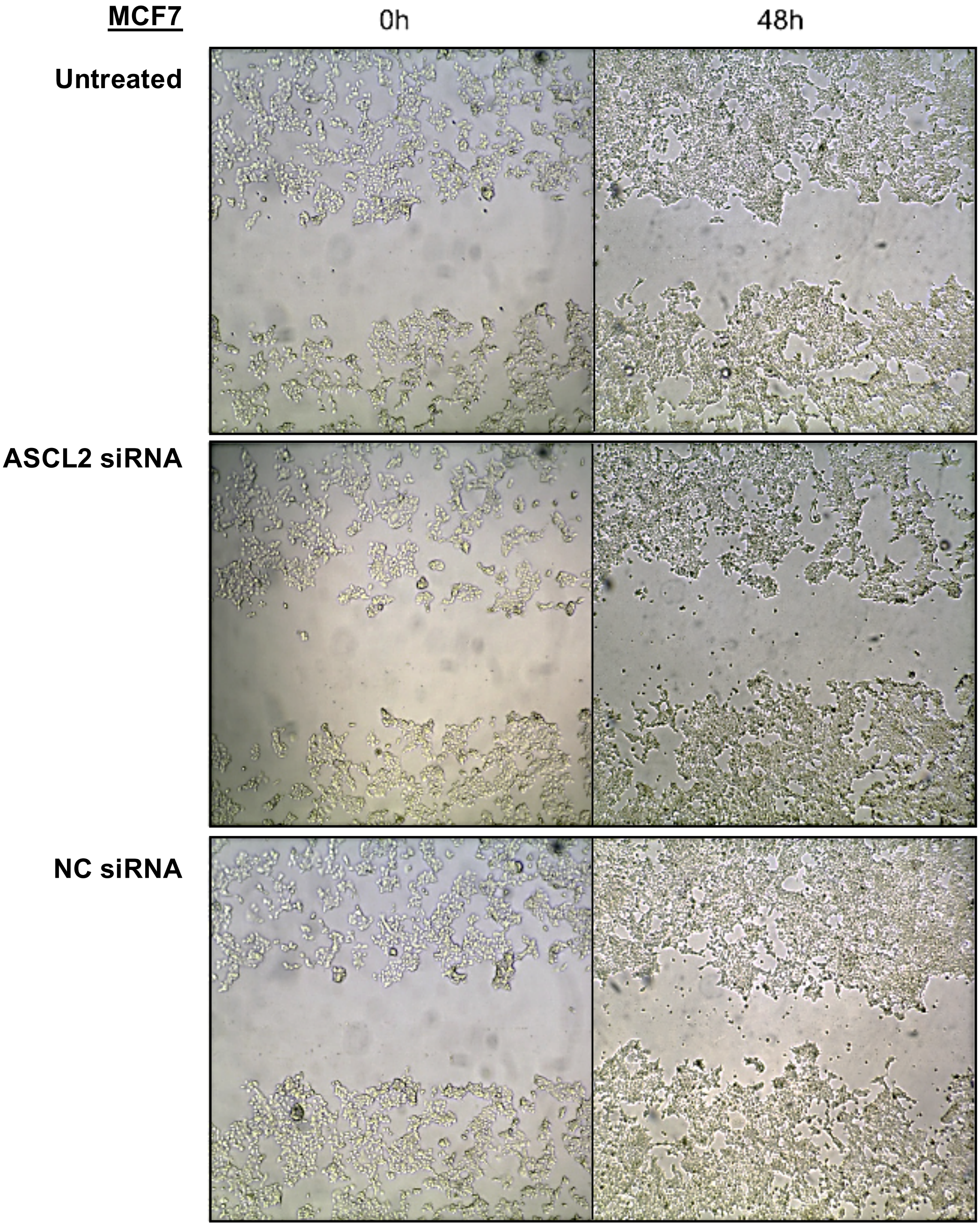


Figure 4.6. Microphotographs of scratches made in untreated cells, *ASCL2* transfected cells and NC transfected cells, at 0h and 48h, under a bright-field light microscope, x4 objective.

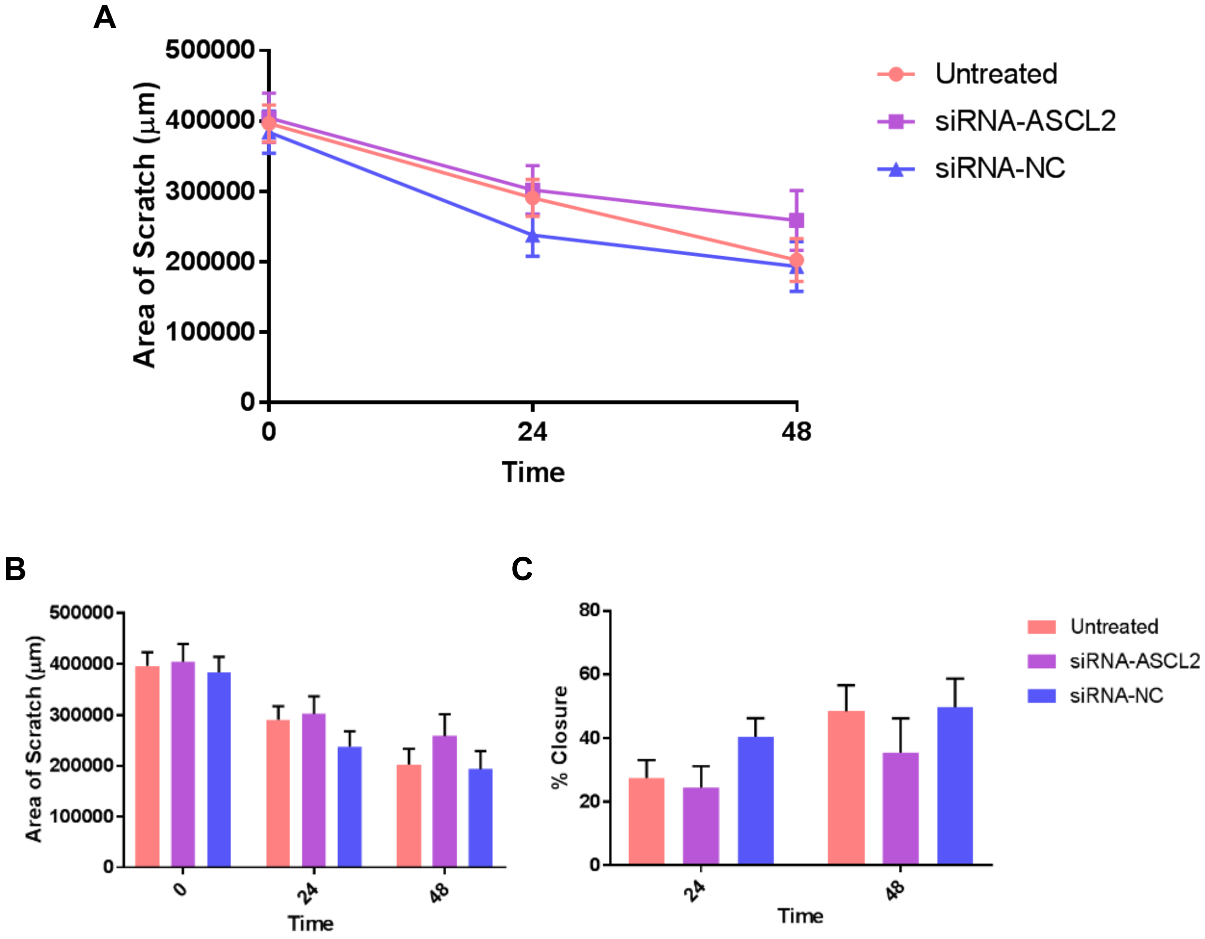


Figure 4.7. A & B. Rate of wound closure in MCF7 cells (n=6) over 24 h and 48h in *ASCL2* and NC transfected samples, as measured by the area of the scratch. Area of wound measured at each time point across replicates is presented as mean ± SEM. C. Percentage closure of scratches measured at 24h and 48h.

The same pattern was observed in Luminal A T47D cells in Figure 4.8 and 4.9. Although this appeared less apparent from the microphotographs in Figure 4.8 compared with MCF7 cells in Figure 4.6, this was seen in the quantitative data in Figure 4.9. The graphs in Figure 4.9 highlight that wound closure is slowed in T47D cells after *ASCL2* silencing with a mean difference between siRNA-ASCL2 and siRNA-NC transfected samples of 10.22% after 48h. This difference was statistically significant (p=0.03), signifying that *ASCL2* may contribute to enhanced cellular movement and migration.

Contrary to the trend observed in MCF7 and T47D cells, *ASCL2* did not appear to be associated with wound healing and migration in SKBR3 cells. Figure 4.10 and 4.11 highlights that there was no change in the rate of wound closure (area of scratch) or percentage closure over 48h.

A challenge faced using SKBR3 and T47D cell lines was achieving the correct cell density for analysis; these cells would have benefitted from being grown to a greater confluence to gain a better picture of the growth and movement of cells. However, although this was attempted, this resulted in greater cell death and thus obstruction of the wound area with detached cells. T47D data collection was also problematic due to high volumes of cell death post transfection, hence only two replicates were able to be analysed for data collected at 48h. Therefore, further optimisation is required for the future.

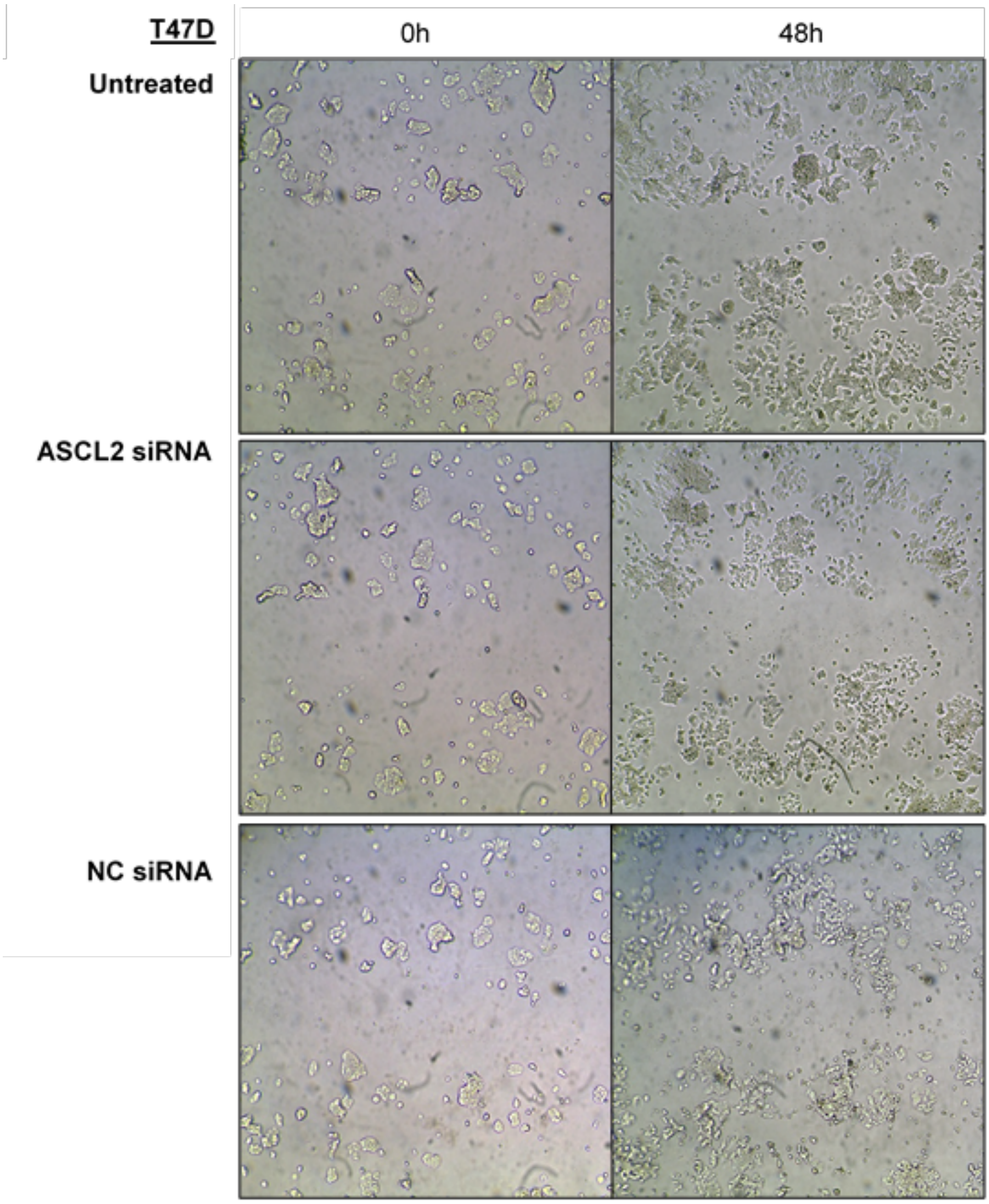


Figure 4.8. Microphotographs of scratches made in untreated cells, *ASCL2* transfected cells and NC transfected cells, at 0h and 48h, under a bright-field light microscope, x4 objective.

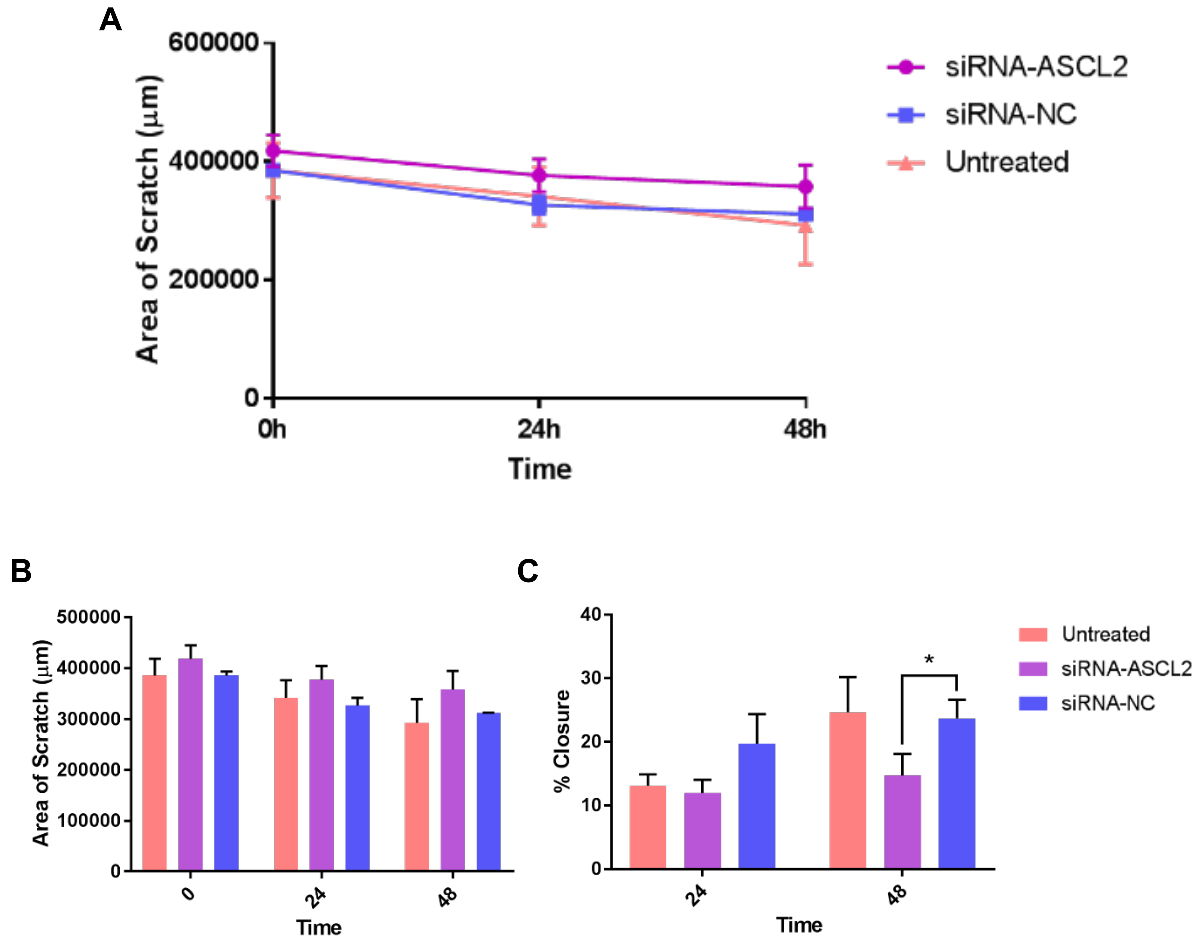


Figure 4.9. A & B. Rate of wound closure in T47D cells over 24h (n=3) and 48h (n=2) in *ASCL2* and NC transfected samples, as measured by the area of the scratch. Area of wound measured at each time point across replicates is presented as mean ± SEM. C. Percentage closure of scratches measured at 24h and 48h, **p<0.05*.

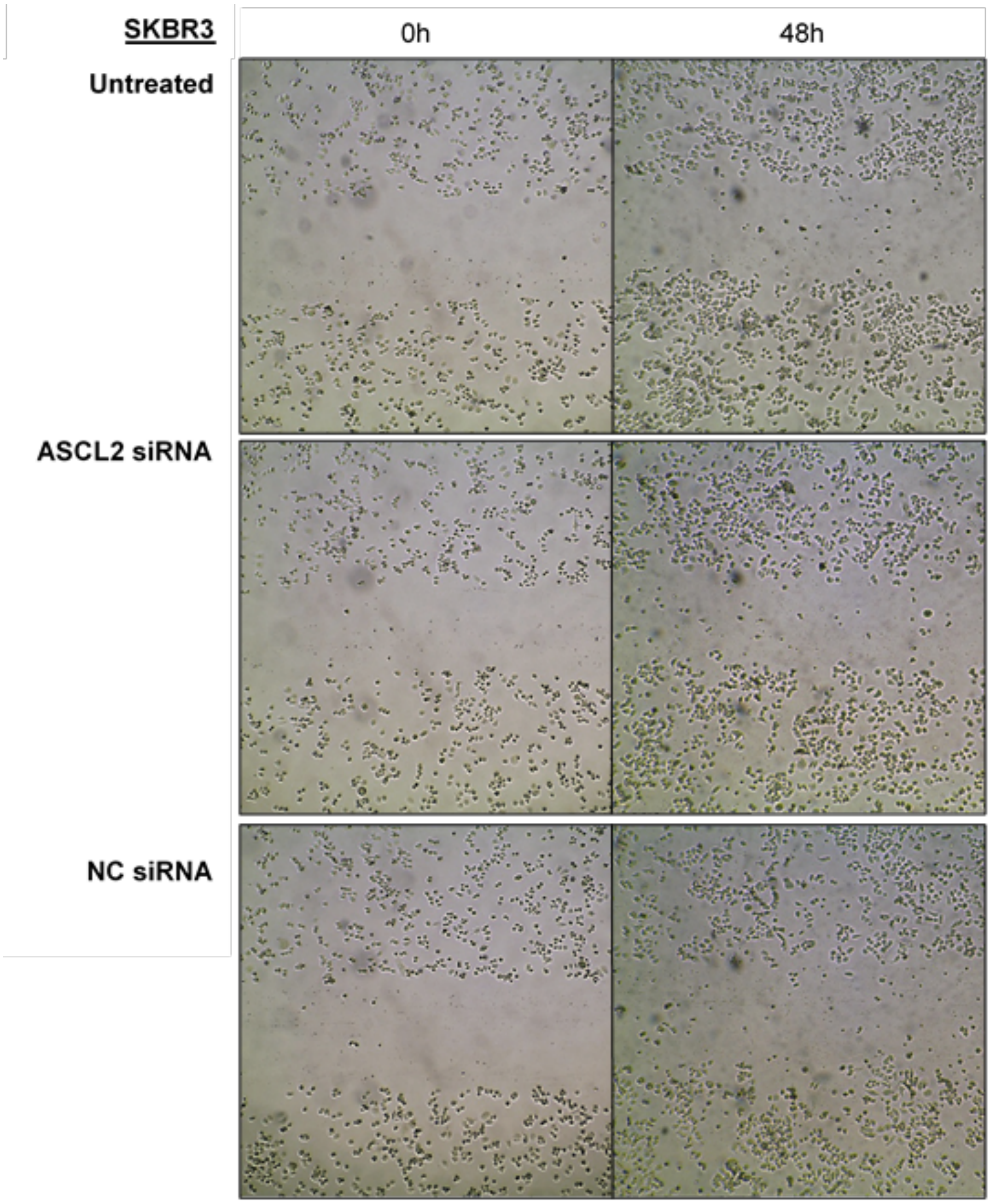


Figure 4.10. Microphotographs of scratches made in untreated cells, *ASCL2* transfected cells and NC transfected cells, at 0h and 48h, under a bright-field light microscope, x4 objective.

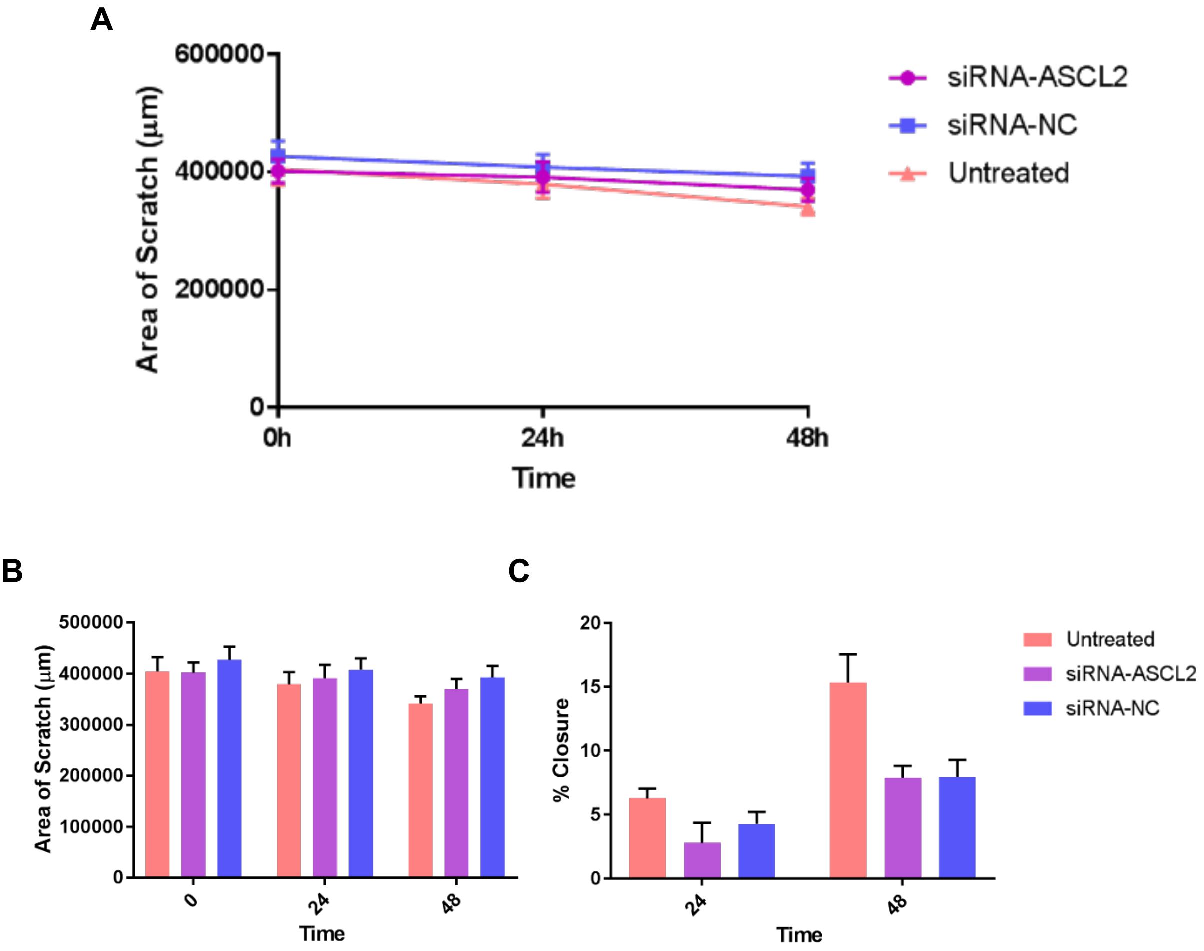


Figure 4.11. A & B. Rate of wound closure in SKBR3 cells (n=3) over 24h and 48h in *ASCL2* and NC transfected samples, as measured by the area of the scratch. Area of wound measured at each time point across replicates is presented as mean ± SEM. C. Percentage closure of scratches measured at 24h and 48h, **p<0.05*.

4.2.6 The Effect on Wnt-target Genes and 'Stemness' Markers Following ASCL2 Silencing

To further shed light on the relationship between *ASCL2* and the Wnt signalling pathway in breast cancer, and to confirm if *ASCL2* interference in breast cancer cells inhibited stem-like properties of cancer cells, 6 genes were selected for investigation. The genes *C-MYC, CCND1, CD44, CTNNB1, LGR5* and *SURV (BIRC5)* were selected based on their relation/role in Wnt signalling and their roles regarding their cellular behaviour (Chen, *et al.*, 2016; Kim, *et al.*, 2017; Wei, *et al.*, 2017; Zhu, *et al.*, 2012), intending to mirror the functional investigations in this study.

The differential expression of genes involved in Wnt signalling was measured by RT-qPCR. Figure 4.12 A shows a significant reduction in the gene expression of the markers *CD44, CTNNB1, LGR5* and *SURV* upon silencing of *ASCL2* in MCF7 cells; conversely, expression of the genes *C-MYC* and *CCND1* were not reduced after *ASCL2* silencing, but were instead significantly increased compared to the expression in negative control cells (siRNA-NC). Expression of the gene *SURV* also showed a marked decrease after *ASCL2* knockdown which may back the trend from the Caspase-Glo 3/7 assay, suggesting that *ASCL2* may play a role in the evasion of apoptosis in breast cancer.

Likewise, Figure 4.12 B illustrates a significant reduction in expression of all Wnt markers in T47D cells, following *ASCL2* silencing, however little change was observed in SKBR3 cells other than in *LGR5* in which an increased expression was observed, in contrast to Luminal A cells.

The greatest reduction of mRNA expression after *ASCL2* knockdown was seen in the genes *CD44* and *LGR5*, in MCF7 and T47D cells. These genes are widely known members of Wnt signalling and cancer stem cell markers, therefore this data, in line with the literature, suggests that the knockdown of *ASCL2* may inhibit the action of CSCs or may reduce breast cancer 'stemness'.

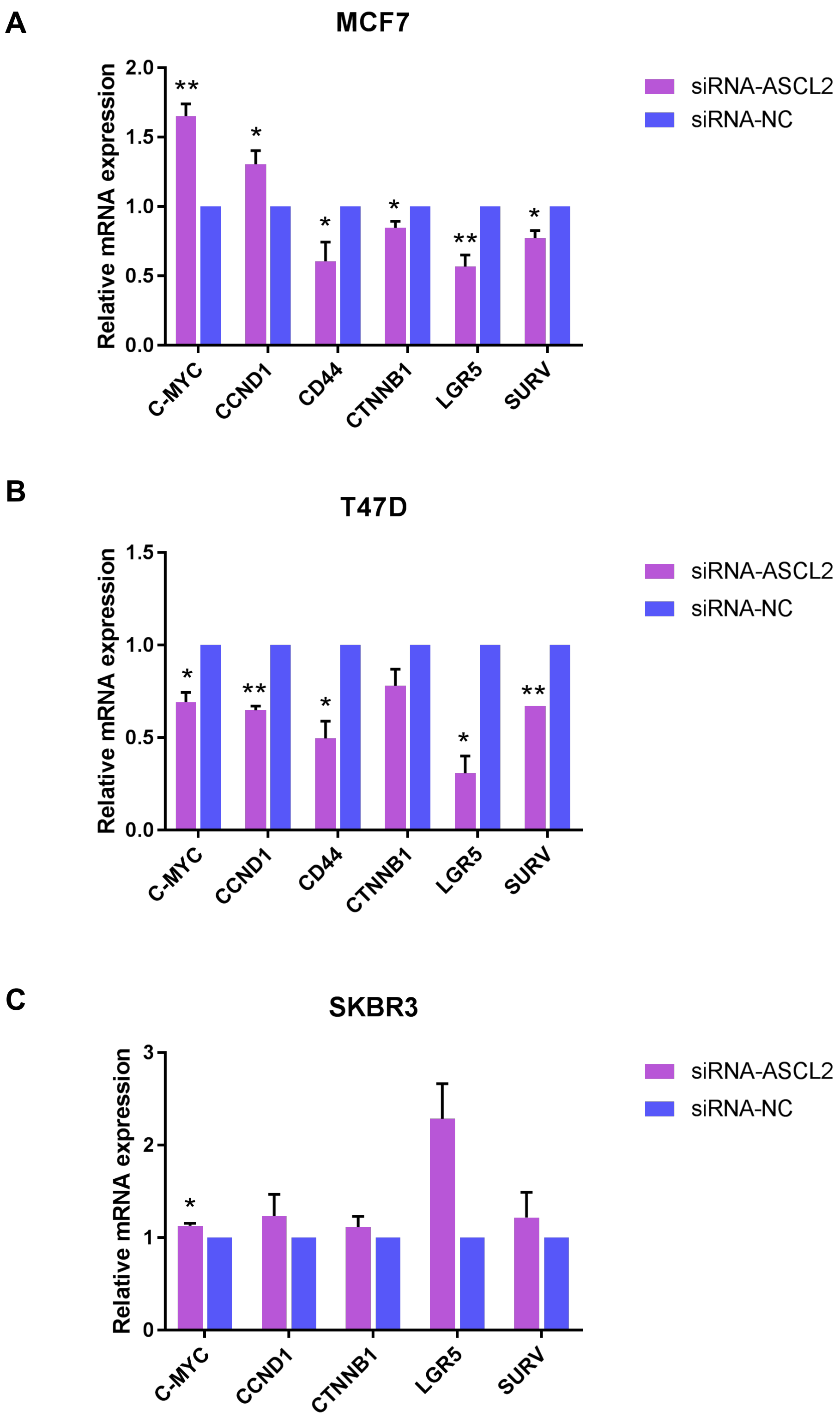


Figure 4.12. Changes in relative mRNA expression (fold change, mean ± SEM) of Wnt pathway markers and 'stemness' genes after silencing *ASCL2* compared to NC *(*p<0.05,**p<0.01)*. A. MCF7 cells (n=3), the genes *CD44, CTNNB1, LGR5*, and *SURV* exhibited significantly decreased gene expression when *ASCL2* was silenced. Conversely, *C-MYC* and *CCND1* showed significantly increased expression. B. T47D cells (n=2), the expression of all genes was reduced after *ASCL2* knockdown. C. SKBR3 cells (n=2), the expression of most genes appeared unchanged after *ASCL2* silencing; although *LGR5* showed overexpression, no significant difference was observed. Data for *CD44* was inconclusive.

4.2.7 Validation of ASCL2 Protein Knockdown in MCF7 Cells

To ensure the completeness and rigour of the study, the protein expression of *ASCL2* was examined to confirm that siRNA knockdown had been translated through to the protein level. Immunostaining of MCF7 cells was performed, and it was observed that ASCL2 was located within the cytoplasm upon fixation (Figure 4.13); however, protein knockdown could not be definitively confirmed. Possible reasons for this outcome are discussed in greater detail in Section 4.3

Although these results were not as anticipated, and functional changes are not usually to be expected without an observed change in protein levels after gene silencing, the previous experiments maintain validity, as a number of processes or feedback pathways may be affected that are currently unknown. The significance of these experiments, and this Chapter as a whole, in assessing the outcome of *ASCL2* knockdown in breast cancer cells is therefore justified by the scarcity of information currently available within the literature.

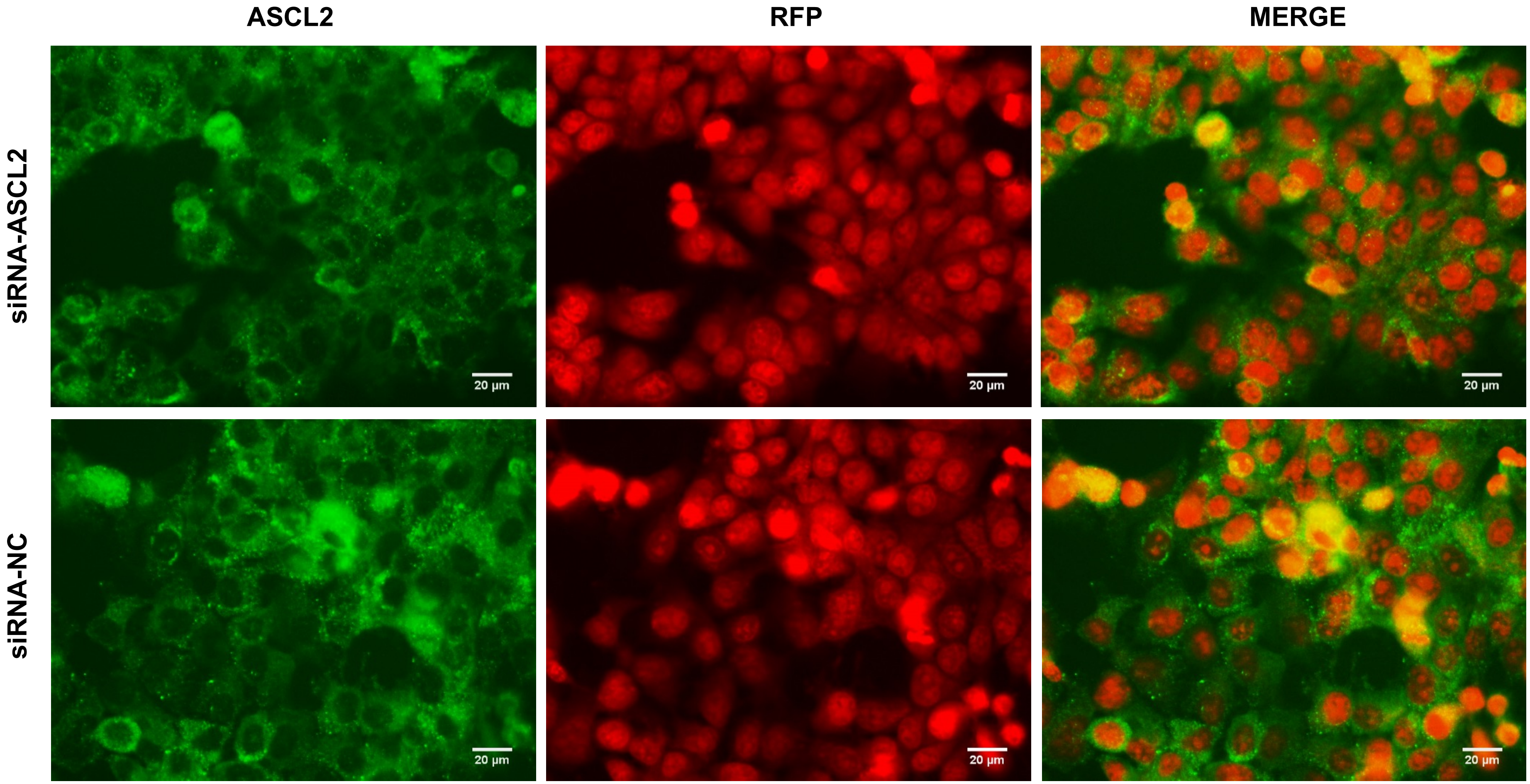


Figure 4.13. Immunofluorescence displaying no change in protein expression after *ASCL2* gene knockdown using siRNA in MCF7 cells (48h post transfection). It was observed that at the time of fixation, ASCL2 protein was predominantly located in the cytoplasm.

4.3 Discussion

Regarding siRNA experimental data collection in this study, untreated and non-targeting negative control (NC) duplex transfected cells were used as controls. For gene expression analysis and to assess statistical significance in functional assays, siRNA-NC treated cells were used for comparison. The siRNA-NC treated cells were favoured for experimental comparison over untransfected cells as this allowed any non-specific effects to be observed and distinguished from sequence-specific effects.

Prior to beginning functional investigation of *ASCL2* it was important to validate and optimise the process of gene knockdown. As shown in Appendix 2, a large body of time was devoted to optimising this procedure, as this formed the foundation of functional investigation. To ensure that the functional effects of gene knockdown could be observed, a knockdown of approximately 70% was required, and optimisation continued until this could be consistently achieved – unfortunately, despite best efforts, knockdown efficiency still varied between experiments, and this seems to be reflected in the variation seen between biological replicates. Many influencing factors on transfection efficiency were scrutinised in the process, including but not limited to cell seeding density, siRNA delivery, incubation times, RNA purity and extraction, and cDNA concentration. However, transfection efficiency can still vary dramatically from one experiment to another due to the inherent and unavoidable cellular toxicity caused by transfection methods and reagents; achieving the very fine balance between adequate siRNA delivery and toxicity was challenging (Biocompare, 2012). With regards to variable transfection efficiency, it is worthwhile mentioning that the difficulties faced to maintain knockdown consistency may be exacerbated by the intra-tumour heterogeneity of breast cancer cells. To combat this variability, multiple replicates were carried out, and this was kept in mind throughout the study when analysing data and drawing conclusions.

One prominent limitation of this analysis was the inability to demonstrate *ASCL2* knockdown at the protein level in MCF7 cells, despite seeing some phenotypic effects of the gene silencing. This was attempted using two methods – Western Blot (not presented) and immunostaining. Although protein expression of ASCL2 was observed in MCF7 cells using both techniques, a knockdown of protein expression could not be detected. However, in model disease systems, whereby a plethora of factors and circumstances are simultaneously at play at any given time, unanticipated results requiring troubleshooting are to be expected.

With regards to the observed lack of protein knockdown, changes in gene expression levels do not always accurately reflect the protein level, and in fact, this correlation is often weak (Maier *et al.*, 2009; Vaklavas *et al.*, 2020). this inconsistency between mRNA and protein levels has been investigated in transcriptomic and proteomic studies, whereby research has found that typically, cellular concentrations of proteins to their corresponding mRNAs only correlate by approximately 40% (Vogel & Marcotte, 2013).

A likely explanation may be due to the half-life of *ASCL2;* very stable proteins have a longer half-life, therefore may be highly transcribed or degraded at a slower rate, resulting in a longer time required for mRNA reduction to translate to the reduction of protein (Boettcher, & McManus, 2015). In this study, protein based experiments were measured after 48 hours, but may have benefitted from measurement after 72 hours or longer, as protein stability can vary from minutes to days. In contrast, the rate of mRNA degradation is restricted to a much tighter range, as mRNAs are generally less stable than their protein counterparts (half-life = 2.6-7 hours versus 46 hours) (Vogel & Marcotte, 2013). This may also be attributed to the rate at which mRNA is transcribed in comparison to the rate at which protein is translated, which has been estimated to be around 2 copies of mRNA versus dozens of the corresponding protein per hour (Vogel & Marcotte, 2013).

Overall, levels of cellular protein require the orchestration of a number of regulatory processes including the transcription, processing and degradation of mRNA, post-transcriptional events, and translation, localisation and modification of proteins (Vogel & Marcotte, 2013; Kim *et al.*, 2019). Therefore, a change in any one of these events may highly impact the abundance levels of mRNA or protein, especially in cancer cells, where the cross-talk of a number of pathways is likely to be over activated and highly dynamic in nature.

From an experimental standpoint, development of the stable expression of lentiviral shRNA would be more beneficial to circumvent the issue of transient transfection using siRNA, and eventually lead to full protein depletion (Boettcher,

& McManus, 2015). Another possible reason was that these experiments could have had a poor transfection efficiency prior to protein extraction for western blots; to improve certainty of knockdown in samples, the RNA (for RT-qPCR validation of efficiency) could have been simultaneously extracted while purifying out the corresponding protein for Western Blot analysis. Other reasons may include but are not limited to antibody specificity, or the presence of multiple different transcripts of the same gene (Bass, *et al.*, 2017). Antibody specificity was not validated by testing different antibodies prior to performing western blot or immunostaining analyses. Though, this will be done in the future to ensure that changes in protein expression can be accurately determined. By correctly identifying the most specific antibody, observed expression changes (even no expression change) can be more confidently relied upon to reflect the biological picture, and rules out the potential of technical or experimental artefacts. In addition to this, future work in the continuation of this study will also include the MCF10A cell line as comparison control cells in immunofluorescence analysis.

Previous data from Zhu, *et al.*, (2012) highlighted that *ASCL2* knockdown inhibits proliferation in colon cancer cells. Although in this study, the results from the Alamar Blue cell viability assay showed no change in viability (representing proliferation) after knockdown in breast cancer cell lines, the suitability of this assay for measurement of proliferation may be scrutinised. The main reason for this is that the assay was based on redox changes – this assumes that only live cells were metabolically active, where a reduction of Alamar Blue was proportional to the amount of 'viable' cells (Bio-Rad Laboratories, Inc, 2016). Hence, the reliance on metabolic changes in the cells representing changes in cell proliferation is problematic for a multitude of reasons.

Firstly, there are a large number of enzymes present in cells that may be responsible for the reduction of Alamar Blue; this makes it difficult to assess whether Alamar Blue reduction is due to genetic knockdown altering processes like cell death and proliferation, or just a change in cellular metabolism (Rampersad, 2012). Secondly, cells use a large amount of energy to push invasion and migration in cancer cells; assuming that *ASCL2* knockdown results in the decrease of migration (based on other findings in this chapter), it follows that this could be causing a great demand for energy in cells (Han, *et al.*, 2013). To address these shortfalls, the Trypan Blue dye exclusion assay was used to estimate cell viability based on membrane integrity, rather than metabolic activity; this confirmed Alamar Blue data and the indication that *ASCL2* did not enhance cell proliferation. However, using a more direct and sensitive method of quantifying cell proliferation may have been more valuable in this case, for example, measuring DNA synthesis using an 5-ethynyl-2′-deoxyuridine (EdU) staining assay and flow cytometry (Salic, & Mitchison, 2008). Jubb, *et al.*, (2006) suggested that ASCL2 promoted cell progression through the G2/M checkpoint of the cell cycle in intestinal neoplasia; investigating this using EdU staining would be particularly useful for further research into the role of *ASCL2* in breast cancer. Still, after 48 hours of gene knockdown in this study, no change in protein levels could be detected, as well as no effect on cell viability. Therefore, the relationship between *ASCL2* and cell viability could not be determined.

Results from the Caspase-Glo 3/7 assay indicated a slight elevation of apoptosis in siRNA-ASCL2 treated cells compared to the negative control cells in both MCF7 and SKBR3 cell lines. Although this insinuated a possible role of *ASCL2* in apoptosis in breast cancer, evidence was not sufficient to confirm this. As previously identified by Wang, *et al.*, (2018) in gastric cancer, it was anticipated that *ASCL2* silencing would increase apoptosis in breast cancer cells. In contrast to this notion, Zhongfeng, *et al.*, (2018) demonstrated that overexpression of ASCL2 increased levels of Caspase 3 in neuronal stem cells, thereby promoting apoptosis in these cells. Although, it is possible that *ASCL2* may act in a context dependent manner within different cells. Within this study, there were some technical reasons why *ASCL2* knockdown did not result in the significant increase of apoptosis in breast tumour cells.

Sundquist, *et al.*, (2006) demonstrated that caspase 3/7 activity was time dependent, and increased over the first 7 hours in their study. Therefore, they highly recommended that the optimal peak activity was determined consistently over a broad time period prior to data collection, as monitoring caspase activity prematurely or after its peak could result in a weakened signal leading to false conclusions. In this study, apoptosis was measured 24 hours post siRNA transfection, which may have been too late to capture peak caspase 3/7 activity; measuring caspase 3/7 activity at an earlier time may have shown a more pronounced effect of *ASCL2* knockdown on apoptosis. Although, collecting experimental data too soon post-transfection may result in taking measurements when cell activity is most perturbed by other reagents, for example, apoptotic markers may increase as a consequence of transfection-induced toxicity. However, this knowledge can be carried forward to improve this study and solidify knowledge of *ASCL2* in breast cancer.

It was also noted after experimental analysis that the functional caspase 3 gene product is absent in MCF7 cells, which may have been another contributing factor in the potential underestimation of apoptosis induced by *ASCL2* silencing (Jänicke, *et al.*, 1998; Jänicke, 2008; Sundquist, *et al.*, 2006). In light of this notion, it could be deliberated that the results seen in this chapter were not entirely demonstrative of the real cellular behaviour relating to apoptosis in MCF7 cells; with further investigation using a different method, such as flow cytometry, a greater change in apoptosis after *ASCL2* silencing may be observed. However, seeing as this is subject to debate, further research is required to confirm this.

The wound-healing scratch assay was chosen as a core technique to assess cell migration in an extremely convenient and economical manner. As discussed by Jonkman, *et al.*, (2014), the lack of a standardised method poses a challenge to researchers, however the guidelines outlined in this paper were followed within the means of resources available. Despite demonstrating clear trends towards decreased migration subsequent to *ASCL2* silencing in Luminal A cells, the difficulties in managing the variability of transfection efficiency between replica compromised reproducibility of the assay and posed difficulties for statistical analysis in MCF7 cells. Although every effort was made to address these reproducibility issues within the assay design, the manual nature of the assay meant that there were still a number of limitations.

For example, a challenge faced was ensuring that scratches made in each well for each replicate were the exact same width (the same tip pressure and angle) – however, this was combatted by calculating and comparing the percent closure of each well (making the scratch at the beginning and end of the assay, relative to each well individually). One method to combat this difficulty is the use of silicon inserts to ensure gap consistency. As well as this, there was a risk of manually imaging a different section of the scratch for each well. Although, every effort was made to ensure images were captured precisely for each time point; for example, marks were made on plate lids to ensure repeat images were as precise as possible, and a 4X objective was used to maximise the field of view of the wound area. However, the reproducibility of the assay could have been improved by using an automated live imaging digital camera and more specialised 2D image analysis software. Sampling at multiple positions with automated stage control would have also been useful to eliminate user-bias. Another complexity of this method was that while manually imaging, cells had to be removed from the incubator for each time interval, possibly impacting cell growth. Making the procedure automated by using an environment-controlled microscope would ensure optimal culture conditions were maintained consistently throughout the process, eliminating any effects on cell microenvironment and physiology (Gough, *et al.*, 2011; Johnston, *et al.*, 2014; Jonkman, *et al.*, 2014).

Other reproducibility issues could have been attributed to possible variation in transfection efficiency for each biological replicate – for example, gene silencing of *ASCL2* with a transfection efficiency of 65% compared to 80% could have had profound differences on the functional effect. This was likely to be causing the large variation in scratch migration between biological replicates in MCF7 cells, which in turn affected statistical significance in this study (despite the same trends across replicates observed, that *ASCL2* silencing decreased migration compared to negative controls). In spite of best efforts of quality control and experimental handling, fluctuating transfection efficiency between biological replicates could have given rise to the large variation in the closure of the wound.

Examining the raw data in Appendix 4, although MCF7 cells closed on average a greater percentage between *ASCL2* and NC transfected cells compared to the same conditions in T47D cells, the variation of data between biological replicates was more apparent in MCF7 cells. For example, in MCF7 cells, the mean percentage closure after 48h between the two conditions (siRNA-ASCL2 vs siRNA-NC) was 13.8%, with a SEM of ±11.53. However, in T47D cells, the difference between the mean percentage closures of each condition was 10.22% ± 3.108. This is reflected in the coefficient of determination (R^2) values of 0.1332 and 0.7299 in MCF7 and T47D cells respectively. Therefore, despite 6 replicates in MCF7 cells vs 2 replicates in T47D cells, and a clear graphical trend, the large variation has affected the statistical significance in MCF7 cells. It can therefore be assumed that repeating this work with more sophisticated technology would yield a sounder conclusion and confirm that *ASCL2* enhances migration in Luminal A breast cancers.

With regards to the SKBR3 (HER2+) cell line, *ASCL2* silencing did not influence cell migration. Untreated cells appeared to close the wound gap much quicker than transfected cells, which may have been attributed to the toxicity of the transfection process. As there was no change observed in cellular behaviour in SKBR3 cells, this may suggest a subtype specific effect of *ASCL2*. This is contrary to work published by Xu, *et al.*, (2017), who claimed to find no differences in *ASCL2* gene expression between breast cancer subtypes in patient tissue samples, but suggest the use of *ASCL2* as a prognostic marker in patients. Although it could be argued that patient tissue samples are much more representative of tumours than cell lines, cell lines are comparatively less susceptible to high heterogeneity and provide a relatively stable genetic basis for exploration; additionally, the differential expression of *ASCL2* has been verified at multiple levels in this study (transcriptomic pathway analysis and *in vitro* analysis). Therefore, it is plausible to say that *ASCL2* is expressed in breast cancer cell lines in a subtype-specific manner.

Overall, this analysis has presented some evidence that *ASCL2* may be involved in the collective migration of Luminal A breast cancer cells, however, further assessment of migration would improve the reliability of these results. This could be done using previously mentioned silicon gap inserts, or using the Boyden chamber assay, or Dunn cell chambers, both of which use a chemical concentration gradient to follow the movement and migration of cells; however, both of these would require time-lapse recording equipment.

It was summarised in Section 1.2.3 that *ASCL2* controls intestinal stem cell fate via the downstream effects of Wnt signalling (van der Flier, *et al.*, 2009; Zhu, *et al.*, 2012). In colon cancer, *ASCL2* regulates cell self-renewal and plasticity through EMT, and selective blockade of *ASCL2* has been suggested to contribute to the reversal of EMT and thus cancer progression (Tian, *et al.*, 2014). With this in mind, its role in the 'stemness' of colon cancer has received much attention within the literature, yet has been neglected in breast cancer. In order to highlight this link between *ASCL2* and stemness in breast cancer, and its activity within the Wnt pathway, genes with Wnt involvement were examined after *ASCL2* silencing.

Multiple genes were selected in the hope to broadly represent various cancer hallmarks and functional parameters relating to cancer within the Wnt signalling pathway. The genes *CD44, LGR5, CTNNB1* and *C-MYC* were chosen as Wnt signalling markers of stemness (Kim, *et al.*, 2017; Yang, *et al.*, 2015a; Jang, *et al.*, 2015; Zhao, *et al.*, 2017). *SURV* was chosen as a marker of apoptosis (Chen, *et al.*, 2016), and *CCND1* was chosen on the basis of being widely overexpressed in breast cancer and due to its role in cell cycle regulation and progression (Roy, & Thompson, 2006). In primitive terms, these have been chosen to complement the investigation of migration (wound-healing), apoptosis and proliferation respectively, however it is acknowledged that these genes are all multifaceted in their functions.

The data from this study indicated that upon *ASCL2* silencing, the expression of stemness-related genes *CD44, CTNNB1* and *LGR5* decreased in both MCF7 and T47D cells. This not only supported the findings from the wound-healing assay in this study, but also supported the idea that *ASCL2* may be a key gene involved in breast cancer stemness via the Wnt signalling pathway. In T47D cells, the expression of *C-MYC* and *CCND1* was also reduced after *ASCL2* silencing, however the opposite trend was observed in MCF7 cells. Although *C-MYC* has been shown to mediate cancer stem cells via sustained activity in triple-negative breast cancers (Yin, *et al.*, 2017), amongst other tumour types, the observation of overexpression after *ASCL2* silencing may be due to another compensatory mechanism, as *C-MYC* also functions to regulate, for example, cell growth and proliferation. As both *C-MYC* and *CCND1* are ubiquitous in their functions, a single trait cannot be attributed or measured in these genes. Therefore, in this case, the mechanism by which *C-MYC* and *CCND1* functions in stemness may be context dependent, or may only be present within a small subpopulation of the MCF7 tumour cells (Yin, *et al.*, 2017).

The expression of the gene SURV was also measured as a means to estimate the effect of *ASCL2* on apoptosis. This gene is also multifunctional but is principally recognised for controlling cell division and the inhibition of apoptosis, and is associated with therapy resistance and poor prognosis in breast cancer (Chen, *et al.*, 2016). Downregulation of *SURV* subsequent to *ASCL2* silencing may add weight to the trend observed in the Caspase-Glo 3/7 assay, signifying that *ASCL2* may contribute to the evasion of apoptosis in breast cancer cells. However, the mechanisms of apoptosis involving *SURV* in tumourigenesis is highly complex, therefore this can only be hypothesised rather than confirmed in this study – to elaborate on this, other apoptosis associated genes such as *BCL-2* could also be measured after *ASCL2* knockdown.

Overall, this data suggested that *ASCL2* silencing may have inhibited the action of cancer stem cells or may have reduced breast cancer ‘stemness’. Results demonstrated that *ASCL2* did have an effect on the Wnt target genes in breast cancer, and therefore may work within this pathway to drive tumourigenesis. To strengthen this part of the study, more stemness markers such as *Oct4, Sox2* and *CD133* could be added for investigation (Zhu, *et al.*, 2012)

To summarise the findings of this chapter, as demonstrated, sufficient gene knockdown was achieved using anti-*ASCL2* siRNAs in all cell lines. *In vitro* functional assays suggested that reducing the expression of this gene had the potential to lessen cellular migration within MCF7/T47D Luminal A subtypes, and may contribute to the evasion of apoptosis in these breast cancer cells (however, this requires additional confirmation). Conversely, it appeared that *ASCL2* gene knockdown did not impact cell proliferation in the same way. The data described also demonstrated that *ASCL2* silencing resulted in reduced expression of Wnt signalling associated genes; the effect of silencing on *CTNNB1* (the gene encoding the β-catenin protein, a key player in Wnt signalling), highlighted that *ASCL2* was likely to exercise its effect on breast tumourignesis, as confirmed in colon cancer via the action of the canonical Wnt pathway. In particular, knockdown of *ASCL2* resulted in the greatest reduction of the stemness marker genes *CD44* and *LGR5*, as well as the apoptosis gene *SURV*, thus confirming the trends seen in functional assays and echoing evidence published in colon cancer studies (Zhu, *et al.*, 2012). Ultimately, this data provides a sufficient body of evidence that *ASCL2* is involved in breast tumourigenesis, and although mechanisms have not yet been fully elucidated, provides a basis for which further hypotheses and research can be conducted.

Chapter V

Assessment of molecular features and survival outcomes associated with ASCL2 in patient breast tumours, to evaluate potential as a clinical or prognostic marker

5.1 Introduction

In recent years, considerable efforts have been made to explore the gene expression profiles underlying the distinctive breast cancer subtypes, as well as possible markers associated with poorer survival, therapeutic sensitivity and clinical outcomes (Dai, *et al.*, 2015; Prat, & Perou, 2011; Reaz, *et al.*, 2018). The utility of such genetic markers has been demonstrated in the clinic for some time, for example, the expression of the proliferative gene, Ki67, has been used as a determinant of chemotherapy response and as an indicator of prognosis (Yerushalmi, *et al.*, 2010). Additionally, the identification of novel molecular markers for predicting aggressive phenotypes could provide new opportunities for future therapy development, or the personalised management of tumours across patients (Reaz, *et al.*, 2018).

However, as knowledge and understanding of the genetic diversity of breast cancer advances, the current intrinsic subtype classification has come under scrutiny (Russnes, *et al.*, 2017). In light of this, a more extensive and integrated classification system was proposed termed the integrative clusters. These 10 clusters were each linked with likely molecular markers, variable causal biology, and thus discrete clinical outcomes in which personalised management and treatment approaches could be tailored towards (Dawson, *et al.*, 2013). Still, inter- and intra-tumour heterogeneity presents important challenges to researchers due to the resultant variations in molecular and clinical characteristics (Bedard, *et al.*, 2013). In this respect, the pursuit of a variety of functional oncogenic markers to develop biological understanding, improve cancer risk models, and enhance marker-based therapies are of significant interest within the field of precision medicine for breast cancer (Kalia, 2015).

The recent application of high-throughput technologies for gene expression profiling, and the curation of publically available repositories, means it is now possible for large published studies to share clinically relevant datasets. Researchers may use these perpetually for candidate gene investigation, to yield insights into novel molecular markers and targets, and assist in the clinical translatability of oncology research (Cheng, *et al.*, 2015; Y. Yang, *et al.*, 2015b). One such resource designed to facilitate gene exploration and discovery is the cBioPortal, an open-access online platform, providing access to large-scale datasets from 246 cancer studies (Cerami, *et al.*, 2012; Gao, *et al.*, 2013).

Within the literature, *ASCL2* has been implicated in a number of cancers, alluding to the notion that it may be an attractive gene worth exploring as a novel marker in breast cancer. Previous studies have reported evidence suggesting the role of *ASCL2* as a prognostic indicator in tumours; high ASCL2 protein expression was associated with advanced tumour stage and poorer differentiation status in lung squamous cell carcinoma, and poorer overall and metastasis-free survival in osteosarcoma (Hu, *et al.*, 2015; Liu, *et al.*, 2016). However, little is known about the role of ASCL2 in breast cancer. One of the few studies exploring the relationship between ASCL2 and clinical outcomes in breast cancer was that by Xu, *et al.*, (2017); this study used semi-quantitative immunohistochemical staining in a small cohort of patients, suggesting that increased ASCL2 protein expression was correlated with poorer survival and relapse.

In the present study, previous analysis of cell line data indicated that *ASCL2* expression was elevated in some breast tumour cells vs non-tumourigenic cells (Chapters 3 and 4). To develop these findings, it was hypothesised whether elevated *ASCL2* expression may be correlated with advanced or aggressive breast tumours, as well as poorer patient survival. Further, the association of *ASCL2* with the distinct intrinsic subtypes or integrative clusters of breast cancer, hence its specificity as a potential marker, has not yet been examined. Exploring such relationships may be used to identify promising avenues for further research, and enable more focussed study of *ASCL2*.

The aim of this chapter was therefore to determine the suitability of *ASCL2* as a clinical or prognostic marker. In order to execute this, the expression levels of *ASCL2* were analysed among a large cohort of patient breast tumours (n=1904) from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) study (Curtis, *et al.*, 2012), obtained via the cBioportal (Gao, *et al.*, 2013). In this Chapter, the association of *ASCL2* mRNA expression with clinical features such as, subtype, receptor status, age of onset, and overall survival was examined to assess the prognostic significance of *ASCL2 i*n breast cancer.

5.2 Results

The work presented in Chapter 3 used *in silico* methods to identify *ASCL2* as a potential tumourigenic candidate gene in breast cancer, primarily within the Luminal A subtype. This was followed using cell lines *in vitro* (Chapter 4), demonstrating a functional role of *ASCL2* in cellular migration, hypothetically within the Wnt signalling pathway.

To accompany the work in cell lines, gene expression data from patient samples were explored to examine the clinical impact of *ASCL2* and association between clinicopathologic features. Data obtained from the METABRIC study (Section 2.2) (Curtis, *et al.*, 2012; Pereira, *et al.*, 2016) was accessed through the cBioPortal web application (Gao *et al.,* 2013). Expression level Z-score data of ±2 was used as a threshold to classify clinical breast cancer cases into 3 groups according to the expression level of *ASCL2*: upregulated (overexpressed), downregulated and unaltered. The METABRIC dataset contained 2509 samples from primary tumours, of which, gene expression data for *ASCL2* was recorded in 1904 patients. Of these, 3.3% (82) of patient samples were shown to overexpress *ASCL2* with only one case (0.04%) shown to downregulate *ASCL2*.

Descriptive statistics for all clinical parameters and corresponding *ASCL2* expression in the METABRIC cohort is presented in Table 5.1 and 5.2, and can also be found in Appendix 5.

Table 5.1. Population distribution of the METABRIC study, and association between *ASCL2* expression and clinicopathologic characteristics of breast

METABRIC Study	Total	Downregulated (≤ -2)	Unaltered (-2 to +2)	Upregulated (≥ +2)	P value
		N (%)	N (%)	N (%)	
Age	1,904	1	1821	82	*0.497*
Mean (Years)		75.13	61.05	61.75	
Range		-	21.93, 96.29	32.99, 87.18	
Survival	1,904	1	1821	82	*0.311*
Mean (Months)		58.67	125.53	114.70	
Range		-	0.00, 355.20	9.60, 297.23	
Stage					*0.224*
0	4	-	4 (0.3)	-	
1	475	-	460 (34.2)	15 (25.9)	
2	800	-	766 (57)	34 (58.6)	
3	115	-	106 (7.9)	9 (15.5)	
4	9	-	9 (15.5)	-	
Grade					*0.45*
1	165	-	160 (9.1)	5 (6.3)	
2	740	-	711 (40.6)	29 (36.7)	
3	927	-	882 (50.3)	45 (57)	
4	0	-	-	-	
PAM 50 Subtype					*<0.001**
Luminal A	679	1 (100)	660 (36.4)	18 (22)	
Luminal B	461	-	434 (23.9)	27 (32.9)	
HER2 +	220	-	193 (10.6)	27 (32.9)	
Basal	199	-	192 (10.6)	7 (8.5)	
Claudin-Low	199	-	199 (11)	-	
Normal-like	140	-	137 (7.5)	3 (3.7)	
Integrative Cluster					*0.010**
1	132	-	126 (6.9)	6 (7.3)	
2	72	-	68 (3.7)	4 (4.9)	
3	282	-	275 (15.1)	7 (8.5)	
4ER-	244	-	238 (13.1)	6 (7.3)	
4ER+	74	-	71 (3.9)	3 (3.7)	
5	184	-	163 (9)	21 (25.6)	
6	84	-	81 (4.4)	3 (3.7)	
7	182	-	176 (9.7)	6 (7.3)	
8	288	1 (100)	279 (15.3)	9 (11)	
9	142	-	132 (7.2)	10 (12.2)	
10	219	-	212 (11.6)	7 (8.5)	
ER Status					*0.838*
Positive	1458	1 (100)	1396 (76.7)	62 (75.6)	
Negative	445	-	425 (23.3)	20 (24.4)	
PR Status					*0.222*
Positive	1008	1 (100)	971 (53.3)	37 (45.1)	
Negative	895	-	850 (46.7)	45 (54.9)	
HER2 Status					*0.001**
Positive	236	-	215 (11.8)	21 (25.6)	
Negative	1668	1 (100)	1606 (88.2)	61 (74.4)	

* For each data type the total number of cases may differ due to missing values or incomplete data within the METABRIC dataset.

Table 5.2. Association between *ASCL2* expression (Z-score) and clinicopathologic characteristics of breast cancers in the METABRIC dataset.

ASCL2 Z-score	Downregulated (≤ -2)		Unaltered (-2 to +2)			Upregulated (≥+2)		
	N	Mean	N	Mean	95% CI	N	Mean	95% CI
Stage								
0	-	-	4	0.21	-2.57, 2.99	-	-	-
1	-	-	460	-0.2	-0.28, -0.13	15	2.43	2.25, 2.60
2	-	-	766	-0.04	-0.10, 0.03	34	2.40	2.29, 2.51
3	-	-	106	0.18	0.02, 0.34	9	2.43	2.22, 2.65
4	-	-	9	-0.29	-1.04, 0.45	-	-	-
Grade								
1	-	-	160	-0.28	-0.41, -0.15	5	2.64	2.27, 3.00
2	-	-	711	-0.24	-0.30, -0.18	29	2.45	2.32, 2.57
3	-	-	882	0.07	0.01,0.13	45	2.37	2.28, 2.46
4	-	-	-	-	-	-	-	-
PAM 50 Subtype								
Luminal A	1	-2.08	660	-0.19	-0.25, -0.13	18	2.35	2.24, 2.45
Luminal B	-	-	434	-0.34	-0.42, -0.26	27	2.49	2.35, 2.62
HER2 +	-	-	193	0.53	0.41, 0.65	27	2.38	2.25, 2.50
Basal	-	-	192	0.00	-0.12, 0,12	7	2.35	2.07, 2.63
Claudin-Low	-	-	199	0.04	-0.06, 0.13	-	-	-
Normal-like	-	-	137	-0.03	-0.17, 0.11	3	2.49	1.17, 3.80
Cluster								
1	-	-	126	-0.28	-0.43, 0.13	6	2.45	2.13, 2.77
2	-	-	68	-0.38	-0.56, -0.20	4	2.30	1.87, 2.74
3	-	-	275	-0.12	-0.22, -0.03	7	2.23	2.11, 2.36
4ER-	-	-	238	-0.13	-0.24, -0.03	6	2.28	2.08, 2.47
4ER+	-	-	71	0.34	0.13, 0.55	3	2.10	1.91, 2.29
5	-	-	163	0.47	0.33, 0.60	21	2.42	2.26, 2.58
6	-	-	81	-0.30	-0.49, -0.11	3	2.38	1.32, 3.44
7	-	-	176	-0.26	-0.38, -0.14	6	2.36	2.15, 2.58
8	1	-2.08	279	-0.29	-0.39, -0.20	9	2.56	2.32, 2.79
9	-	-	132	0.17	0.01, 0.34	10	2.65	2.39, 2.91
10	-	-	212	-0.05	-0.15, 0.05	7	2.35	2.10, 2.60
ER Status								
Positive	1	-2.08	1394	-0.19	-0.24, -0.15	62	2.43	2.35, 2.51
Negative	-	-	425	0.24	0.16, 0.32	20	2.33	2.19, 2.47
PR Status								
Positive	1	-2.08	971	-0.23	-0.28, -0.18	37	2.42	2.31, 2.52
Negative	-	-	850	0.07	0.01, 0.13	45	2.40	2.31, 2.50
HER2 Status								
Positive	-	-	215	0.40	0.28, 0.52	21	2.43	2.27, 2.59
Negative	1	-2.08	1606	-0.16	-0.20, -0.12	61	2.40	2.32, 2.48

* For each data type the total number of cases may differ due to missing values or incomplete data within the METABRIC dataset.

5.2.1 Distribution of clinicopathological features based on ASCL2 expression

The average age of breast cancer onset between tumours with overexpressed, unaltered and downregulated *ASCL2* was first analysed, as patients presenting at a younger age (before 40 years) are known to have more aggressive tumours and a reduced overall survival compared to older women (Anders, *et al.*, 2009). However, the mean age of onset was approximately 61 years for all *ASCL2* expression groups - a one-way ANOVA (with Tukey post hoc multiple comparisons test) showed no significant difference in age of breast cancer onset between the upregulated, downregulated and unaltered expression groups of *ASCL2* (p=0.497, F=0.699).

Stage and grade were other clinical parameters used to assess the invasive capacity and aggression of tumours in patients. Likewise, no difference was observed, other than a slight tendency for *ASCL2* overexpressing tumours to be associated with a higher grade and a more advanced stage (stage 3 and 4 more frequently observed where *ASCL2* is upregulated (16%) compared with unaltered *ASCL2* expression (8%)). Analysis of stage and histological grade of tumours across *ASCL2* expression groups did not show statistical significance (Pearson Chi-Square, χ^2, p=0.224 and p=0.45 respectively). Overall, breast cancer was most frequently diagnosed as Stage 2 across *ASCL2* unaltered and upregulated tumours.

As breast cancer is a clinically and biologically heterogeneous disease, highlighting the possible genes involved and the underlying gene expression patterns of each subtype can provide a clearer portrait to guide clinical management. Therefore, the association of *ASCL2* expression with intrinsic subtype and integrative cluster (10-subtype classification by Curtis, *et al.*, (2012)) distribution was investigated.

The greatest frequency of tumours with overexpressed *ASCL2* were found to be classified as Luminal B or HER2 Positive (32.9% for each). Luminal A tumours represented 22% of samples with increased expression in *ASCL2* mRNA, whilst the lowest frequency of tumours overexpressing *ASCL2* (8.5%) was observed within the triple negative subtype (basal and claudin low) (Figure 5.1). The highest mean *ASCL2* expression was seen in HER2 positive tumours (Figure 5.1). Despite the large spread of data within each subtype (Figure 5.1 A), the distribution of breast cancer subtypes appeared to be statistically different (*$p<0.001$*, χ^2) between the three *ASCL2* expression groups.

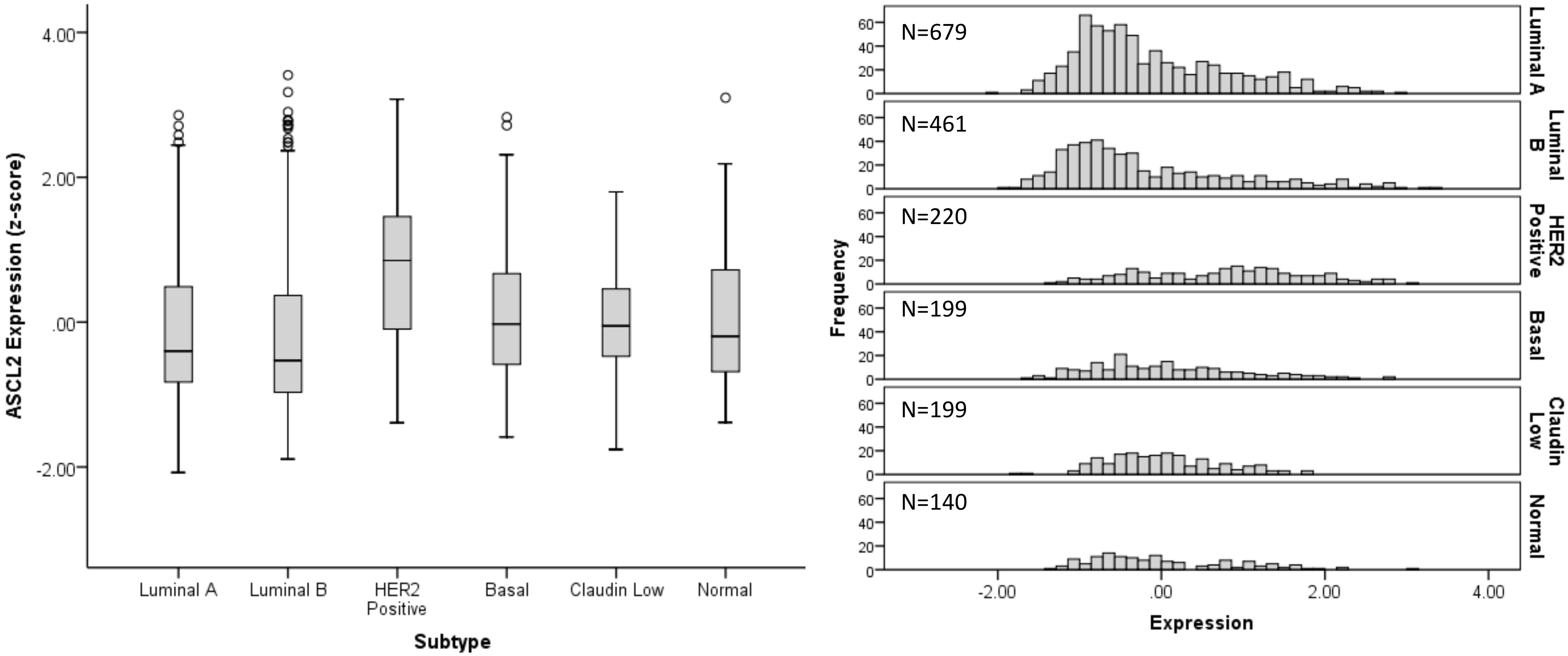


Figure 5.1. A. Boxplots showing the distribution of *ASCL2* gene expression across the intrinsic subtypes. HER2+ cancers appear to exhibit the highest expression of *ASCL2.* B. A histogram showing the frequency of tumours at various expression levels. The majority of HER2+ samples were shifted to the right of the histogram, compared to the other subtypes exhibiting a skew towards the left of the histogram.

Although the current diagnostic standard of classification is based on the PAM50 intrinsic subtypes, Curtis, *et al.*, (2012) previously proposed a new and more refined integrated classification of breast cancer, based on the combined analysis of genomic and transcriptomic information. This defined 10 distinct integrative clusters, with different genetic profiles and clinical courses.

Figure 5.2 illustrates the distribution of *ASCL2* gene expression across the 10 integrative clusters, where it can be seen that cluster 5 exhibited the highest mean expression of *ASCL2*, as well as the greatest percentage of *ASCL2* overexpressing tumours compared to all other clusters (25.6%). Cluster 5 represented tumours that were HER2 positive (can be ER-/+) with a poor prognosis, presenting early and of a higher grade according to (Dawson, *et al.*, 2013); this was also concordant with subtype data presented in Figure 5.1. Among tumours with an unaltered *ASCL2* expression, cluster 8 was more common, representing 15.3% of cases; cluster 8 was characterised by a 1q gain, 16q loss, the presence of hormone receptors, and were more likely to be low grade Luminal A tumours with a good prognosis. Consistent across all data was that the triple negative/basal phenotype was the least associated with *ASCL2* overexpression. The varying distributions of the integrative cluster classifications were shown to be statistically significant between expression groups (p=0.01, χ^2).

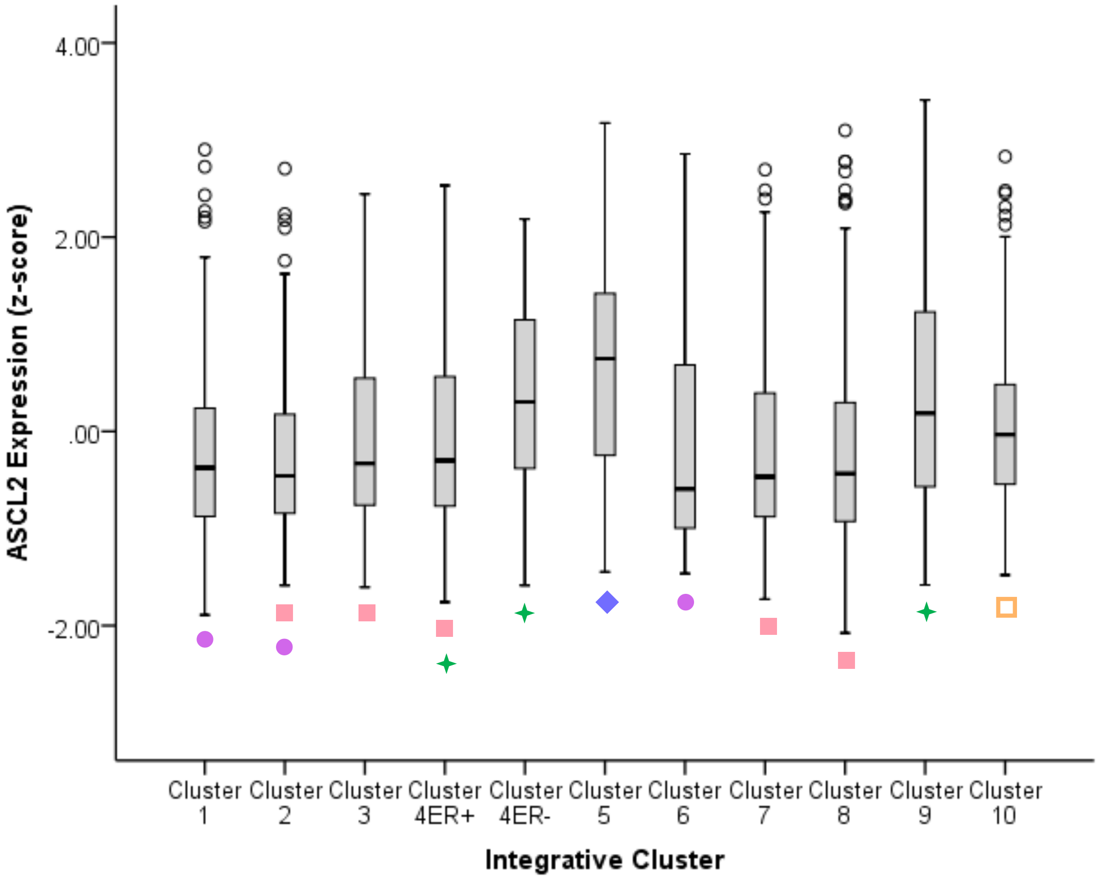


Figure 5.2. Boxplots showing the distribution of *ASCL2* gene expression across the integrative clusters. Clusters have been mapped to the most dominant PAM50 intrinsic subtype. ■ Luminal A ● Luminal B ◆ HER2 + □ TNBC/Basal ✦ Mix

Analysis of receptor status highlighted that tumours overexpressing *ASCL2* were more likely to be ER positive, HER2 negative, and PR negative (based on percentage distribution, Appendix 5). However, those that were HER2+ were more likely to be correlated with a higher gene expression of *ASCL2*, which is illustrated in Figure 5.3, by the upward shift of *ASCL2* expression in the distribution of HER2+ tumours. Statistical analysis using a Pearson Chi-Square (^{X2}) test revealed that ER and PR receptor status did not show a significant difference between expression groups (p=0.838 and 0.222 respectively). In addition, logistic regression analysis was performed to model the relationship between *ASCL2* expression and HER2 status. Results of this model indicated that *ASCL2* expression was less associated with negative HER2 status (Odds Ratio [OR] = 0.546; 95% confidence interval [CI] = 0.5, to 0.6; *p<0.001*), indicating that *ASCL2* expression is approximately 80% more likely to be associated with HER2+ tumours.

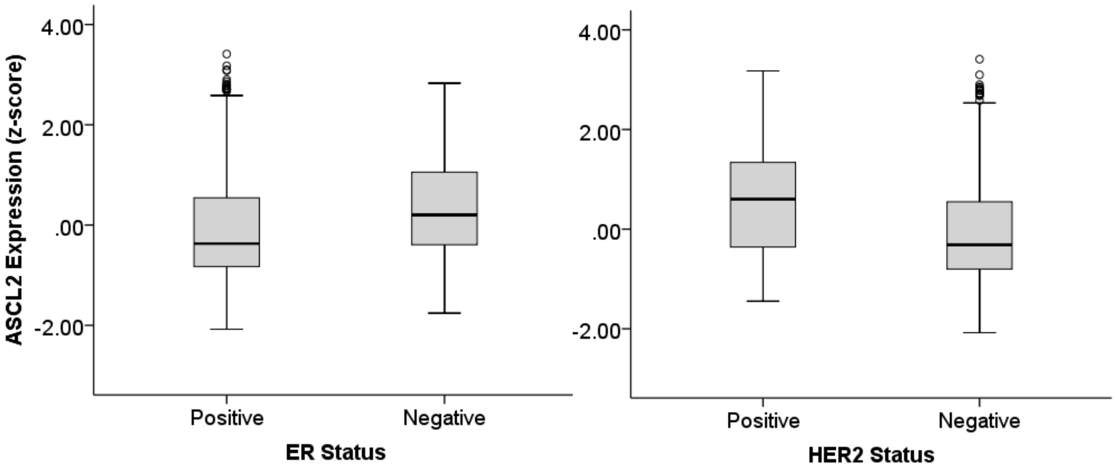


Figure 5.3. Boxplots showing the distribution of *ASCL2* gene expression between ER/HER2 positive and negative tumours. There is evidence to show that *ASCL2* may correlate with HER2 positivity in breast tumours.

5.2.2 Analysis of ASCL2 as an indicator of survival in patients

The impact of *ASCL2* overexpression on overall patient survival within the entire METABRIC cohort was evaluated. Within the whole patient population in this study, there was one patient exhibiting a downregulated expression of *ASCL2*. This patient survived for approximately 59 months, which was roughly half the time of the rest the cohort (with unaltered or overexpressed *ASCL2*, Table 5.1). However, as this was an isolated case (n=1), it could not be considered as a true representation of survival linked to downregulated *ASCL2* expression, and was therefore excluded from the majority of analyses.

When evaluated alone, patients with tumours overexpressing *ASCL2* had a lower mean survival (114.7 months) compared to tumours with unaltered *ASCL2* expression (125.5 months) (Appendix 5). Yet, no statistical significance was found to suggest that *ASCL2* overexpression in tumours was associated with poorer overall survival in breast cancer, in comparison to tumours with unaltered *ASCL2* expression (Hazard ratio [HR] = 1.12; 95% confidence interval [CI] = 0.9, to 1.4; *p>0.05*). The Kaplan-Meier plot in Figure 5.4 illustrates no significant difference in overall survival between tumours with unaltered and upregulated *ASCL2* gene expression, determined by a log-rank test. A one-way ANOVA (with Tukey post hoc multiple comparisons test) also indicated no significant difference (p=0.311, F=1.167) in overall survival between the 3 expression groups. However, there is a possibility that these findings could have been attributed to large differences between the size of each group (82 tumours overexpressing *ASCL2* vs 1821 with unaltered expression).

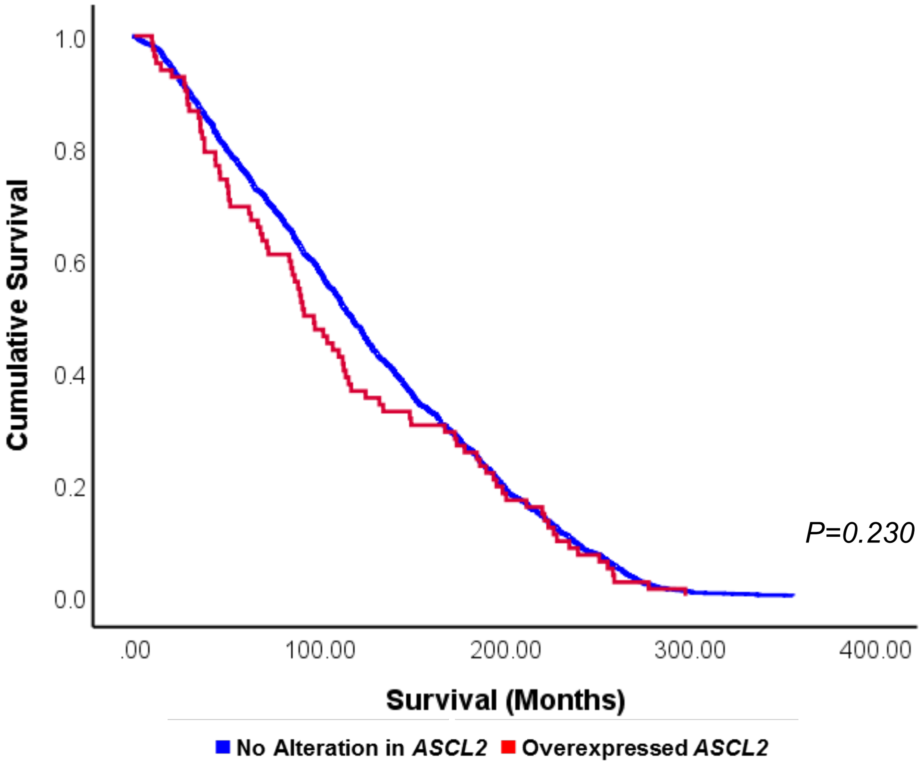


Figure 5.4. Kaplan–Meier overall survival analysis comparing breast tumours with no alterations or overexpression in *ASCL2* within the complete METABRIC cohort. No statistically significant difference was observed for overall survival based on *ASCL2* expression, calculated based on the Mantel-Cox log-rank test, and cox regression analysis (HR = 1.12; 95% CI = 0.9, to 1.4; *p>0.05*). Overall survival was defined as the time of diagnosis to the time of death.

Next, the relationship between subtype (PAM50 intrinsic subtyping vs integrative clustering) and patient survival was explored. Firstly, it was determined whether *ASCL2* overexpression affected overall survival, and therefore patient outcome, based on intrinsic subtype classification. Figure 5.5 illustrates the differences between overall survival trends in each subtype. Overall, the results revealed that there was no statistically significant impact of unaltered or increased *ASCL2* expression on survival across the subtypes, therefore no additional prognostic value was provided when subtype was considered (Table 5.3). However, among all subtypes, tumours with *ASCL2* overexpression did appear to exhibit a decreased survival time.

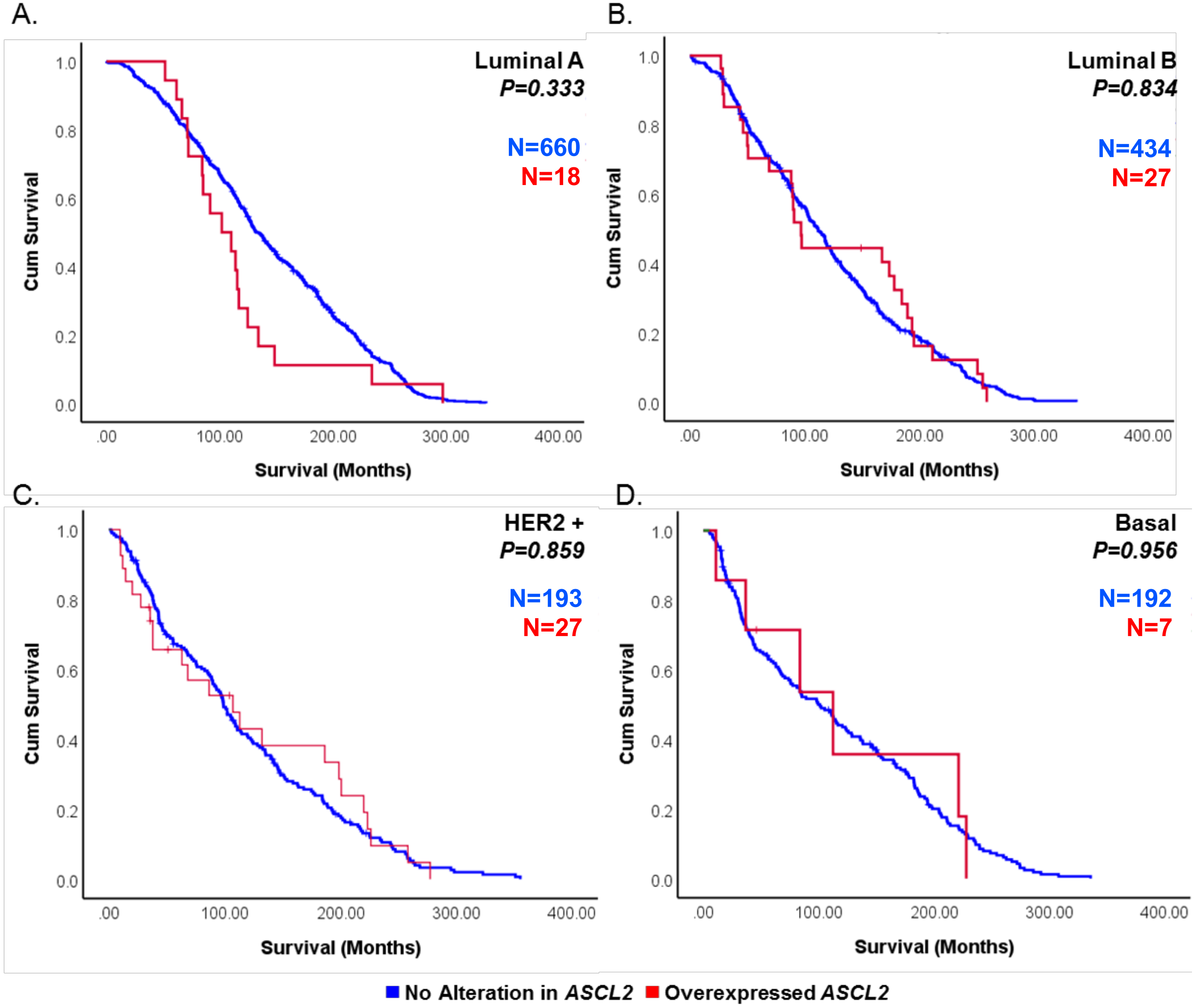


Figure 5.5. Kaplan–Meier curves for overall survival comparing breast tumours with no alterations or overexpression in *ASCL2* within the intrinsic subtypes A. Luminal A, B. Luminal B, C. HER2+, D. Basal. Overall survival for the claudin low subtype is not presented as none of these tumours exhibited an increased expression of *ASCL2*. No significant differences were observed for overall survival depending on *ASCL2* expression between the intrinsic subtypes. P values were calculated based on the Mantel-Cox log rank test.

With regards to the integrative clusters, although Kaplan-Meier curves illustrate that overexpression of *ASCL2* was associated with overall survival in cases classified as cluster 4ER- (Figure 5.6A, *p<0.05*) and cluster 6 (Figure 5.6C, *p<0.05*), no significant association was exhibited using cox regression analysis (Table 5.3). Cluster 4ER- represents tumours which are ER negative and a mixture of the intrinsic subtypes, with low level of genomic instability and a favourable outcome. By contrast, cluster 6 tumours are considered aggressive ER positive/HER2 negative, Luminal A or B tumours with high genetic instability and an intermediate prognosis (Dawson, *et al.*, 2013). However, in these two clusters, it was observed that the frequency of tumours with *ASCL2* overexpression were much smaller than the number of unaltered cases (cluster 4ER -, no alteration, n=238 vs overexpressed, n=6; cluster 6, no alteration, n=81 vs overexpressed, n=3). As the number of tumours overexpressing *ASCL2* were so low, this is likely to be influencing the estimation of significance compared to the other clusters, and therefore calls into question the reliability of the Mantal-Cox log rank test statistics here.

Despite cluster 5 tumours presenting the majority of cases overexpressing *ASCL2* and the highest mean *ASCL2* expression, no significant association between *ASCL2* and overall survival was found (Figure 5.6 B, Table 5.3). As previously mentioned, within tumours with an unaltered *ASCL2* expression, the greatest frequency was classed as cluster 8. However, in this survival analysis, cluster 8 tumours with an increased *ASCL2* expression appeared to exhibit an improved overall survival in this cluster (Figure 5.6 D), contrary to the other clusters; nonetheless, this estimation is statistically insignificant, and therefore, no significant association between *ASCL2* and overall survival was found in cluster 8.

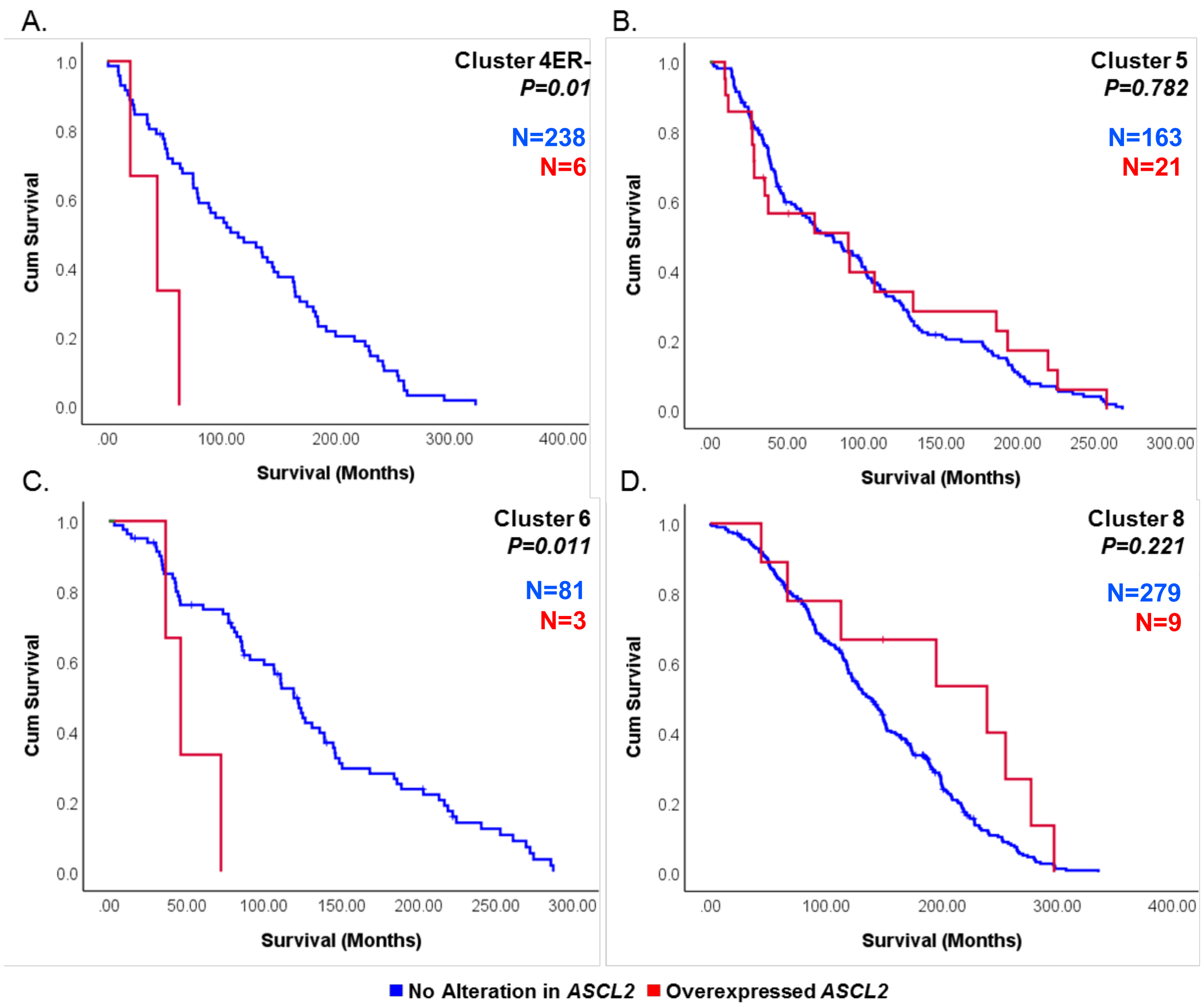


Figure 5.6. Kaplan–Meier curves for overall survival comparing breast tumours with no alterations or overexpression in *ASCL2* within the integrative clusters. A. Cluster 4ER-, B. Cluster 5, C. Cluster 6, D. Cluster 8. Overall survival for all intrinsic subtypes are not presented, however were analysed. No significant differences were observed for overall survival depending on *ASCL2* expression in cluster 5 despite being highlighted as the cluster most associated with *ASCL2* overexpression. P values were calculated based on the Mantel-Cox log rank test.

Finally, to evaluate the association of *ASCL2* on overall survival, as a single factor in the context of clinicopathologic factors, or adjusted for clinicopathologic factors, cox regression analysis was used. This showed that when evaluated alone *ASCL2* expression was not significantly associated with overall survival in breast cancer patients (univariate, HR = 1.04, 95% CI = 0.99, to 1.09; *p>0.05*), similarly to when adjusted for other clinical factors (multivariate, HR = 1.00, 95% CI = 0.94, to 1.08; p>0.05) (Table 5.3).

Overall, these results show that there was no statistically significant relationship between *ASCL2* expression and overall survival, and therefore no additional prognostic value is provided when considering *ASCL2* expression in patients (Table 5.3).

Table 5.3. Hazard ratio univariate and multivariate cox regression analysis of the relationship between *ASCL2* expression and overall survival in patients of the METABRIC study. Statistical significance was calculated using the Cox proportional hazards regression test.

	Univariate		**Multivariate**	
Factor	**HR, 95% CI**	***P* value**	**HR, 95% CI**	***P* value**
ASCL2 Expression	1.04 (0.99 – 1.09)	0.11	1.00 (0.94 - 1.08)	0.878
ASCL2 Expression Group *Unaltered vs* *Upregulated*	1.12 (0.90 – 1.40)	0.30	0.97 (0.70 – 1.34)	0.840
Age				
≥60 vs	1.09 (1.02 – 1.17)	0.02*	1.36 (1.21 – 1.52)	0.001*
< 60	1.01 (0.95 – 1.07)	0.80		
ER Status				
positive vs	1.05 (0.99 – 1.10)	0.11	0.78 (0.62 – 0.99)	0.038*
negative	0.95 (0.86 – 1.06)	0.36		
PR Status				
positive vs	1.00 (0.93 – 1.07)	0.99	0.94 (0.81 – 1.08)	0.342
negative	1.04 (0.98 – 1.11)	0.23		
HER2 Status				
positive vs	1.04 (0.92 – 1.17)	0.57	1.13 (0.85 – 1.50)	0.410
negative	1.01 (0.96 – 1.07)	0.61		
PAM 50 Subtype				0.16
Luminal A	1.04 (0.95 – 1.13)	0.41	1.00	-
Luminal B	1.05 (0.97 – 1.14)	0.26	1.05 (0.89 – 1.23)	*0.58*
HER2 +	1.04 (0.91 – 1.18)	0.56	0.80 (0.62 – 1.04)	*0.10*
Basal	0.91 (0.78 – 1.07)	0.25	1.00 (0.74 – 1.33)	*0.97*
Claudin-Low	0.94 (0.77 – 1.16)	0.59	1.01 (0.79 – 1.29)	*0.94*
Normal-like	1.05 (0.88 – 1.27)	0.57	1.22 (0.96 – 1.54)	*0.10*
Integrative Cluster				0.25
1	1.07 (0.91 – 1.26)	0.40	1.00	-
2	0.89 (0.69 – 1.15)	0.36	1.09 (0.78 -. 1.52)	*0.63*
3	1.11 (0.96 – 1.28)	0.17	0.91 (0.70 – 1.20)	*0.52*
4ER-	0.99 (0.77 – 1.27)	0.93	0.91 (0.69 – 1.20)	*0.51*
4ER+	1.09 (0.94 – 1.25)	0.26	0.87 (0.59 – 1.29)	*0.49*
5	1.07 (0.94 – 1.22)	0.32	1.34 (0.93 – 1.93)	*0.11*
6	1.28 (1.01 – 1.63)	0.04*	0.95 (0.69 – 1.33)	*0.78*
7	0.97 (0.82 – 1.15)	0.75	0.92 (0.69 – 1.23)	*0.59*
8	0.89 (0.79 – 1.00)	0.06	0.91 (0.69 – 1.19)	*0.48*
9	1.05 (0.90 – 1.21)	0.56	1.08 (0.80 – 1.46)	*0.61*
10	0.86 (0.73 – 1.02)	0.08	0.75 (0.55 – 1.02)	*0.07*
Grade				0.94
1	0.96 (0.82 – 1.14)	0.65	1.00	-
2	1.05 (0.97 – 1.13)	0.28	0.98 (0.80 – 1.21)	*0.88*
3	1.03 (0/96 – 1.10)	0.42	0.97 (0.77 – 1.21)	*0.75*
4	-	-	-	-
Stage				-
0	0.75 (0.32 – 1.76)	0.51	1.00	-
1	1.06 (0.96 – 1.17)	0.26	0.88 (0.12 – 6.37)	*0.90*
2	0.97 (0.91 – 1.05)	0.48	1.27 (0.17 – 9.22)	*0.82*
3	1.05 (0.89 – 1.24)	0.59	1.97 (0.27 – 14.5)	*0.51*
4	2.0 (0.78 – 5.01)	0.15	3.22 (0.40 – 26.1)	*0.27*

5.3 Discussion

The results in this chapter suggest that when analysed as a single entity in patient tumour samples, higher *ASCL2* expression was most associated with HER2+ tumours in breast cancer. However, no statistically significant findings suggested that *ASCL2* overexpression was associated with poorer patient survival in breast cancer, when analysed in the whole study population or in the context of other clinical features. Therefore, this gene cannot be considered a prognostic marker.

Results from primary tumour samples indicated that increased expression of *ASCL2* had a significant correlation with HER2 receptor expression in breast cancer (Figures 5.1, 5.2, 5.3). However, this did not exclude tumours with upregulated *ASCL2* primarily expressing hormone receptors (ER/PR) or of Luminal A or B subtypes. Overall, HER2+ tumours appeared to be more likely to aberrantly overexpress *ASCL2* compared to other tumour subtypes. In addition to this, approximately a quarter of all tumours with upregulated *ASCL2* were classified into integrative cluster 5 associated with HER2+ tumours (whereas all other tumours with increased *ASCL2* expression were equally distributed among the remaining clusters). Although many drugs have emerged in the last decade that target HER2, these have been met with challenges including acquired resistance (Vu, & Claret, 2012). Hence, exploring the relationship between *ASCL2* and HER2+ tumours further may shed light on possible mechanisms that lead to resistance, to improve patient response to these drugs. Although this remains a tenuous link, it may be an avenue worth pursuing. Tumours with unaltered expression were mainly associated with cluster 8 (15%), however distribution is more equal among clusters, and conversely do not show a strong affinity with one cluster in particular.

To assess the impact of *ASCL2* on patient outcomes and as a possible prognostic marker, overall survival was examined. Xu, *et al.*, (2017) investigated the clinical relevance of ASCL2 in breast cancer by examining specimens from 191 breast cancer cases using immunohistochemical staining. The study concluded that higher levels of ASCL2 correlated with poorer overall survival, greater tumour recurrence and relapse in patients.

In the present study, overall, patients with tumours overexpressing *ASCL2* survived for approximately 11 months less, which, in clinical terms, could be considered critical. However, statistical analysis indicated otherwise, concluding that there was little to no evidence to show that *ASCL2* expression was associated with patient survival. Though, consistent with the literature, *ASCL2* expression did not affect overall survival based on the breast cancer intrinsic subtypes.

Although the study by Xu, *et al.*, (2017) presented significant survival trends relating to ASCL2 expression, their study focussed on protein expression scored semi-quantitatively into high and low groups, whereas the present study relied on quantitative gene expression data (with an objective, numerical threshold value), a larger sample size, and also included an 'unaltered *ASCL2*' group of patients. Therefore, it could be assumed that the classification of *ASCL2* overexpression had a higher and more stringent cut off in the present study. In addition, it could be argued that the present study, inclusive of almost 2000 patients, is more robust than the study by Xu, *et al.*, (2017) focussing on just under 200 samples. However, the findings in this chapter neither support nor reject the work carried out by Xu, *et al.*, (2017) due to the differences in study design.

Nevertheless, with the previous study in mind, to strengthen and expand the estimation of the prognostic significance of *ASCL2,* future work may include investigation into tumour recurrence, or disease-specific survival (distinguishing patients dying from breast cancer specifically, in comparison to other possible causes); this data was not available from the METABRIC dataset. Further exploration of this hypothesis could include more datasets such as those from The Cancer Genome Atlas (TCGA), available from the cBioPortal, which did follow-up on tumour relapse.

Upon analysing the results obtained from the METABRIC study, there was a slight discrepancy in the data mining of primary tumours compared to the data mining of cell lines (Chapter 3) and *in vitro* laboratory work (Chapter 4). Extreme variation analysis presented in Chapter 3 pointed to the highest *ASCL2* expression in MCF7 cells. Likewise, RT-qPCR gene expression analysis revealed the highest *ASCL2* expression in MCF7 and T47D cells (Luminal A), and SKBR3 cells (HER2+) respectively, with low expression in BT474 cells (Luminal B). Functional analysis also highlighted a potential role of *ASCL2* in the migration of Luminal A cells, yet no impact was observed in HER2+ cells.

Therefore, it was considered that *ASCL2* expression was the most associated with Luminal A tumours.

Yet, analysis of clinical cases revealed the strongest correlation (according to distribution) with HER2+ tumours and Luminal B tumours. Nevertheless, there was evidence supporting the association with Luminal A tumours; *ASCL2* was overexpressed in 22% of these tumours and approximately three quarters of tumours with *ASCL2* expression were ER positive. Additionally, analysis of receptor status highlighted that tumours overexpressing *ASCL2* were more likely to be ER positive, HER2 negative, and PR negative (based on percentage distribution). From this perspective, data was consistent with the MCF7 cell line used for *ASCL2* investigation in Chapter 4. However, considering the variability in trends observed between cell lines and patient tumour data in this study, it remains uncertain whether overexpression of *ASCL2* is correlated with HER2 positivity. Hence, further analysis of the relationship between *ASCL2* and HER2 positivity is required to confirm this. Overall, the possibility that *ASCL2* overexpression may result in susceptibility to a specific breast cancer subtype was not entirely confirmed in this study.

Though the methodology used in this chapter was beneficial within the scope of this project, the shortfall with this type of analysis was that it investigated the overexpression of *ASCL2* as an isolated event in cancer. Cancer is an extremely complex and multifaceted process, yet this analysis reduced the complex nature of cancer down to a single gene relating to patient survival. The magnitude of the impact of *ASCL2* overexpression as a single entity is likely to be minute in comparison to the expression of the entire genome (Prat, *et al.*, 2014). However, it is completely plausible to consider that *ASCL2* could be one part of a larger puzzle influencing survival; therefore, *ASCL2* could be investigated as part of a multi gene signature. As of yet, the mechanistic role of this gene in breast cancer remains to be discovered, so pinpointing other genes to investigate alongside *ASCL2* would be the first step to achieving this. It must also be noted that the absence of a statistically significant trend between *ASCL2* expression and patient survival should not be considered as the only marker of clinical outcome. For example, *ASCL2* may be a contributor to other measures such as therapy resistance, or the spread of secondary tumours. Likewise, the predictive potential of *ASCL2* on these variables could be further explored in the future.

Of particular note when analysing the data presented in this chapter was the large variation, diversity and spread of data from patient tumours. An explanation for this may have been due to the inter-tumour heterogeneity seen between clinical breast tumours, compared with *in vitro* cell culture, which generally consists of a uniform cell population (Sun, & Yu, 2015). A further complication of data collection and analysis in patients is the genetic and epigenetic heterogeneity reported in different parts of the same tumour; oncologists tend to rely on the molecular characterisation of a small sample of tumour tissue, which is unlikely to represent the true heterogeneity seen within and between patients (Bedard, *et al.*, 2013; Zardavas, *et al.*, 2015). This has been observed in many multi-omic and single-cell transcriptome profiling studies (Bareche, *et al.*, 2018; Bedard, *et al.*, 2013; Chung, *et al.*, 2017). The heterogeneity observed between the samples of this dataset exemplify the challenging nature of gene investigation that researchers are attempting to navigate and overcome.

Despite the use of cell lines being scrutinised over the years, mainly due to their questionable representation of tumour biology *in situ*, and their translatability of research findings into the clinic (Choi, *et al.*, 2014), their use in this study was warranted. Firstly, because on the whole, laboratory findings have been consistent with cell line gene expression data from Array Express and Gene Expression Omnibus (GEO) (Chapter 3); additionally, the Human Protein Atlas also verified expression of *ASCL2* in MCF7, T47D and SKBR3 cells. Secondly, the clinical data mining component of this study complements laboratory evidence, by taking into account the tumour biology of primary tumours and the effect of *ASCL2* on clinical parameters in patients. These results can also be used to direct the trajectory of further study in the future. Although cell line models are not ideal, they are a good pre-clinical model and basis for initial candidate gene investigation; ultimately, using cell lines can inform researchers about tumour biology prior to investigation using more advanced models (Gillet, *et al.*, 2013; Katt, *et al.*, 2016).

Although clinical investigation added to the breadth of this study, there were drawbacks to this analysis which need to be addressed in the future to strengthen conclusions. The issue with small sample size is of high priority here. Despite 2509 patient samples downloaded from the METABRIC study, not all samples had full clinical information provided, and therefore not all samples were able to be analysed. This led to missing information that needed to be culled prior to statistical analysis. As well as this, owing to the low prevalence of *ASCL2* overexpressing tumours within the METABRIC cohort (3.3%, representing 82 tumours), a small sample size for such a varied population meant that some parameters or ‘events’ could only be measured at very low frequencies. This broad spread of data, divided into smaller unbalanced subsets (for example, Table 5.1, 18 Luminal A tumours overexpressing *ASCL2* vs 660 Luminal A tumours with no alteration in *ASCL2* expression), complicated statistical analysis and may have resulted in a bias towards statistical insignificance (Block, *et al.*, 2018; Ogden, *et al.*, 2017). Small sample sizes do not have the power to detect subtle gene expression changes, and are in essence only applicable when expression changes are large (Biau, *et al.*, 2008; van Iterson, *et al.*, 2009). Therefore, a greater sample size of tumours overexpressing *ASCL2* would lead to more reliable conclusions. To target this shortcoming, further data could be downloaded from other studies to compile a larger dataset from multiple sources, thus integrating *ASCL2* gene expression data and clinical information from a greater number of primary tumours, rather than from a single dataset (Prat, *et al.*, 2014). Although caution must be taken to account for differences between study designs, this may represent a more relevant approach to investigating the clinical impact of *ASCL2* in patient tumours.

Another improvement to further develop this study, could focus on sourcing a dataset which includes information on modifiable (age, family history) and non-modifiable (BMI, childbirths) factors in patients, as well as clinicopathologic features of tumours. In the same vein, the results from this dataset may have benefitted from the inclusion of factors such as menopausal state, allowing for the adjustment of data by other risk factors, potentially improving the estimation of *ASCL2* as a prognostic indicator. Statistical analysis could also be elaborated by the inclusion of power analysis, for the determination of the appropriate sample size.

The cBioPortal web tool was used to access METABRIC data, and was selected as it allowed the refined analysis of a single gene (or a select group of genes) within a large-scale cancer dataset (Zhang, *et al.*, 2018). This tool was also open-access, allowing the use of existing public data, as well as being suitable for use by researchers at any level, even those with little knowledge in bioinformatics.

These factors were in-line with themes discussed in Chapter 3 regarding the accessibility and usability of bioinformatics tools to enhance research (Zhang, *et al.*, 2018).

To conclude this chapter, it must be noted that identifying clinically relevant, specific and robust markers that translate from the laboratory to the clinic is a prominent and pertinent challenge for researchers. Overall, the present study was well performed, indicating no statistically significant evidence to support the role of *ASCL2* as a clinical marker in patient tumours; the hypothesis that higher *ASCL2* expression may be related to poorer survival in patients was not supported by the data presented in this study. However, likely causes influencing these findings may be heterogeneity between patient tumours, and the low prevalence of elevated *ASCL2* expression, therefore further analysis will be important for the confirmation of this data. The widespread and varied expression of *ASCL2* across the subtypes also discounts the possibility of aberrant expression being subtype-specific, yet does suggest that this gene may play a more fundamental role during tumour development; a view also expressed by Xu, *et al.*, (2017). Consistent across all data in this Chapter was that the triple negative/basal phenotype was the least associated with *ASCL2* overexpression. Therefore, even considering the limitations discussed, the work presented in this Chapter has contributed to the knowledge of *ASCL2* in breast cancer, concluding that this gene is not associated with survival.

Chapter VI

Overall Discussion, Future Work & Conclusions

6.1 Overall Discussion

Research directed towards improved understanding, identification of novel markers and stratification of the complex heterogeneity of breast cancer has led to increased patient survival over the past decade; as of 2014, almost 90% of patients survived for 5 years or more (National Cancer Institute, 2018). The research outlined in this thesis aimed to use an integrated approach to identify a novel candidate gene in breast cancer, and investigate its biological function and potential as a clinical marker.

This research provides novel evidence that *ASCL2* may be involved in breast tumourigenesis by way of influencing cellular migration via the Wnt signalling pathway. The evidence gathered in Chapters 3 and 4 employing bioinformatics and RT-qPCR, demonstrated the differential expression of *ASCL2* across varying breast cancer cell lines. Notably *ASCL2* expression levels were significantly increased in MCF7 cells compared to non-tumourigeneic cells (MCF10A). In addition to this, it was proposed that *ASCL2* knockdown may have had an anti-migratory effect in Luminal A cell lines; further supporting this was the enrichment of GO terms relating to the regulation of cell migration and motility, wound healing, morphogenesis and EMT in breast cancer cell lines (DAVID and GO analysis, Chapter 3). Data described also exhibited the reduced expression of Wnt signalling associated genes as a result of *ASCL2* silencing, in particular, the stemness marker genes *CD44* and *LGR5, and CTNNB1 (*β-catenin) suggesting that *ASCL2* is an upstream regulator of these Wnt pathway genes. This inhibition of migration resulting from *ASCL2* silencing could be considered as an effect of the downregulation of Wnt signalling in breast cancer, however further confirmation is needed. Together with previous studies in colon cancer (Tanaka, *et al.*, 2019), these findings indicate that the Wnt/*ASCL2* pathway may harbour key targets for the management of the Wnt pathway in cancer.

Within a broader context, it could be inferred from this primary analysis that *ASCL2* may be an oncogene contributing to the migratory and invasive properties of certain breast cancer cells and cancer stem cells that push progression through the EMT. The EMT is a fundamental biological process within embryogenesis, development and wound-healing, yet is also considered a malignant driver (Chaffer, *et al.*, 2016). Likewise, *ASCL2* has been established as a developmental gene within neurogenesis and embryogenesis (García-Bellido, & de Celis, 2009; Guillemot, *et al.*, 1994; Oh-McGinnis, *et al.*, 2011). In intestinal stem cells it has been shown that *ASCL2* correlates directly with *LGR5* expression to regulate stemness (van der Flier, *et al.*, 2009; Yan, *et al.*, 2015). Giakountis, *et al.*, 2016 and Schuijers, *et al.*, 2015 also demonstrated that alongside TCF4/β-catenin, *ASCL2* can activate stem cell gene expression programmes in intestinal stem cells through an auto-regulatory positive feedback loop involving Wnt signalling in colon cancer. Evidence in colon cancer also suggests a role of *ASCL2* in EMT (Tian, *et al.*, 2014). Therefore, it could be hypothesised that the overexpression of this gene may act in a similar way to activate transcriptional programmes in breast cancer, and possibly more specifically in breast cancer stem cells (Smith, *et al.*, 2017). Overall, this may result in tumour initiation and progression, by over activity of the EMT program, modulation of plasticity, increased cell migration, and activation of further downstream oncogenic genes within the Wnt pathway (Schuijers, *et al.*, 2015; Tian, *et al.*, 2014; Zuo, *et al.*, 2018). Although the relevance of *ASCL2* to breast cancer stem cells is yet to be established, this study highlights an area requiring further attention.

The work undertaken in Chapter 5 looked to evaluate the expression of *ASCL2* as a potential prognostic indicator and clinical marker. Considering the previously observed effects on cellular migration in MCF7 and T47D cells (Chapter 4), one might have expected to have seen a correlation between high *ASCL2* expression and more aggressive tumours in patients (for example, earlier age of onset, advanced stage and grade, poorer survival), particularly within Luminal A tumours. However, patient data had a largely varied distribution, and no significant association between *ASCL2* and survival outcomes were identified. Thus, the exact impact of *ASCL2* on clinicopathologic features remains to be elucidated, and this study concluded that *ASCL2* cannot be said to be associated with patient survival. Though, the variation observed in this clinical data acted to highlight the multifaceted complexity and diversity of tumourigenesis within patients, as well as the inter-tumour heterogeneity between patients within a population (Bedard, *et al.*, 2013).

Although previous reports have documented *ASCL2* as a potential prognostic indicator in other cancers (Hu, *et al.*, 2015; Liu, *et al.*, 2016; Xu, *et al.*, 2017), the results from this study suggests that this relationship may be more complex in breast cancer. This work suggests that *ASCL2* may play more of a core role in breast tumourigenesis that encompasses and underpins a variety of tumours, rather than a specific role attributed to a single subtype. In this sense, it seems unlikely that *ASCL2* could be considered as a specific molecular marker, but may be investigated as part of a predictive signature or multi-gene test in the future. For example, the Oncotype DX Breast Cancer Assay uses RT-qPCR to quantify gene expression of a 21-gene panel, for the estimation of overall survival, recurrence risk and response to chemotherapy (Cronin, *et al.*, 2007; Harbeck, *et al.*, 2014). Moreover, while a minority of breast cancers may be attributed to a single specific genetic aberration, most cases are triggered by a combination of genetic alterations. Complex diseases like cancer are caused by multiple genes that are dysregulated at different points of disease development, resulting in the presentation of diverse symptoms and different responses to treatment.

Across the literature, *ASCL2* is involved in a number of developmental processes, and has been implicated in some cancers. However its role remains varied and occasionally conflicting between tissues, species, and pathologies (Wang, *et al.*, 2017; Zhongfeng, *et al.*, 2018). This gene has essential roles in trophoblast development within the placenta (Bogutz, *et al.*, 2018), epidermal development (Moriyama, *et al.*, 2008), maintaining stemness of intestinal stem cells, follicular T-helper cell development (Liu, *et al.*, 2014), as an inhibitor of myogenic differentiation (Wang, *et al.*, 2017), and as a negative regulator of Schwann cell proliferation (Küry, *et al.*, 2002). These examples appear to be reflective of the complexity of *ASCL2*, suggesting both a microenvironment and tissue dependent function (Zhongfeng, *et al.*, 2018).

Within the context of breast cancer, evidence presented in this work and from the literature discussed throughout may lead to the future investigation of *ASCL2* as an ‘accessory driver’, or an epidriver. Currently, no driver mutations have been identified in *ASCL2*, thus it cannot be considered as a driver gene (affirmed by the COSMIC database, v87 – www.cancer.sanger.ac.uk). Yet, it is plausible to suggest that *ASCL2* may have an epistatic effect with varying potency on a number of other genes, including driver genes, thus investigation into mutual exclusive interactions and genetic co-occurrence could yield significant results. An example of an epistatic interaction is the synthetic lethality between BRCA mutations and PARP inhibition. This knowledge is now used in clinical practice for breast cancer treatment utilising PARP inhibitors (Gonzalez, *et al.*, 2016; Park, & Lehner, 2015). A growing body of evidence suggests that *ASCL2* contributes to tumourigenesis in a number of cancer types, and has been identified as a causative gene in colorectal cancer involved in therapeutic resistance (Tanaka, *et al.*, 2019). Although the link between *ASCL2* and therapeutic efficacy is yet to be examined in breast cancer, this represents an attractive area for further research.

In addition to the research centred on *ASCL2*, a simple analysis pipeline was presented in this thesis that is practical for use on its own, or may be flexibly adapted to suit the needs or expertise of the researcher. The advent of rapidly evolving omics technologies have given rise to an exponentially growing amount of biological data. Studies exploiting these technologies generally use extremely convoluted computational workflows, which although lead to the discovery of novel and interesting genes, yield little functional knowledge. This pipeline has demonstrated that gene inquiry can be kept modest without diminishing comprehension. A common value presently held by scientists, is the sharing of data and accessibility of tools. Thus, this study prioritised the use of open databases for the extraction of freely available transcriptomic cell line and patient data for gene investigation.

The merit of this type of study was the integrated approach consisting of a variety of bioinformatics methods for gene identification, laboratory analysis for functional investigation, and data mining of patient samples, aiming to bridge the gap between laboratory research and clinical application. The study design and methods utilised within this project also allowed for the broad investigation of a single gene which had been largely overlooked in breast cancer in the past. It is anticipated that this study may provoke future research in *ASCL2*.

Another strength of this study was the multiple levels of validation. Despite their assets, microarray studies are prone to issues relating to reproducibility due to the large number of gene probes in comparison to small samples sizes. Therefore, combining multiple studies can increase reliability and achieve greater precision when estimating differential gene expression (Ramasamy, *et al.*, 2008). In this work, microarray data from cell lines were pooled from a number of studies to address this. In addition, various pathway tools were compared and combined to allow for comprehensive identification of a candidate gene with oncogenic potential. This data was subsequently validated in cell lines and the role of this gene was explored *in vitro*. Finally, a large cohort of patient data was also obtained to explore the association of this gene on clinical outcomes in patient tumours.

The *in vitro* component of this study relied on the growth and maintenance of well-established breast cancer cell lines (Lacroix, & Leclercq, 2004). However, it is widely acknowledged that cell lines are prone to a number of limitations compared to other experimental cancer models, such as primary cell lines, patient derived tissue, or animal models. Such limitations include genetic and phenotypic drift over passages, the absence of a tumour microenvironment, and the lack of biological heterogeneity in comparison to patient tumours (Holliday, & Speirs, 2011). Yet, the use of a cell line model was practical for such initial investigation, due to their convenience, control over experimental variables, and ability to directly compare between experiments and replicates (Burdall, *et al.,* 2003). This research may be expanded using other tumour models, to uncover a greater breadth of knowledge regarding *ASCL2* in breast cancer.

As discussed throughout this thesis, a significant challenge within the area of gene investigation and marker identification for the implementation of personalised medicine and diagnostics, is tumour heterogeneity. Interpatient tumour heterogeneity has been acknowledged for some time, resulting in the employment of the intrinsic subtype classification system in the clinic. More recently, the issue of intra-tumour heterogeneity has been recognised, meaning that the predictive biomarkers present in a tumour may be different depending on location, and could be prone to change during progression or metastasis (Bedard, *et al.,* 2013). Evidence has also shown that breast cancer cells may exhibit subtype plasticity by exhibiting the ability to interconvert between subtypes (Yeo, & Guan, 2017). However, despite this challenge, huge strides in progress have been made over the last 10 years by fully exploiting omics technologies and harnessing the growing amount of biological data already available. This is reflected in the growing number of women surviving beyond 5 years (National Cancer Institute, 2018).

6.2 Recommendations for Future Work

The work presented in this thesis has identified some interesting, yet somewhat overlooked, implications in breast cancer, but also highlights a number of opportunities for further research. This work has shed light on *ASCL2* as a novel gene involved in breast carcinogenesis, however some findings and proposed mechanisms in this study require additional confirmation and research which could be executed in a number of ways.

To improve on the work detailed in this thesis, it would first be beneficial to optimise the consistency of gene knockdown, by performing permanent knockdowns via the use of lentiviral construction to improve 'loss of function' experiments (Boettcher, & McManus, 2015). This could be done by using stable expression of short hairpin RNA (shRNA) or the CRISPR/Cas9 system; the effects of shRNA transfection are more prolonged than the use of siRNA, whereas CRISPR generally yields a more stable, consistent and robust knockdown or knockout demonstrating a stronger effect on phenotypes (Boettcher, & McManus, 2015). Both of these techniques may also yield validation of a consistent protein knockdown. In addition to this, the study of apoptosis and wound-healing in Chapter 4 could be improved to strengthen conclusions; flow cytometry and silicon gap inserts along with automated time-lapse recording equipment could address these issues respectively. Lastly, data from other breast cancer studies could be pooled with METABRIC data to increase the sample size of tumours with *ASCL2* overexpression. As mentioned in the previous section, future work including more biologically relevant tumour models, such as *in vivo* models or 3D cell culture (Chen, *et al.*, 2012), would also be essential to combat the limitations associated with cell lines.

To expand on the work detailed in this thesis, the mechanism that leads to the overexpression of *ASCL2* in MCF7 cells could be investigated. It has been previously reported in other cancers that *ASCL2* may be epigenetically regulated by DNA methylation, the action of miRNAs and the long non-coding RNA, WiNTRLINC1 (Conway, *et al.*, 2014; Giakountis, *et al.*, 2016; Tian, *et al.*, 2014; Zhu, *et al.*, 2012). Therefore, it may be interesting to pursue this further and explore this prospect *in vitro* in breast cancer. A small number of studies have also shown that *ASCL2* is involved in therapeutic resistance in colon cancer and postulated its role as a chemoresistance biomarker (Juarez, *et al.*, 2018; Kwon, *et al.*, 2013; Tanaka, *et al.*, 2019). With this in mind, it would be worth studying this relationship in breast cancer, potentially by assessing the effects of *ASCL2* expression on common chemotherapeutic agents used in the treatment of breast cancer, or even the endocrine therapy, tamoxifen, for treatment of Luminal A tumours. This could also be expanded to include the monoclonal antibody, trastuzumab, targeting the HER2 receptor, as resistance to this drug remains a common challenge (Esteva, *et al.*, 2010; Vu, & Claret, 2012). A final consideration would be the eventual use of tissue microarrays to assess ASCL2 protein expression changes in correspondence with clinical parameters in a high-throughput manner.

6.3 Conclusions

Identifying and pursuing critical associations within the complex system that is breast cancer remains a considerable challenge for researchers. At the time of writing this thesis, a PubMed search of *ASCL2* returned 171 published studies (March 2019). Searching the keywords “ASCL2 cancer” returned 70 results, while “ASCL2 breast cancer” revealed 6 results. By comparison, searching “HER2 breast cancer” gave 17,354 results. This exemplifies how little is known about the *ASCL2* gene in relation to breast cancer, and highlights a research niche that requires attention.

This project has succeeded in its intentions, yet has also acknowledged a number of technical and field-related challenges throughout. Namely, the variability and complexity of bioinformatics analysis, the difficulties associated with gene investigation, and the vast heterogeneity observed in breast cancer. Ultimately, the work presented in this thesis, being the first comprehensive and integrated study to examine *ASCL2* in breast cancer, has contributed to the understanding of the multifaceted function and role of *ASCL2* in breast tumourigenesis. This study has brought to the forefront the potential of *ASCL2* as a novel gene involved in breast cancer development, whilst highlighting a number of avenues for further research.

Appendices

Appendix I – Candidate Gene Lists Derived from Individual Pathway Analysis Tools

DAVID									GO/PANTHER			GSEA
A400TCF4	*CAV2*	*CXCL8*	*FOSL1*	*IL6*	*MME*	*PREX1*	*SMARCA1*	*TNFAIP3*	*ANXA9*	*PDGFC*	*GCNT2*	*ASCL1*
ACSL4	*CBLN2*	*CXCR4*	*FOXA1*	*INPP5D*	*MMP14*	*PRLR*	*SNAI2*	*TNFRSF10D*	*ASCL1*	*PDGFD*	*HOMB13*	*CXCL12*
ADCY1	*CCRL2*	*CYP1A1*	*FOXC2*	*IRF6*	*MMP28*	*PROS1*	*SNCA*	*TNNI2*	*ASCL2*	*PPARG*	*IGFBP5*	*SOX3*
ADRB2	*CD109*	*DAB2*	*FOXQ1*	*IRX1*	*MPPED2*	*PSMB9*	*SOX2*	*TNS1*	*BDNF*	*RET*	*IL6*	*BMP5*
AFF3	*CD74*	*DAPK1*	*FPR1*	*ITGB8*	*MSLN*	*PTPRM*	*SOX3*	*TP63*	*BMP5*	*RGS20*	*JUP*	*SYT1*
AKR1B1	*CDH11*	*DCLK1*	*FREM2*	*IVL*	*MSN*	*PTPRZ1*	*SOX7*	*TPM2*	*BMP7*	*ROR1*	*KLK12*	*DSCAM*
AKR1C2	*CDH13*	*DCN*	*GAL*	*JUP*	*MYBPC1*	*RAC2*	*SPANXB1*	*UNC5B*	*BRINP2*	*ROS1*	*KLK5*	*OLFM1*
AKR1C3	*CDKN1C*	*DHRS2*	*GAS1*	*KCNJ3*	*MYLK*	*RARRES2*	*SPARC*	*VCAM1*	*BST2*	*SERPINA5*	*KRT13*	*FOXA1*
AL928654.3	*CEACAM5*	*DHRS9*	*GATA4*	*KCNJ8*	*NAV2*	*RASGEF1A*	*SPESP1*	*VCAN*	*CASP1*	*SERPINB13*	*KRT16*	*SOX2*
ALDH3A1	*CEACAM6*	*DLX1*	*GBP1*	*KCNK2*	*NLRP2*	*RASGRP1*	*SPINK4*	*VNN1*	*CAV2*	*SERPINB5*	*KRT23*	*PCP4*
ALOX15B	*CEBPD*	*DNALI1*	*GCNT2*	*KIF1A*	*NNMT*	*RBM24*	*SPINK6*	*WIPF1*	*CCND2*	*SERPINB7*	*KRT4*	*ASCL2*
ANO1	*CEMIP*	*DNER*	*GFRA1*	*KLK5*	*NOG*	*RBP1*	*SPOCK1*	*WLS*	*COL5A1*	*SERPINB9*	*KRT6A*	*ID4*
ANOS1	*CFB*	*DOCK8*	*GJA1*	*KLK6*	*NR4A2*	*RET*	*SPRR1A*	*WNT5A*	*CSTA*	*SERPINE2*	*KRT6B*	*KLK6*
ANPEP	*CFH*	*DOK5*	*GJB5*	*KLK7*	*NRCAM*	*RGCC*	*SPRR1B*	*ZBTB16*	*CXCL1*	*SOX2*	*KRT6C*	*BMP7*
ANXA8	*CFI*	*DSC2*	*GPM6B*	*KLK8*	*NRG1*	*RGS20*	*SPRR3*	*ZEB1*	*CXCL2*	*SOX7*	*MSN*	*COL3A1*
APOE	*CHST2*	*DSC3*	*GPX1*	*KRT13*	*NRK*	*RLN2*	*SRCIN1*	*ZIC1*	*CXCL8*	*SPATA18*	*OLFM1*	*PREX1*
ASCL1	*CLDN1*	*DSCAM*	*GSDME*	*KRT14*	*NRXN3*	*ROR1*	*SRGN*		*EGFR*	*SPRR1A*	*PDPN*	*EPHA7*
ASCL2	*CLDN3*	*DST*	*GSTM3*	*KRT16*	*NUP210L*	*ROS1*	*ST8SIA4*		*EPHA7*	*SPRR1B*	*PGR*	*NRCAM*
AXL	*CLMP*	*DUOX1*	*GSTP1*	*KRT19*	*OLFM1*	*RTN1*	*SULF1*		*ERBB2*	*SPRR3*		*PACSIN1*
BCL11A	*CLTA*	*DUSP6*	*HBEGF*	*KRT4*	*OLFML3*	*S100A2*	*SULF2*		*ETS1*	*SULF1*		*RET*
BDNF	*CNN3*	*EDIL3*	*HIPK1*	*KRT5*	*PACSIN1*	*S100A7*	*SYT1*		*ETS2*	*SULF2*		*DLX1*
BIN1	*COL13A1*	*EDNRB*	*HMGA2*	*KRT6A*	*PCDH9*	*S100A8*	*TBC1D30*		*ETV1*	*TCF4*		*TBC1D30*
BIRC3	*COL4A1*	*EFHD1*	*HNMT*	*KRT6B*	*PCDHB16*	*SAA1*	*TBX2*		*FOXA1*	*TENM3*		*SBF2*
BMP5	*COL4A2*	*EGFR*	*HOXB13*	*LAMA3*	*PCDHB2*	*SATB1*	*TBX3*		*FOXC2*	*TFP1*		*NRG1*
BMP7	*COL5A1*	*ELF5*	*HOXB2*	*LAMC2*	*PCDHB5*	*SCEL*	*TCF4*		*FOXQ1*	*TFP12*		*LPAR1*
BNC1	*COL8A1*	*EMP1*	*HOXC10*	*LIN28A*	*PCP4*	*SERPINA5*	*TENM3*		*HBEGF*	*TGFB2*		*ABCA1*
BRINP2	*CREB3L1*	*EPHA7*	*HPGD*	*LOX*	*PDE4D*	*SERPINB13*	*TFCP2L1*		*HES2*	*TGFB3*		*TGFB2*
BST2	*CRIP1*	*ERBB2*	*HSD17B2*	*LOXL2*	*PDGFC*	*SERPINB2*	*TFPI*		*ID4*	*TIMP3*		*PLCE1*
C3	*CRISPLD1*	*ESR1*	*ID4*	*LPAR1*	*PDGFD*	*SERPINB5*	*TFPI2*		*IFIT1*	*TYRP1*		
C4BPB	*CSGALNACT1*	*ETS1*	*IFI16*	*LRRN3*	*PDLIM3*	*SERPINB7*	*TGFB1I1*		*IFIT3*	*ALOX15B*		
CA2	*CST1*	*ETV1*	*IFI27*	*LY6K*	*PDPN*	*SERPINB9*	*TGFB2*		*KCNMA1*	*APOE*		
CADM1	*CST4*	*EXT1*	*IGF2BP2*	*MAGEA3*	*PI3*	*SERPINE1*	*TGFB3*		*LIFR*	*CAV1*		
CALB1	*CST6*	*F2R*	*IGF2BP3*	*MALL*	*PID1*	*SERPINE2*	*TGFBI*		*LOX*	*CREB3L1*		
CALCR	*CSTA*	*F3*	*IGFBP1*	*MAMLD1*	*PLAC8*	*SFRP1*	*TGFBR3*		*LOXL2*	*CTGF*		
CALD1	*CTGF*	*FAP*	*IGFBP5*	*MAOB*	*PLAGL1*	*SH3PXD2A*	*TGM1*		*LOXL4*	*DAB2*		
CALML5	*CTHRC1*	*FAT2*	*IGFBP7*	*MAP1B*	*PLAT*	*SHANK2*	*TGM2*		*MX2*	*DSC2*		
CARD16	*CXCL1*	*FBLN1*	*IL15*	*MCC*	*PLAU*	*SLC16A1*	*THBD*		*MYLK*	*DSC3*		
CASP1	*CXCL12*	*FBN2*	*IL18*	*MELTF*	*PLCE1*	*SLC1A3*	*TIMP3*		*NOG*	*F2R*		
CAV1	*CXCL14*	*FLRT3*	*IL1A*	*MFAP5*	*PLXNA2*	*SLC40A1*	*TLR3*		*OASL*	*F3*		
	CXCL2	*FOS*	*IL1RN*	*MGP*	*PPARG*	*SLPI*	*TNC*		*OSMR*	*FAP*		

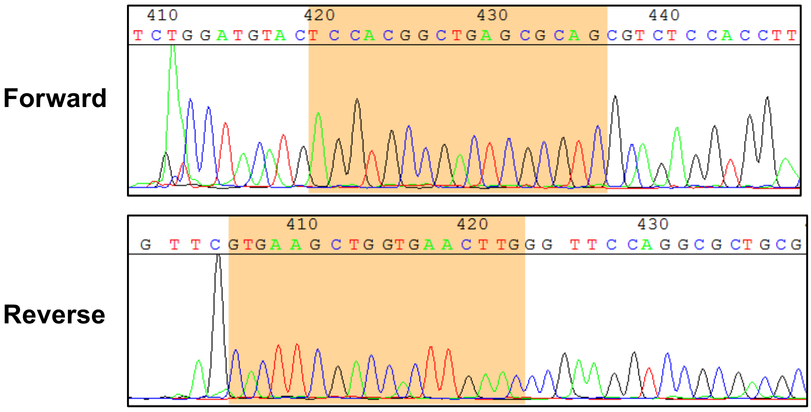
Appendix II – Optimisation of siRNA Knockdown of *ASCL2*

Experiment & Action	Troubleshooting Conditions	Notes
Used positive control (*HPRT*) to determine best siRNA dosage Fluorescence (TYE 563) used to check efficiency prior to qPCR	10nM, 1nM, 0.1nM *HPRT* & with corresponding Dharmafect (DF) 3x10^4 cells per well (24 well plate)	Fluorescence observed Highest knockdown was at 10nm, however poor efficiency (45%)
Tested a higher DF concentration	10nM siRNA + 25nM DF 10nM siRNA + 10nM DF 3x10^4 cells per well (24 well)	Fluorescence observed Knockdown improved with higher concentration of transfection reagent (DF), however poor efficiency
Checked effect of cell density on knockdown efficiency	10nM siRNA/25nM DF 3, 4, & 5 x10^4 per well (24 well)	Fluorescence observed No knockdown
New transfection reagent tested (Lipofectamine RNAiMAX) Increased cell density to account for high volume of cells dying during transfection	10pmol siRNA + 1.5µl Lipofectamine Tested against 10nM siRNA + 10nM DF 2x10^5 cells per well (24 well)	Fluorescence observed Inconclusive results - RNA concentration was too poor. Cells were aggregating in the centre of the well and were not growing or adhering evenly. Cells had almost all detached 48 hours after transfection.
Larger wells (6 well plate) to see if greater RNA would enhance qPCR estimation	30pmol siRNA + 9µl Lipofectamine 1x10^6 cells per well	Fluorescence observed RNA purity was extremely poor for the *HPRT* transfected sample. This housekeeping gene is crucial for cell development, and therefore affecting ability of cells to stay alive. Therefore, *ASCL2* will be tested, before considering a new kit.
Checked individual *ASCL2* siRNA oligonucleotides	10pmol *HPRT, ASCL2* + 1.5 ul Lipofectamine	Fluorescence observed Inconclusive knockdown RNA purity poor – this needed to be addressed
Attempted reverse transfection	10pmol *HPRT, ASCL2* + 1.5 ul Lipofectamine	Fluorescence could not be observed as cells had died Reverse transfection was not effective, therefore reverted back to original transfection.
Pooled 3x*ASCL2* oligos Seeded more cells and care taken to ensure cells seeded equally – 5x10^4 New RNA Microprp kit was used and cells were scraped from wells	10nM *HPRT* and pooled *ASCL2* (x3 oligos) + 1.5 ul Lipofectamine	Fluorescence observed Sufficient knockdown was measured. ~70% Cells had grown evenly and RNA purity was good. The Microprep kit was used for the proceeding experiments and care was taken to seed cells evenly going forward
Extracted 24hrs post transfection (rather than 48hrs) to minimise amount of cells detaching	10nM *HPRT* and pooled *ASCL2* (x3 oligos) + 1.5 ul Lipofectamine	Fluorescence observed Knockdown successful >70% Transfection time of 24hrs was used for proceeding experiments

Appendix III – Sanger Sequencing of *ASCL2* & siRNA sequences

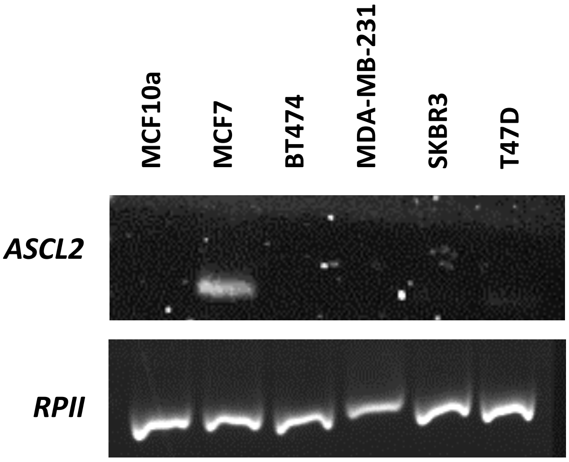
GATC LightRun Sanger Sequencing

Chromatogram confirming the sequence of primers and PCR products as *ASCL2*. The forward and reverse primers are shown.



PCR Analysis

Varied gene expression of *ASCL2* across breast cancer cell lines, demonstrated by PCR analysis. This was also used to demonstrate sample integrity, using a reference gene, *RPII* as a control.



Gene sequence of *ASCL2* with siRNA sequences highlighted

Sequence from Ensembl (www.ensembl.org)

```
CAGACCTCCAGGCCCTCCGGGTTAAGGTGCCGCCCAGAGCCCTCAGGCCGGGGGCGCACGGAAACCACAGGCAG
GGTGCGCGTGGAGGGACGGGGAAAGCGGGGCGGGTTGGGGAAGGCGCCCCGGGAACCTGAACCTCCCACCCCGC
CTCAGTCTCGACCACTCCTTAAGCCCCACCCCGCCCCAGGTAAGGCGCAGTCCACCCCCATTCCCAGTAGATTA
ACGCACAGGTGGGGGCGCGCTCGGGACATAGCTGCGCTAGGGGACAGCGCGCCCAGCCCAGTCGCGGGGGCGAG
GAGCAGGGCGGGGCCCAGCAGGAACCCAGCTTTGTTAGCGATGCTCCCCGTGAGCCACGCGCCACGCGTACGCG
CTTCCTCAATGGGGCCGGGCGTGGAGCCGCGCCCTGCGCGATTGGCCAAACGGGTGGCCCACGATTGGCTGAGA
CCCTGGCCCCCGCCTCCTCGGCCCCAGGAGGGTGGGGCGTGGGTGTGGGCTGCGCGGCGCGTGCTGCCCCCGGG
GATCTTGCGCGCCTCCCGAACAGCCGTGTTGTCGCCAGGGCCGCGCCTTCCCTCCCACAGCGCGCGCTGCGCGT
GCGAAGGT
```

EXON 1

```
CTGGCGGCTCTTGGGACTGGCGGGGCTGCGCGCGGGGTTAGGGTGGGGGTACGGGAAGGCTCAACCCAGGACCT
GCGTACCTTGCTTTGGGGGCGCACTAAGCACCTGCCGGGAGCAGGGGGCGCACCGGGAACTCGCAGATTTCGCC
AGTTGGGCGCACTGGGGATCTGTGGACTGCGTCCGGGGGATGGGCTAGGGGGACATGCGCACGCTTTGGGCCTT
ACAGAATGTGATCGCGCGAGGGGGAGGGCGAAGCGTGGCGGGAGGGCGAGGCGAAGGAAGGAGGGCGTGAGAAA
GGCGACGGCGGCGGCGCGGAGGAGGGTTATCTATACATTTAAAAACCAGCCGCCTGCGCCGCGCCTGCGGAGAC
CTGGGAGAGTCCGGCCGCACGCGCGGGACACGAGCGTCCCACGCTCCCTGGCGCGTACGGCCTGCCACCACTAG
GCCTCCTATCCCCGGGCTCCAGACGACCTAGGACGCGTGCCCTGGGGAGTTGCCTGGCGGCGCCGTGCCAGAAG
CCCCCTTGGGGCGCCACAGTTTTCCCCGTCGCCTCCGGTTCCTCTGCCTGCACCTTCCTGCGGCGCGCCGGGAC
CTGGAGCGGGCGGGTGGATGCAGGCGCGATGGACGGCGGCACACTGCCCAGGTCCGCGCCCCCTGCGCCCCCCG
TCCCTGTCGGCTGCGCTGCCCGGCGGAGACCCGCGTCCCCGGAACTGTTGCGCTGCAGCCGGCGGCGGCGACCG
GCCACCGCAGAGACCGGAGGCGGCGCAGCGGCCGTAGCGCGGCGCAATGAGCGCGAGCGCAACCGCGTGAAGCT
GGTGAACTTGGGCTTCCAGGCGCTGCGGCAGCACGTGCCGCACGGCGGCGCCAGCAAGAAGCTGAGCAAGGTGG
AGACGCTGCGCTCAGCCGTGGAGTACATCCGCGCGCTGCAGCGCCTGCTGGCCGAGCACGACGCCGTGCGCAAC
GCGCTGGCGGGAGGGCTGAGGCCGCAGGCCGTGCGGCCGTCTGCGCCCCGCGGGCCGCCAGGGACCACCCCGGT
CGCCGCCTCGCCCTCCCGCGCTTCTTCGTCCCCGGGCCGCGGGGGCAGCTCGGAGCCCGGCTCCCCGCGTTCCG
CCTACTCGTCGGACGACAGCGGCTGCGAAGGCGCGCTGAGTCCTGCGGAGCGCGAGCTACTCGACTTCTCCAGC
TGGTTAGGGGGCTACTGAGCGCCCTCGACCTATGAG
```

```
GTAACAGCCGGGAGGCAGGGAGGAGGGAGGGCCGGGGGCCGGGGTGGAGGGACGGGGTGGGCAGGCCCGGCGGG
TCGCGCCCCCAGGAGCCCGCGGAGCCGAGCGCCAGGCCCGAGCGATGGCTTCGATTTCGCTCACTCTTCATTTC
CCCCAAAGTTTTTCAAGCCCGTGCAAGACCGGCGTTTGTTTGTCCGGGATTGCAAAACTTCCCCTCGCGGCTCA
GCCGCCGACGAGGGAGGGGTAGACGAGGGGAGGGGAGCGGCCGTCGGGCCGTTGAGGTCTCTAGTGCTGGCGGA
TCCTGGGGCAGATTGGGGTGCTGGAGGCGGGGTGACTTTGCATTGCAAATCGCGCTCCCGGGCCGGGGCGGCAG
AAATGAGTCGGCGGGCGCGGAGCCCTGACTCACCGCGGCTCCGAGCGCCCGCCCCGCCCCCGCCGTGTCTCAGA
CCGAGTCGCGGCACCCACGGACTCAAGACTCCAAAACCAACCGAGCAAACGAAACTGCCGACTTCGCTTGGGGG
AGGTGCGGGCAGGGCCGGCCCGGGCGGGGTCTGCCCCGGGCCCGCGCCCGCGTTGACGCGCGTTTGGTTCCCCA
CCTTCCCCCCGCAG
```

EXON 2

```
CCTCAGCCCCGGAAGCCGAGCGAGCGGCCGGCGCGCTCATCGCCGGGGAGCCCGCCAGGTGGACCGGCCCGCGC
TCCGCCCCCAGCGAGCCGGGGACCCACCCACCACCCCCCGCACCGCCGACGCCGCCTCGTTCGTCCGGCCCAGC
CTGACCAATGCCGCGGTGGAAACGGGCTTGGAGCTGGCCCCATAAGGGCTGGCGGCTTCCTCCGACGCCGCCCC
TCCCCACAGCTTCTCGACTGCAGTGGGGCGGGGGGCACCAACACTTGGAGATTTTTCCGGAGGGGAGAGGATTT
TCTAAGGGCACAGAGAATCCATTTTCTACACATTAACTTGAGCTGCTGGAGGGACACTGCTGGCAAACGGAGAC
CTATTTTTGTACAAAGAACCCTTGACCTGGGGCGTAATAAAGATGACCTGGACCCCTGCCCCCACTATCTGGAG
TTTTCCATGCTGGCCAAGATCTGGACACGAGCAGTCCCTGAGGGGCGGGGTCCCTGGCGTGAGGCCCCCGTGAC
AGCCCACCCTGGGGTGGGTTTGTGGGCACTGCTGCTCTGCTAGGGAGAAGCCTGTGTGGGGCACACCTCTTCAA
GGGAGCGTGAACTTTATAAATAAATCAGTTCTGTTTACCA
```

```
GTGGCTCCTATCACCTACACTTCCCAGGTGACGGCCAGACTTCCGTGGTCACTACTCCTCAAACCCTGCTGCCT
CCTCCGTAGGGTGGGTCTGGGTGAGATCTGGAGTGCAGCCAGGCCGTTGATAGCGGAGCCATTGGGACACCTTG
TGAGGCTGGGGGCATCCTCCAGGAGGTGGTGGGCTGGTGGGTTGTCCAGACAGGGCTACTCGCTGGCTTGGAAG
CTGCAGGCTGGAGGCTGCTGACCCATCCCGAGGGCTGGGGTAAGTGCTGGGTGTGGGGCTAGGCTGAGGTGGTC
TGACCAGAGAGCACCGGCTGTGGGGCTGAGGGCATGGGCTCCTGCGCAGGCCACCACGCTCAGATCTCCACTAA
CGTGGCAGCTGGGCAGCCCAGGGCAAGTGGGTTAACTTGCAAATGGGTTTGACCAGACCCACCTCAACGGCCTC
TGGGAGGAGTTAGTGAGAGGTGCCTGGAGGCTGCCCTCTCGCTAGCTTTGGGTTTTGCCCGCACTGGGGAGGCC
CTGCAGGTCTCCGCTCACCTGAATTCTAAGAGCGGCTCTTGAAAGGAACAAGGAAGGCTTGGAAGCTTTGCGCC
AGGCTCCC
```

Appendix IV – Experimental Descriptive Data

ASCL2 Extreme Variation Analysis Descriptive Statistics

Cell Line	Sample	Mean Log 2 Expression, *ASCL2*	Std. Deviation	P
MCF10A	n=10	2.5	0.16	
BT474	n=7	9.16	0.33	<0.0001
MCF7	n=9	8.84	0.72	<0.0001
MDA-MB-231	n=4	2.60	0.28	0.46
T47D	n=7	5.11	0.32	<0.0001
ZR-75	n=3	5.41	2.51	0.0018

ASCL2 RT-qPCR Descriptive Statistics

Cell Line	ΔCT				Relative Fold Change			
	n=1	n=2	Mean	Std. Deviation	n=1	n=2	Mean	Std. Deviation
MCF10A	13.26	14.1	13.68	0.59	1	1	1	0
MCF7	7.06	6.17	6.615	0.63	73.26	244.72	158.99	121.24
MDA-MB-231	13.6	14.48	14.04	0.62	0.79	0.77	0.78	0.01
BT474	12	14.03	13.015	1.44	2.39	1.05	1.72	0.95
SKBR3	9.70	10.1	9.9	0.28	11.79	16	13.90	2.97
T47D	8.89	9.25	9.07	0.25	20.61	28.94	24.77	5.89

ASCL2 Knockdown Descriptive Statistics

Relative mRNA expression of *ASCL2* in experimental samples (siRNA-ASCL2) compared to non-targeting control (siRNA-NC), across cell lines.

Cell Line	Sample	n=1	n=2	n=3	n=4	n=5	Mean Fold Change	Percentage Knockdown	Std. Deviation
MCF7 n=5	ASCL2 NC	0.32 1	0.30 1	0.38 1	0.47 1	0.23 1	0.34 1	66% 0	0.09 0
T47D n=4	ASCL2 NC	0.19 1	0.35 1	0.43 1	0.21 1	-	0.30 1	70% 0	0.11 0
SKBR3 n=3	ASCL2 NC	0.09 1	0.34 1	-	-	-	0.20 1	80% 0	0.12 0

Wound-Healing Assay Descriptive Statistics

Area of wound (μm)

		siRNA	n=1	n=2	n=3	n=4	n=5	n=6	Mean Area (μm)	Std. Deviation
MCF7	0h	ASCL2	383679	551558	362271	449259	374399	304557	404287	85721
		NC	378105	448258	409788	426729	398355	241273	383751	73799
	24h	ASCL2	211266	378455	359391	398369	234676	228101	301710	85618
		NC	165160	247292	321914	325117	205800	161382	237778	73390
	48h	ASCL2	127900	336536	332646	380924	193438	178806	258375	104130
		NC	90722	189955	300190	289975	174943	113718	193250	87149
T47D	0h	ASCL2	444636	391440	-	-	-	-	418038	37615
		NC	393166	376998	-	-	-	-	385082	11432
	24h	ASCL2	404135	349472	-	-	-	-	376804	38653
		NC	341591	311557	-	-	-	-	326574	21237
	48h	ASCL2	394491	320579	-	-	-	-	357535	52263
		NC	310502	312128	-	-	-	-	311315	1149
SKBR3	0h	ASCL2	378740	383884	442193	-	-	-	401606	35244
		NC	409191	478284	393282	-	-	-	426919	45189
	24h	ASCL2	357963	372861	442766	-	-	-	391197	45277
		NC	389790	451297	383435	-	-	-	408174	37480
	48h	ASCL2	342535	359288	407798	-	-	-	369874	33895
		NC	368388	437729	372131	-	-	-	392749	38999

Percentage closure after 48h

	siRNA	n=1	n=2	n=3	n=4	n=5	n=6	Mean % Closure
MCF7	ASCL2	66.7	39	8.2	15	48.3	41.3	13.8
	NC	76	57.6	26.7	32	56.1	52.9	
	% Closure	9.3	18.6	18.5	17	7.8	11.6	
T47D	ASCL2	11.3	18.1	-	-	-	-	9.01
	NC	26.6	20.8	-	-	-	-	
	% Closure	15.3	2.7	-	-	-	-	
SKBR3	ASCL2	9.6	6.4	7.8	-	-	-	0.028
	NC	10	8.5	5.4	-	-	-	
	% Closure	0.4	2.1	-2.4	-	-	-	

Appendix V – Data Mining Descriptive Statistics & Frequency Tables of *ASCL2* Expression

Age of Onset	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Downregulated	1	75.1300	-	-	-	-	75.13	75.13
Unaltered	1821	61.0495	12.98345	0.30425	60.4528	61.6463	21.93	96.29
Upregulated	82	61.7487	12.92019	1.42680	58.9098	64.5875	32.99	87.18

Overall Survival	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Downregulated	1	58.6667	-	-	-	-	58.67	58.67
Unaltered	1821	125.5265	76.25047	1.78685	122.0220	129.0310	0.00	355.20
Upregulated	82	114.7004	78.03784	8.61784	97.5536	131.8472	9.60	297.23

Stage		.00	Stage 1	Stage 2	Stage 3	Stage 4
Downregulated	Count	0	0	0	0	0
	Row N %	0.0%	0.0%	0.0%	0.0%	0.0%
Unaltered	Count	4	460	766	106	9
	Row N %	0.3%	34.2%	57.0%	7.9%	0.7%
Upregulated	Count	0	15	34	9	0
	Row N %	0.0%	25.9%	58.6%	15.5%	0.0%

Grade		Grade1	Grade 2	Grade 3	Grade 4
Downregulated	Count	0	0	0	0
	Row N %	0.0%	0.0%	0.0%	0.0%
Unaltered	Count	160	711	882	0
	Row N %	9.1%	40.6%	50.3%	0.0%
Upregulated	Count	5	29	45	0
	Row N %	6.3%	36.7%	57.0%	0.0%

ASCL2 Expression	**Subtype**					
	Luminal A	Luminal B	HER2 Positive	Basal	Claudin Low	Normal
Mean	-0.12	-0.18	0.76	0.08	0.04	0.02
Median	-0.4	-0.53	0.85	-0.03	-0.05	-0.2
SD	0.92	1.07	1.01	0.94	0.69	0.89

Subtype		Luminal A	Luminal B	HER2 Positive	Basal	Claudin Low	Normal
Downregulated	Count	1	0	0	0	0	0
	Row N %	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Unaltered	Count	660	434	193	192	199	137
	Row N %	36.4%	23.9%	10.6%	10.6%	11.0%	7.5%
Upregulated	Count	18	27	27	7	0	3
	Row N %	22.0%	32.9%	32.9%	8.5%	0.0%	3.7%

Integrative Cluster	Downregulated		Unaffected		Upregulated	
	Count	Row N %	Count	Row N %	Count	Row N %
Cluster 1	0	0.0%	126	6.9%	6	7.3%
Cluster 2	0	0.0%	68	3.7%	4	4.9%
Cluster 3	0	0.0%	275	15.1%	7	8.5%
Cluster 4ER+	0	0.0%	238	13.1%	6	7.3%
Cluster 4ER-	0	0.0%	71	3.9%	3	3.7%
Cluster 5	0	0.0%	163	9.0%	21	25.6%
Cluster 6	0	0.0%	81	4.4%	3	3.7%
Cluster 7	0	0.0%	176	9.7%	6	7.3%
Cluster 8	1	100.0%	279	15.3%	9	11.0%
Cluster 9	0	0.0%	132	7.2%	10	12.2%
Cluster 10	0	0.0%	212	11.6%	7	8.5%

Receptor Status		**ER**		**HER2**		**PR**	
		Positive	Negative	Positive	Negative	Positive	Negative
Downregulated	Count	1	0	0	1	1	0
	Row N %	100.0%	0.0%	0.0%	100.0%	100.0%	0.0%
Unaltered	Count	1396	425	215	1606	971	850
	Row N %	76.7%	23.3%	11.8%	88.2%	53.3%	46.7%
Upregulated	Count	62	20	21	61	37	45
	Row N %	75.6%	24.4%	25.6%	74.4%	45.1%	54.9%

References

Ahern, T.P., Lash, T.L., Egan, K.M. and Baron, J.A., (2009). Lifetime tobacco smoke exposure and breast cancer incidence. *Cancer Causes & Control,* 20(10), 1837-1844.

Alexandrov, L.B., Ju, Y.S., Haase, K., Van Loo, P., Martincorena, I., Nik-Zainal, S., Totoki, Y., Fujimoto, A., Nakagawa, H., Shibata, T., Campbell, P.J., Vineis, P., Phillips, D.H. and Stratton, M.R., (2016). Mutational signatures associated with tobacco smoking in human cancer. *Science*, 354(6312), 618–622.

Alhamdoosh, M., Ng, M., Wilson, N.J., Sheridan, J.M., Huynh, H., Wilson, M.J. and Ritchie, M.E., (2017). Combining multiple tools outperforms individual methods in gene set enrichment analyses. *Bioinformatics*, 33(3), 414–424.

Ali, S.M., Watson, J., Wang, K., Chung, J.H., McMahon, C., Ross, J.S. and Dicke, K.A., (2016). A Combination of Targeted Therapy with Chemotherapy Backbone Induces Response in a Treatment-Resistant Triple-Negative MCL1-Amplified Metastatic Breast Cancer Patient. *Case Reports in Oncology*, 9(1), 112–118.

Anastas, J.N. and Moon, R.T., (2013). WNT signalling pathways as therapeutic targets in cancer. *Nature reviews. Cancer*, 13(1), 11–26.

Anders, C.K., Johnson, R., Litton, J., Phillips, M. and Bleyer, A., (2009). Breast cancer before age 40 years. *Seminars in oncology*, 36(3), 237–49.

Antoniou, A.C., Casadei, S., Heikkinen, T., Barrowdale, D., Pylkäs, K., Roberts, J., Lee, A., Subramanian, D., De Leeneer, K., Fostira, F., Tomiak, E., Neuhausen, S.L., Teo, Z.L., Khan, S., Aittomäki, K., Moilanen, J.S., Turnbull, C., Seal, S., Mannermaa, A., *et al.,* (2014). Breast-cancer risk in families with mutations in PALB2. *The New England journal of medicine*, 371(6), 497–506.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., *et al.,* (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1), 25–29.

Auer, H., Newsom, D.L. and Kornacker, K., (2009). Expression Profiling Using Affymetrix GeneChip Microarrays. In Humana Press, 35–46.

Augustyn, A., Borromeo, M., Wang, T., Fujimoto, J., Shao, C., Dospoy, P.D., Lee, V., Tan, C., Sullivan, J.P., Larsen, J.E. & Girard, L., (2014). ASCL1 is a lineage oncogene providing therapeutic targets for high-grade neuroendocrine lung cancers. *Proceedings of the National Academy of Sciences*, 111(41), 14788-14793.

Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D. a and Shendure, J., (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nature reviews. Genetics*, 12(11), 745–55.

Bareche, Y., Venet, D., Ignatiadis, M., Aftimos, P., Piccart, M., Rothe, F. and Sotiriou, C., (2018). Unravelling triple-negative breast cancer molecular heterogeneity using an integrative multiomic analysis. *Annals of Oncology*, 29(4), 895–902.

Barrett, T., Suzek, T.O., Troup, D.B., Wilhite, S.E., Ngau, W.-C., Ledoux, P., Rudnev, D., Lash, A.E., Fujibuchi, W. and Edgar, R., (2005). NCBI GEO: mining millions of expression profiles--database and tools. *Nucleic acids research*, 33(Database issue), D562-6.

Bass, J.J., Wilkinson, D.J., Rankin, D., Phillips, B.E., Szewczyk, N.J., Smith, K. and Atherton, P.J., (2017). An overview of technical considerations for Western blotting applications to physiological research. *Scandinavian Journal of Medicine & Science in Sports*, 27(1), 4–25.

Basu, S., Gavert, N., Brabletz, T. and Ben-Ze'ev, A., (2018). The intestinal stem cell regulating gene ASCL2 is required for L1-mediated colon cancer progression. *Cancer Letters*. 28(424), 9-18.

Basu, S., Cheriyamundath, S. and Ben-Ze'ev, A., (2018). Cell–cell adhesion: linking Wnt/β-catenin signaling with partial EMT and stemness traits in tumorigenesis. *F1000Research*, 7, 1488.

Basu, S., Haase, G. and Ben-Ze'ev, A., (2016). Wnt signaling in cancer stem cells and colon cancer metastasis. *F1000 Research*, 5.

Bayer, I., Groth, P. and Schneckener, S., (2013). Prediction Errors in Learning Drug Response from Gene Expression Data – Influence of Labeling, Sample Size, and Machine Learning Algorithm V. Brusic, ed. *PLoS ONE*, 8(7), e70294.

Beaber, E.F., Buist, D.S.M., Barlow, W.E., Malone, K.E., Reed, S.D. and Li, C.I., (2014). Recent oral contraceptive use by formulation and breast cancer risk among women 20 to 49 years of age. *Cancer research*, 74(15), 4078–89.

Bedard, P.L., Hansen, A.R., Ratain, M.J. and Siu, L.L., (2013). Tumour heterogeneity in the clinic. *Nature*, 501(7467), 355–364.

Biau, D.J., Kernéis, S. and Porcher, R., (2008). Statistics in brief: the importance of sample size in the planning and interpretation of medical research. *Clinical orthopaedics and related research*, 466(9), 2282–8.

Bio-Rad Laboratories, Inc, (2016). *AlamarBlue® technical datasheet*, Available from: https://www.bio-rad-antibodies.com/static/uploads/ifu/buf012a.pdf.

Biocompare, (2012). Cellular Toxicity Caused by Transfection: Why is it important? Available from: https://www.biocompare.com/Bench-Tips/121111-Cellular-Toxicity-Caused-by-Transfection-Why-is-it-important/.

Block, I., Burton, M., Sørensen, K.P., Andersen, L., Larsen, M.J., Bak, M., Cold, S., Thomassen, M., Tan, Q. and Kruse, T.A., (2018). Association of miR-548c-5p, miR-7-5p, miR-210-3p, miR-128-3p with recurrence in systemically untreated breast cancer. *Oncotarget*, 9(10), 9030–9042.

Boettcher, M. and McManus, M.T., (2015). Choosing the Right Tool for the Job: RNAi, TALEN, or CRISPR. *Molecular Cell*, 58(4), 575–585.

Bogutz, A.B., Oh-McGinnis, R., Jacob, K.J., Ho-Lau, R., Gu, T., Gertsenstein, M., Nagy, A. and Lefebvre, L., (2018). Transcription factor ASCL2 is required for development of the glycogen trophoblast cell lineage G. S. Barsh, ed. *PLOS Genetics*, 14(8), e1007587.

Bolós, V., Mira, E., Martínez-Poveda, B., Luxán, G., Cañamero, M., Martínez-A, C., Mañes, S. and de la Pompa, J.L., (2013). Notch activation stimulates migration of breast cancer cells and promotes tumor growth. *Breast cancer research : BCR*, 15(4), R54.

Bos, P.D., Zhang, X.H.-F., Nadal, C., Shu, W., Gomis, R.R., Nguyen, D.X., Minn, A.J., van de Vijver, M.J., Gerald, W.L., Foekens, J.A. and Massagué, J., (2009). Genes that mediate breast cancer metastasis to the brain. *Nature*, 459(7249), 1005–1009.

Borromeo, M.D., Savage, T.K., Kollipara, R.K., He, M., Augustyn, A., Osborne, J.K., Girard, L., Minna, J.D., Gazdar, A.F., Cobb, M.H. & Johnson, J.E., (2016). ASCL1 and NEUROD1 reveal heterogeneity in pulmonary neuroendocrine tumors and regulate distinct genetic programs. Cell reports, 16(5), 1259-1272.

Braune, E.-B., Seshire, A., Lendahl, U., Braune, E.-B., Seshire, A. and Lendahl, U., (2018). Notch and Wnt Dysregulation and Its Relevance for Breast Cancer and Tumor Initiation. *Biomedicines*, 6(4), E101.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A. and Jemal, A., (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6), 394–424.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., Gaasterland, T., Glenisson, P., Holstege, F.C.P., Kim, I.F., Markowitz, V., Matese, J.C., Parkinson, H., Robinson, A., Sarkans, U., *et al.,* (2001). Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature Genetics*, 29(4), 365–371.

Brennan, S.F., Cantwell, M.M., Cardwell, C.R., Velentzis, L.S. and Woodside, J. V, (2010). Dietary patterns and breast cancer risk: a systematic review and meta-analysis. *The American journal of clinical nutrition*, 91(5), 1294–302.

Broad Institute, (2019). GSEA User Guide. *Gene Set Enrichment Analysis*. Available from: https://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/FAQ.

Buermans, H.P.J. and den Dunnen, J.T., (2014). Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1842(10), 1932–1941.

Bumgarner, R., (2013). Overview of DNA microarrays: types, applications, and their future. *Current protocols in molecular biology*, Chapter 22, Unit 22.1.

Burdall, S.E., Hanby, A.M., Lansdown, M.R.J. and Speirs, V., (2003). Breast cancer cell lines: friend or foe? *Breast cancer research : BCR*, 5(2), 89–95.

Byler, S., Goldgar, S., Heerboth, S., Leary, M., Housman, G., Moulton, K. and Sarkar, S., (2014). Genetic and epigenetic aspects of breast cancer progression and therapy. *Anticancer research*, 34(3), 1071–7.

Cancer Research UK, (2019). Breast Cancer Statistics. Available from: http://publications.cancerresearchuk.org/cancerstats/statsbreast/kfbreast.html.

Catsburg, C., Miller, A.B. and Rohan, T.E., (2015). Active cigarette smoking and risk of breast cancer. *International Journal of Cancer*, 136(9), 2204–2209.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A.P., Sander, C. and Schultz, N., (2012). The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data: Figure 1. *Cancer Discovery*, 2(5), 401–404.

Cerami, E.G., Gross, B.E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., Schultz, N., Bader, G.D. and Sander, C., (2011). Pathway Commons, a web resource for biological pathway data. *Nucleic acids research*, 39(Database issue), D685-90.

Chaffer, C.L., San Juan, B.P., Lim, E. and Weinberg, R.A., (2016). EMT, cell plasticity and metastasis. *Cancer and Metastasis Reviews*, 35(4), 645–654.

Chen, L., Xiao, Z., Meng, Y., Zhao, Y., Han, J., Su, G., Chen, B. and Dai, J., (2012). The enhancement of cancer stem cell properties of MCF-7 cells in 3D collagen scaffolds for modeling of cancer and anti-cancer drugs. *Biomaterials*, 33(5), 1437–1444.

Chen, X., Duan, N., Zhang, C. and Zhang, W., (2016). Survivin and Tumorigenesis: Molecular Mechanisms and Therapeutic Strategies. *Journal of Cancer*, 7(3), 314–23.

Chen, Y., Shi, H.Y., Stock, S.R., Stern, P.H. and Zhang, M., (2011). Regulation of Breast Cancer-induced Bone Lesions by β-Catenin Protein Signaling. *Journal of Biological Chemistry*, 286(49), 42575–42584.

Cheng, P., Dummer, R. and Levesque, M., (2015). Data mining The Cancer Genome Atlas in the era of precision cancer medicine. *Swiss Medical Weekly*, 145(3738).

Choi, S.Y.C., Lin, D., Gout, P.W., Collins, C.C., Xu, Y. and Wang, Y., (2014). Lessons from patient-derived xenografts for better in vitro modeling of human cancer. *Advanced drug delivery reviews*, 79–80, 222–37.

Christensen, B.C., Kelsey, K.T., Zheng, S., Houseman, E.A., Marsit, C.J., Wrensch, M.R., Wiemels, J.L., Nelson, H.H., Karagas, M.R., Kushi, L.H., Kwan, M.L. and Wiencke, J.K., (2010). Breast cancer DNA methylation profiles are associated with tumor size and alcohol and folate intake. *PLoS genetics*, 6(7), e1001043.

Christie, J.D., (2005). Microarrays. *Critical care medicine*, 33(12 Suppl), S449-52.

Chung, W., Eum, H.H., Lee, H.-O., Lee, K.-M., Lee, H.-B., Kim, K.-T., Ryu, H.S., Kim, S., Lee, J.E., Park, Y.H., Kan, Z., Han, W. and Park, W.-Y., (2017). Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nature Communications*, 8, 15081.

Conway, K., Edmiston, S.N., May, R., Kuan, P.F., Chu, H., Bryant, C., Tse, C.-K., Swift-Scanlan, T., Geradts, J., Troester, M.A. and Millikan, R.C., (2014). DNA methylation profiling in the Carolina Breast Cancer Study defines cancer subclasses differing in clinicopathologic characteristics and survival. *Breast cancer research : BCR*, 16(5), 450.

Cronin, M., Sangli, C., Liu, M.-L., Pho, M., Dutta, D., Nguyen, A., Jeong, J., Wu, J., Langone, K.C. and Watson, D., (2007). Analytical Validation of the Oncotype DX Genomic Diagnostic Test for Recurrence Prognosis and Therapeutic Response Prediction in Node-Negative, Estrogen Receptor–Positive Breast Cancer. *Clinical Chemistry*, 53(6), 1084-91.

Curtis, C., Shah, Sohrab P., Chin, S.-F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., Gräf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., Langerød, A., Green, A., Provenzano, E., *et al.,* (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403), 346–352.

Dai, X., Li, T., Bai, Z., Yang, Y., Liu, X., Zhan, J. and Shi, B., (2015). Breast cancer intrinsic subtype classification, clinical use and future trends. *American journal of cancer research*, 5(10), 2929–43.

Dawson, S.-J., Rueda, O.M., Aparicio, S., Caldas, Carlos, El-Rehim, D.A., Ball, G., Pinder, S., Rakha, E., Paish, C., Robertson, J., Macmillan, D., Blamey, R., Ellis, I., Ali, H., Dawson, S., Blows, F., Provenzano, E., Pharoah, P., Caldas, C., *et al.,* (2013a). A new genome-driven integrated classification of breast cancer and its implications. *The EMBO journal*, 32(5), 617–28.

Dawson, S., Rueda, O.M., Aparicio, S. and Caldas, C., (2013). A new genome-driven integrated classification of breast cancer and its implications. *The EMBO Journal*, 32(5), 617-28.

Desmedt, C., Yates, L. and Kulka, J., (2016). Catalog of genetic progression of human cancers: breast cancer. *Cancer and Metastasis Reviews*, 35(1), 49–62.

Dewangan, J., Srivastava, S. and Rath, S.K., (2017). Salinomycin: A new paradigm in cancer therapy. *Tumor Biology*, 39(3), 101042831769503.

van Dijk, E.L., Auger, H., Jaszczyszyn, Y. and Thermes, C., (2014). Ten years of next-generation sequencing technology. *Trends in Genetics*, 30(9), 418–426.

Ding, Y., Xu, L., Jovanovic, B.D., Helenowski, I.B., Kelly, D.L., Catalona, W.J., Yang, X.J., Pins, M. and Bergan, R.C., (2007). The methodology used to measure differential gene expression affects the outcome. *Journal of biomolecular techniques : JBT*, 18(5), 321–30.

Eroles, P., Bosch, A., Pérez-Fidalgo, J.A. and Lluch, A., (2012). Molecular biology in breast cancer: intrinsic subtypes and signaling pathways. *Cancer treatment reviews*, 38(6), 698–707.

Esteva, F.J., Yu, D., Hung, M.-C. and Hortobagyi, G.N., (2010). Molecular predictors of response to trastuzumab and lapatinib in breast cancer. *Nature Reviews Clinical Oncology*, 7(2), 98–107.

Fadoukhair, Z., Zardavas, D., Chad, M. a, Goulioti, T., Aftimos, P. and Piccart, M., (2015). Evaluation of targeted therapies in advanced breast cancer: the need for large-scale molecular screening and transformative clinical trial designs. *Oncogene*, 35(April), 1–7.

Fakruddin, M. and Chowdhury, A., (2012). Pyrosequencing-an alternative to traditional Sanger sequencing. *American Journal of Biochemistry and Biotechnology*, 8(1), 14–20.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E. and Mello, C.C., (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, 391(6669), 806–811.

van der Flier, L.G., van Gijn, M.E., Hatzis, P., Kujala, P., Haegebarth, A., Stange, D.E., Begthel, H., van den Born, M., Guryev, V., Oving, I., van Es, J.H., Barker, N., Peters, P.J., van de Wetering, M. and Clevers, H., (2009). Transcription Factor Achaete Scute-Like 2 Controls Intestinal Stem Cell Fate. *Cell*, 136(5), 903–912.

Freese, J.L., Pino, D. and Pleasure, S.J., (2010). Wnt signaling in development and disease. *Neurobiology of Disease*, 38(2), 148–153.

Friel, A.M., Corcoran, C., Crown, J. and O'Driscoll, L., (2010). Relevance of circulating tumor cells, extracellular nucleic acids, and exosomes in breast cancer. *Breast Cancer Research and Treatment*, 123(3), 613–625.

Frost, H.R. and Amos, C.I., (2018). A multi-omics approach for identifying important pathways and genes in human cancer. *BMC Bioinformatics*, 19(1), 479.

Fu, D., Calvo, J.A. and Samson, L.D., (2012). Balancing repair and tolerance of DNA damage caused by alkylating agents. *Nature Reviews Cancer*, 12(2), 104–120.

Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., Cerami, E., Sander, C. and Schultz, N., (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science signaling*, 6(269), pl1.

García-Bellido, A. and de Celis, J.F., (2009). The complex tale of the achaete-scute complex: a paradigmatic case in the analysis of gene organization and function during development. *Genetics*, 182(3), 631–9.

García-Campos, M. A., Espinal-Enríquez, J., & Hernández-Lemus, E., (2015). Pathway Analysis: State of the Art. *Frontiers in physiology,* 6, 383.

Garraway, L.A. and Baselga, J., (2012). Whole-Genome Sequencing and Cancer Therapy: Is Too Much Ever Enough? *Cancer Discovery*, 2(9), 766-8.

Gatza, M.L., Silva, G.O., Parker, J.S., Fan, C. and Perou, C.M., (2014). An integrated genomics approach identifies drivers of proliferation in luminal-subtype human breast cancer. *Nature Genetics*, 46(10), 1051–1059.

Gavrilov, K. and Saltzman, W.M., (2012). Therapeutic siRNA: principles, challenges, and strategies. *The Yale journal of biology and medicine*, 85(2), 187–200.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., *et al.,* (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10), R80.

Giakountis, A., Moulos, P., Zarkou, V., Oikonomou, C., Harokopos, V., Hatzigeorgiou, Artemis G, Reczko, M. and Hatzis, P., (2016). A Positive Regulatory Loop between a Wnt-Regulated Non-coding RNA and ASCL2 Controls Intestinal Stem Cell Fate. *Cell reports*, 15(12), 2588–96.

Gibney, E.R. and Nolan, C.M., (2010). Epigenetics and gene expression. *Heredity*, 105(1), 4–13.

Gillet, J.-P., Varma, S. and Gottesman, M.M., (2013). The clinical relevance of cancer cell lines. *Journal of the National Cancer Institute*, 105(7), 452–8.

Goel, S., Wang, Q., Watt, A.C., Tolaney, S.M., Dillon, D.A., Li, W., Ramm, S., Palmer, A.C., Yuzugullu, H., Varadan, V. & Tuck, D., (2016). Overcoming therapeutic resistance in HER2-positive breast cancers with CDK4/6 inhibitors. *Cancer cell*, 29(3), 255-269.

Gonzalez-Perez, A., (2016). Circuits of cancer drivers revealed by convergent misregulation of transcription factor targets across tumor types. *Genome medicine*, 8(1), 6.

Gonzalez-Perez, A. and Lopez-Bigas, N., (2012). Functional impact bias reveals cancer drivers. *Nucleic Acids Research*, 40(21), e169–e169.

Gonzalez, G.H., Tahsin, T., Goodale, B.C., Greene, A.C. and Greene, C.S., (2016). Recent Advances and Emerging Applications in Text and Data Mining for Biomedical Discovery. *Briefings in Bioinformatics*, 17(1), 33–42.

Gough, W., Hulkower, K.I., Lynch, R., Mcglynn, P., Uhlik, M., Yan, L. and Lee, J.A., (2011). A Quantitative, Facile, and High-Throughput Image-Based Cell Migration Method Is a Robust Alternative to the Scratch Assay. *Journal of Biomolecular Screening*, 16(2), 155–163.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G.L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., Edkins, S., O'Meara, S., Vastrik, I., Schmidt, E.E., Avis, T., Barthorpe, S., Bhamra, G., Buck, G., Choudhury, B., *et al.,* (2007). Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132), 153–8.

van der Groep, P., van der Wall, E. and van Diest, P.J., (2011). Pathology of hereditary breast cancer. *Cellular Oncology*, 34(2), 71–88.

Guillemot, F., Nagy, A., Auerbach, A., Rossant, J. and Joyner, A.L., (1994). Essential role of Mash-2 in extraembryonic development. *Nature*, 371(6495), 333–336.

Gupta, G.P. and Massagué, J., (2006). Cancer metastasis: building a framework. *Cell*, 127(4), 679–95.

Haimes, J. and Kelley, M., (2010). Demonstration of a ΔΔCq Calculation Method to Compute Thermo Scientific Relative Gene Expression from qPCR Data | SelectScience. *Technical Note, Dharmacon*.

Han, T., Kang, D., Ji, D., Wang, X., Zhan, W., Fu, M., Xin, H.-B. and Wang, J.-B., (2013). How does cancer cell metabolism affect tumor migration and invasion? *Cell Adhesion & Migration*, 7(5), 395–403.

Hanahan, D. and Weinberg, R.A., (2011). Hallmarks of Cancer: The Next Generation. *Cell*, 144, 646–674.

Hanahan, D. and Weinberg, R.A., (2000). The Hallmarks of Cancer. *Cell*, 100(1), 57–70.

Hansen, A.R. and Bedard, P.L., (2013). Clinical application of high-throughput genomic technologies for treatment selection in breast cancer. *Breast cancer research : BCR*, 15(5), R97.

Harbeck, N., Sotlar, K., Wuerstlein, R. and Doisneau-Sixou, S., (2014). Molecular and protein markers for clinical decision making in breast cancer: Today and tomorrow. *Cancer Treatment Reviews*, 40(3), 434–444.

Hoeflich, K.P., O'Brien, C., Boyd, Z., Cavet, G., Guerrero, S., Jung, K., Januario, T., Savage, H., Punnoose, E., Truong, T., Zhou, W., Berry, L., Murray, L., Amler, L., Belvin, M., Friedman, L.S. and Lackner, M.R., (2009). In vivo Antitumor Activity of MEK and Phosphatidylinositol 3-Kinase Inhibitors in Basal-Like Breast Cancer Models. *Clinical Cancer Research*, 15(14), 4649–4664.

Holliday, D.L. and Speirs, V., (2011). Choosing the right cell line for breast cancer research. *Breast Cancer Research*, 13(4), 215.

Hook, K.E., Garza, S.J., Lira, M.E., Ching, K.A., Lee, N. V., Cao, J., Yuan, J., Ye, J., Ozeck, M., Shi, S.T., Zheng, X., Rejto, P.A., Kan, J.L.C., Christensen, J.G. and Pavlicek, A., (2012). An Integrated Genomic Approach to Identify Predictive Biomarkers of Response to the Aurora Kinase Inhibitor PF-03814735. *Molecular Cancer Therapeutics*, 11(3), 710–719.

Horgan, R.P. and Kenny, L.C., (2011). 'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics. *The Obstetrician & Gynaecologist*, 13(3), 189–195.

Hou, X., Huang, F., Carboni, J.M., Flatten, K., Asmann, Y.W., Ten Eyck, C., Nakanishi, T., Tibodeau, J.D., Ross, D.D., Gottardis, M.M., Erlichman, C., Kaufmann, S.H. and Haluska, P., (2011). Drug Efflux by Breast Cancer Resistance Protein Is a Mechanism of Resistance to the Benzimidazole Insulin-Like Growth Factor Receptor/Insulin Receptor Inhibitor, BMS-536924. *Molecular Cancer Therapeutics*, 10(1), 117–125.

Howe, L.R. and Brown, A.M.C., (2004). Wnt Signaling and Breast Cancer. *Cancer Biology & Therapy*, 3(1), 36–41.

Hu, X., Chen, L., Wang, Q., Zhao, X., Tan, J., Cui, Y., Liu, X., Zhang, X. and Bian, X., (2015). Elevated expression of ASCL2 is an independent prognostic indicator in lung squamous cell carcinoma. *Journal of Clinical Pathology*, 69(4), 313-8.

Huang, D.W., Sherman, B.T., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y., Stephens, R., Baseler, M.W., Lane, H.C. and Lempicki, R.A., (2007). DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Research*, 35(Web Server), W169–W175.

Ille, F. and Sommer, L., (2005). Wnt signaling: multiple functions in neural development. *CMLS Cellular and Molecular Life Sciences*, 62(10), 1100–1108.

International Human Genome Sequencing Consortium, (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011), 931–45.

Iorio, M. V., Ferracin, M., Liu, C.-G., Veronese, A., Spizzo, R., Sabbioni, S., Magri, E., Pedriali, M., Fabbri, M., Campiglio, M., Ménard, S., Palazzo, J.P., Rosenberg, A., Musiani, P., Volinia, S., Nenci, I., Calin, G.A., Querzoli, P., Negrini, M., *et al.,* (2005). MicroRNA Gene Expression Deregulation in Human Breast Cancer. *Cancer Research*, 65(16), 7065–

7070.

van Iterson, M., ’t Hoen, P., Pedotti, P., Hooiveld, G., den Dunnen, J., van Ommen, G., Boer, J. and Menezes, R., (2009). Relative power and sample size analysis on gene expression profiling data. *BMC Genomics*, 10(1), 439.

Jang, G.-B., Kim, J.-Y., Cho, S.-D., Park, K.-S., Jung, J.-Y., Lee, H.-Y., Hong, I.-S. and Nam, J.-S., (2015). Blockade of Wnt/β-catenin signaling suppresses breast cancer metastasis by inhibiting CSC-like phenotype. *Scientific Reports*, 5(1), 12465.

Jänicke, R.U., Sprengart, M.L., Wati, M.R. and Porter, A.G., (1998). Caspase-3 is required for DNA fragmentation and morphological changes associated with apoptosis. *The Journal of biological chemistry*, 273(16), 9357–60.

Jänicke, R.U., (2008). MCF-7 breast carcinoma cells do not express caspase-3. *Breast cancer research and treatment, 117(1), 219-221.*

Johnston, S.T., Simpson, M.J. and McElwain, D.L.S., (2014). How much information can be obtained from tracking the position of the leading edge in a scratch assay? *Journal of The Royal Society Interface*, 11(97), 20140325–20140325.

Jonkman, J.E.N., Cathcart, J.A., Xu, F., Bartolini, M.E., Amon, J.E., Stevens, K.M. and Colarusso, P., (2014). An introduction to the wound healing assay using live-cell microscopy. *Cell Adhesion & Migration*, 8(5), 440–451.

Jordan, N.V., Bardia, A., Wittner, B.S., Benes, C., Ligorio, M., Zheng, Y., Yu, M., Sundaresan, T.K., Licausi, J.A., Desai, R., O’Keefe, R.M., Ebright, R.Y., Boukhali, M., Sil, S., Onozato, M.L., Iafrate, A.J., Kapur, R., Sgroi, D., Ting, D.T., *et al.,* (2016). HER2 expression identifies dynamic functional states within circulating breast cancer cells. *Nature*, 537(7618), 102–106.

Juarez, M., Schcolnik-Cabrera, A. and Dueñas-Gonzalez, A., (2018). The multitargeted drug ivermectin: from an antiparasitic agent to a repositioned cancer drug. *American journal of cancer research*, 8(2), 317–331.

Jubb, A.M., Chalasani, S., Frantz, G.D., Smits, R., Grabsch, H.I., Kavi, V., Maughan, N.J., Hillan, K.J., Quirke, P. and Koeppen, H., (2006). Achaete-scute like 2 (ascl2) is a target of Wnt signalling and is upregulated in intestinal neoplasia. *Oncogene*, 25(24), 3445–57.

Kalia, M., (2015). Biomarkers for personalized oncology: recent advances and future challenges. *Metabolism*, 64(3), S16–S21.

Kamburov, A., Cavill, R., Ebbels, T.M.D., Herwig, R. and Keun, H.C., (2011). Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics (Oxford, England)*, 27(20), 2917–8.

Karagoz, K., Sinha, R. and Arga, K.Y., (2015). Triple negative breast cancer: a multi-omics network discovery strategy for candidate targets and driving pathways. *Omics : a journal of integrative biology*, 19(2), 115–30.

Katt, M.E., Placone, A.L., Wong, A.D., Xu, Z.S. and Searson, P.C., (2016). In Vitro Tumor Models: Advantages, Disadvantages, Variables, and Selecting the Right Platform. *Frontiers in Bioengineering and Biotechnology*, 4, 12.

Kazi, Trivedi, T.I., Kobawala, T.P. and Ghosh, N.R., (2016). The Potential of Wnt Signaling Pathway in Cancer: A Focus on Breast Cancer. *Cancer*

Translational Medicine, 2(2), 55.

Kelsey, J.L., Gammon, M.D. and John, E.M., (1993). Reproductive factors and breast cancer. *Epidemiol.Rev.*, 15(1), 36–47.

Khatri, P., Sirota, M. and Butte, A.J., (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*, 8(2), e1002375.

Kim, H. J., (2019). Cell Fate Control by Translation: mRNA Translation Initiation as a Therapeutic Target for Cancer Development and Stem Cell Fate Control. *Biomolecules*, 9(11), 665.

Kim, J.-H., Park, S.-Y., Jun, Y., Kim, J.-Y. and Nam, J.-S., (2017). Roles of Wnt Target Genes in the Journey of Cancer Stem Cells. *International Journal of Molecular Sciences*, 18(8), 1604.

King, T.D., Suto, M.J. and Li, Y., (2012). The Wnt/β-catenin signaling pathway: a potential therapeutic target in the treatment of triple negative breast cancer. *Journal of cellular biochemistry*, 113(1), 13–8.

Korthauer, K.D. and Kendziorski, C., (2015). MADGiC: a model-based approach for identifying driver genes in cancer. *Bioinformatics (Oxford, England)*, 31(10), 1526–35.

Kuol, N., Stojanovska, L., Apostolopoulos, V. and Nurgali, K., (2018). Role of the nervous system in cancer metastasis. *Journal of experimental & clinical cancer research : CR*, 37(1), 5.

Küry, P., Greiner-Petter, R., Cornely, C., Jürgens, T. and Müller, H.W., (2002). Mammalian achaete scute homolog 2 is expressed in the adult sciatic nerve and regulates the expression of Krox24, Mob-1, CXCR4, and p57kip2 in Schwann cells. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 22(17), 7586–95.

Kwon, O.-H., Park, J.-L., Baek, S.-J., Noh, S.-M., Song, K.-S., Kim, S.-Y. and Kim, Y.S., (2013). Aberrant upregulation of *ASCL2* by promoter demethylation promotes the growth and resistance to 5-fluorouracil of gastric cancer cells. *Cancer Science*, 104(3), 391–397.

Lacroix, M. and Leclercq, G., (2004). Relevance of breast cancer cell lines as models for breast tumours: an update. *Breast cancer research and treatment*, 83(3), 249–89.

Lamb, R., Ablett, M.P., Spence, K., Landberg, G., Sims, A.H. and Clarke, R.B., (2013). Wnt Pathway Activity in Breast Cancer Sub-Types and Stem-Like Cells. *PLoS ONE*, 8(7), e67811.

Lambert, M., Jambon, S., Depauw, S. and David-Cordonnier, M.-H., (2018). Targeting Transcription Factors for Cancer Treatment. *Molecules (Basel, Switzerland)*, 23(6).

Lash, T.L. and Aschengrau, A., (2002). A null association between active or passive cigarette smoking and breast cancer risk. *Breast cancer research and treatment,* 75(2), 181-184.

Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R., Meyerson, M., Gabriel, S.B., Lander, E.S. and Getz, G., (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484), 495–501.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., Kiezun, A., Hammerman, P.S., McKenna, A., Drier, Y., Zou, L., Ramos, A.H., Pugh, T.J., Stransky, N., Helman, E., *et al.,* (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457), 214–218.

Lee, E.Y.H.P. and Muller, W.J., (2010). Oncogenes and tumor suppressor genes. *Cold Spring Harbor perspectives in biology*, 2(10), a003236.

Lehmann, B.D., Bauer, J.A., Chen, X., Sanders, M.E., Chakravarthy, A.B., Shyr, Y. and Pietenpol, J.A., (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *The Journal of clinical investigation*, 121(7), 2750–67.

Li, B., Shin, H., Gulbekyan, G., Pustovalova, O., Nikolsky, Y., Hope, A., Bessarabova, M., Schu, M., Kolpakova-Hart, E., Merberg, D., Dorner, A. and Trepicchio, W.L., (2015). Development of a Drug-Response Modeling Framework to Identify Cell Line Derived Translational Biomarkers That Can Predict Treatment Outcome to Erlotinib or Sorafenib. *PLOS ONE*, 10(6), e0130700.

Li, Y., Guo, M., Fu, Z., Wang, P., Zhang, Y., Gao, Y., Yue, M., Ning, S. and Li, D., (2016). Immunoglobulin superfamily genes are novel prognostic biomarkers for breast cancer. *Oncotarget*, 8(2), 2444-2456.

Liang, Q., Li, W., Zhao, Z. and Fu, Q., (2016). Advancement of Wnt signal pathway and the target of breast cancer. *Open Life Sciences*, 11(1), 98–104.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L. and Law, M., (2012). Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology*, 2012, 1–11.

Liu, M.-H., Cui, Y.-H., Guo, Q.-N. and Zhou, Y., (2016). Elevated ASCL2 expression is associated with metastasis of osteosarcoma and predicts poor prognosis of the patients. *American journal of cancer research*, 6(6), 1431–40.

Liu, S., Wang, H., Zhang, L., Tang, C., Jones, L., Ye, H., Ban, L., Wang, A., Liu, Z., Lou, F., Zhang, D., Sun, H., Dong, H., Zhang, G., Dong, Z., Guo, B., Yan, H., Yan, C., Wang, L., *et al.,* (2015). Rapid detection of genetic mutations in individual breast cancer patients by next-generation DNA sequencing. *Human genomics*, 9, 2.

Liu, X., Chen, X., Zhong, B., Wang, A., Wang, X., Chu, F., Nurieva, R.I., Yan, X., Chen, P., Van Der Flier, L.G., Nakatsukasa, H., Neelapu, S.S., Chen, W., Clevers, H., Tian, Q., Qi, H., Wei, L. and Dong, C., (2014). Transcription factor achaete-scute homologue 2 initiates follicular T-helper-cell development. *Nature*, 507.

Liu, Y. and Hu, Z., (2014). Identification of collaborative driver pathways in breast cancer. *BMC genomics*, 15(1), 605.

Lord, C.J. and Ashworth, A., (2012). The DNA damage response and cancer therapy. *Nature*, 481(7381), 287–94.

Luo, J., Margolis, K.L., Wactawski-Wende, J., Horn, K., Messina, C., Stefanick, M.L., Tindle, H.A., Tong, E. and Rohan, T.E., (2011). Association of active and passive smoking with risk of breast cancer among postmenopausal women: a prospective cohort study. *BMJ*, 342, 1016.

Luo, D., Wilson, J.M., Harvel, N., Liu, J., Pei, L., Huang, S., Hawthorn, L. and Shi, H., (2013). A systematic evaluation of miRNA:mRNA interactions involved in the migration and invasion of breast cancer cells. *Journal of Translational Medicine*, 11(1), 57.

Macgregor, P.F. and Squire, J.A., (2002). Application of microarrays to the analysis of gene expression in cancer. *Clinical chemistry*, 48(8), 1170–7.

Maier, T., Güell, M., & Serrano, L. (2009). Correlation of mRNA and protein in complex biological samples. *FEBS letters*, 583(24), 3966-3973.

Malone, J.H. and Oliver, B., (2011). Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biology*, 9(1), 34.

Mathur, R., Rotroff, D., Ma, J., Shojaie, A. and Motsinger-Reif, A., (2018). Gene set analysis methods: a systematic comparison. *BioData Mining*, 11(1), 8.

Mavaddat, N., Michailidou, K., Dennis, J., Lush, M., Fachal, L., Lee, A., Tyrer, J.P., Chen, T.H., Wang, Q., Bolla, M.K. and Yang, X., (2019). Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *The American Journal of Human Genetics,* 104(1), 21-34.

Mcveigh, T. and George, A., (2017). Personalisation of Therapy – clinical impact and relevance of genetic mutations in tumours. *Cancer Research Frontiers*, 329(10), 29–50.

Mi, H., Muruganujan, A., Ebert, D., Huang, X. and Thomas, P.D., (2019). PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Research*, 47(D1), D419–D426.

Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R.L., Schmidt, M.K., Chang-Claude, J., Bojesen, S.E., Bolla, M.K., Wang, Q., Dicks, E., Lee, A., Turnbull, C., Rahman, N., Fletcher, O., Peto, J., Gibson, L., Dos Santos Silva, I., *et al.,* (2013). Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nature genetics*, 45(4), 353–61.

Michaut, M., Chin, S.-F., Majewski, I., Severson, T.M., Bismeijer, T., de Koning, L., Peeters, J.K., Schouten, P.C., Rueda, O.M., Bosma, A.J., Tarrant, F., Fan, Y., He, B., Xue, Z., Mittempergher, L., Kluin, R.J.C., Heijmans, J., Snel, M., Pereira, B., *et al.,* (2016). Integration of genomic, transcriptomic and proteomic data identifies two biologically distinct subtypes of invasive lobular breast cancer. *Scientific reports*, 6, 18517.

Mina, L.A. and Arun, B., (2019) Polygenic Risk Scores in Breast Cancer. Current Breast Cancer Reports, 11, 117–122.

Mo, Q., Wang, S., Seshan, V.E., Olshen, A.B., Schultz, N., Sander, C., Powers, R.S., Ladanyi, M. and Shen, R., (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*, 110(11), 4245–4250.

Mootha, V.K., Lindgren, C.M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Houstis, N., Daly, M.J., Patterson, N., Mesirov, J.P., Golub, T.R., Tamayo, P., Spiegelman, B., Lander, E.S., Hirschhorn, J.N., *et al.,* (2003). PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34(3), 267–273.

Mørch, L.S., Skovlund, C.W., Hannaford, P.C., Iversen, L., Fielding, S. and Lidegaard, Ø., (2017). Contemporary Hormonal Contraception and the Risk of Breast Cancer. *New England Journal of Medicine*, 377(23), 2228–2239.

Morey, J.S., Ryan, J.C. and Van Dolah, F.M., (2006). Microarray validation: factors influencing correlation between oligonucleotide microarrays and real-time PCR. *Biological procedures online*, 8, 175–93.

Moriyama, M., Durham, A.-D., Moriyama, H., Hasegawa, K., Nishikawa, S.-I., Radtke, F. and Osawa, M., (2008). Multiple Roles of Notch Signaling in the Regulation of Epidermal Development. *Developmental Cell*, 14(4), 594–604.

National Cancer Institute, (2018). Female Breast Cancer - Cancer Stat Facts. *SEER Cancer Stat Facts: Female Breast Cancer*. Available from: https://seer.cancer.gov/statfacts/html/breast.html.

National Cancer Institute, (2011). FDA Approval for Lapatinib Ditosylate. Available from: https://www.cancer.gov/about-cancer/treatment/drugs/fda-lapatinib.

Ng, C.K.Y., Schultheis, A.M., Bidard, F.-C., Weigelt, B. and Reis-Filho, J.S., (2015). Breast cancer genomics from microarrays to massively parallel sequencing: paradigms and new insights. *Journal of the National Cancer Institute*, 107(5), djv015.

Ng, C.K.Y., Martelotto, L.G., Gauthier, A., Wen, H.-C., Piscuoglio, S., Lim, R.S., Cowell, C.F., Wilkerson, P.M., Wai, P., Rodrigues, D.N., Arnould, L., Geyer, F.C., Bromberg, S.E., Lacroix-Triki, M., Penault-Llorca, F., Giard, S., Sastre-Garau, X., Natrajan, R., Norton, L., *et al.,* (2015). Intra-tumor genetic heterogeneity and alternative driver genetic alterations in breast cancers with heterogeneous HER2 gene amplification. *Genome biology*, 16, 107.

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L.B., Martin, S., Wedge, D.C., Van Loo, P., Ju, Y.S., Smid, M., Brinkman, A.B., Morganella, S., Aure, M.R., Lingjærde, O.C., Langerød, A., Ringnér, M., *et al.,* (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534, 47-54.

Ogden, A., Rida, P.C.G. and Aneja, R., (2017). Prognostic value of CA20, a score based on centrosome amplification-associated genes, in breast tumors. *Scientific Reports*, 7(1), 262.

Oh-McGinnis, R., Bogutz, A.B. and Lefebvre, L., (2011). Partial loss of Ascl2 function affects all three layers of the mature placenta and causes intrauterine growth restriction. *Developmental Biology*, 351(2), 277–286.

Omictools, (2017). Omictools. Available from: https://omictools.com/.

Ortiz, A. G., Muñoz, A. S., Parrado, M. R. C., Pérez, M. Á., Entrena, N. R., Dominguez, A. R., & Conejo, E. A. (2019). Deciphering HER2 breast cancer disease: biological and clinical implications. *Frontiers in oncology*, 9, 1124.

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M.R., Zschocke, J. and Trajanoski, Z., (2014). A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in bioinformatics*, 15(2), 256–78.

Park, S. and Lehner, B., (2015). Cancer type-dependent genetic interactions between cancer driver alterations indicate plasticity of epistasis across cell types. *Molecular Systems Biology*, 11(7), 824.

Parker, J.S., Mullins, M., Cheang, M.C.U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenbush, J.F., Stijleman, I.J., Palazzo, J., Marron, J.S., Nobel, A.B., Mardis, E., Nielsen, T.O., Ellis, M.J., Perou, C.M., *et al.,* (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 27(8), 1160–7.

Passarelli, M.N., Newcomb, P.A., Hampton, J.M., Trentham-Dietz, A., Titus, L.J., Egan, K.M., Baron, J.A. and Willett, W.C., (2016). Cigarette Smoking Before and After Breast Cancer Diagnosis: Mortality From Breast Cancer and Smoking-Related Diseases. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 34(12), 1315–22.

Pavlicek, A., Lira, M.E., Lee, N. V., Ching, K.A., Ye, J., Cao, J., Garza, S.J., Hook, K.E., Ozeck, M., Shi, S.T., Yuan, J., Zheng, X., Rejto, P.A., Kan, J.L.C. and Christensen, J.G., (2013). Molecular Predictors of Sensitivity to the Insulin-like Growth Factor 1 Receptor Inhibitor Figitumumab (CP-751,871). *Molecular Cancer Therapeutics*, 12(12), 2929–2939.

Pereira, B., Chin, S.-F., Rueda, O.M., Vollan, H.-K.M., Provenzano, E., Bardwell, H.A., Pugh, M., Jones, L., Russell, R., Sammut, S.-J., Tsui, D.W.Y., Liu, B., Dawson, S.-J., Abraham, J., Northen, H., Peden, J.F., Mukherjee, A., Turashvili, G., Green, A.R., *et al.,* (2016). The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nature Communications*, 7, 11479.

Perou, C.M. and Børresen-Dale, A.-L., (2011). Systems biology and genomics of breast cancer. *Cold Spring Harbor perspectives in biology*, 3(2), a003293.

Piatetsky-Shapiro, G. and Tamayo, P., (2003). Microarray data mining. *ACM SIGKDD Explorations Newsletter*, 5(2), 1.

Piccolo, S.R. and Frampton, M.B., (2016). Tools and techniques for computational reproducibility. *GigaScience*, 5(1), 30.

Pon, J.R. and Marra, M.A., (2015). Driver and passenger mutations in cancer. *Annual review of pathology*, 10, 25–50.

Pontén, F., Jirström, K. and Uhlen, M., (2008). The Human Protein Atlas-a tool for pathology. *The Journal of Pathology*, 216(4), 387–393.

Prat, A., Carey, L.A., Adamo, B., Vidal, M., Tabernero, J., Cortés, J., Parker, J.S., Perou, C.M. and Baselga, J., (2014). Molecular Features and Survival Outcomes of the Intrinsic Subtypes Within HER2-Positive Breast Cancer. *JNCI: Journal of the National Cancer Institute*, 106(8).

Prat, A. and Perou, C.M., (2011). Deconstructing the molecular portraits of breast cancer. *Molecular oncology*, 5(1), 5–23.

Prescott, J., Ma, H., Bernstein, L. and Ursin, G., (2007). Cigarette smoking is not associated with breast cancer risk in young women. *Cancer Epidemiology and Prevention Biomarkers*, 16(3), 620-622.

Puck, T.T. and Marcus, P.I., (1955). A rapid method for viable cell titration and clone production with hela cells in tissue culture: the use of x-irradiated cells to supply conditioning factors. *Proceedings of the National Academy of Sciences of the United States of America*, 41(7), 432–7.

Radonić, A., Thulke, S., Mackay, I.M., Landt, O., Siegert, W. and Nitsche, A., (2004). Guideline to reference gene selection for quantitative real-time PCR. *Biochemical and Biophysical Research Communications*, 313(4), 856–862.

Ramasamy, A., Mondry, A., Holmes, C.C. and Altman, D.G., (2008). Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS medicine*, 5(9), e184.

Rampersad, S.N., (2012). Multiple applications of Alamar Blue as an indicator of metabolic function and cellular health in cell viability bioassays. *Sensors (Basel, Switzerland)*, 12(9), 12347–60.

Reaz, S., Tamkus, D. and Andrechek, E.R., (2018). Using gene expression data to direct breast cancer therapy: evidence from a preclinical trial. *Journal of Molecular Medicine*, 1–7.

Reis-Filho, J.S. and Pusztai, L., (2011). Gene expression profiling in breast cancer: classification, prognostication, and prediction. *Lancet (London, England)*, 378(9805), 1812–23.

Rexer, B.N. and Arteaga, C.L., (2013). Optimal targeting of HER2-PI3K signaling in breast cancer: mechanistic insights and clinical implications. *Cancer research*, 73(13), 3817–20.

Rizzolo, P., Silvestri, V., Falchetti, M. and Ottini, L., (2011). Inherited and acquired alterations in development of breast cancer. *The application of clinical genetics*, 4, 145–58.

Rodriguez-Barrueco, R., Nekritz, E.A., Bertucci, F., Yu, J., Sanchez-Garcia, F., Zeleke, T.Z., Gorbatenko, A., Birnbaum, D., Ezhkova, E., Cordon-Cardo, C., Finetti, P., Llobet-Navas, D. and Silva, J.M., (2017). miR-424(322)/503 is a breast cancer tumor suppressor whose loss promotes resistance to chemotherapy. *Genes & development*, 31(6), 553–566.

Ross, D.T. and Perou, C.M., (2001). A comparison of gene expression signatures from breast tumors and breast tissue derived cell lines. *Disease markers*, 17(2), 99–109.

Roy, P.G. and Thompson, A.M., (2006). Cyclin D1 and breast cancer. *The Breast*, 15(6), 718–727.

Rubio-Perez, C., Tamborero, D., Schroeder, M.P., Antolín, A.A., Deu-Pons, J., Perez-Llamas, C., Mestres, J., Gonzalez-Perez, A. and Lopez-Bigas, N., (2015). In Silico Prescription of Anticancer Drugs to Cohorts of 28 Tumor Types Reveals Targeting Opportunities. *Cancer Cell*, 27(3), 382–396.

Russnes, Hege G, Lingjærde, O.C., Børresen-Dale, A.-L. and Caldas, C., (2017). Breast Cancer Molecular Stratification: From Intrinsic Subtypes to Integrative Clusters. *The American journal of pathology*, 187(10), 2152–

2162.

Salic, A. and Mitchison, T.J., (2008). A chemical method for fast and sensitive detection of DNA synthesis in vivo. *Proceedings of the National Academy of Sciences*, 105(7), 2415–2420.

Sanger, F., Nicklen, S. and Coulson, A.R., (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), 5463–7.

Schuijers, J., Junker, J.P., Mokry, M., Hatzis, P., Koo, B.-K., Sasselli, V., van der Flier, L.G., Cuppen, E., van Oudenaarden, Alexander, Clevers, H., Acar, M., Mettetal, J.T., Oudenaarden, A. van, Andreu, P., Colnot, S., Godard, C., Gad, S., Chafey, P., Niwa-Kawakita, M., *et al.,* (2015). Ascl2 acts as an R-spondin/Wnt-responsive switch to control stemness in intestinal crypts. *Cell stem cell*, 16(2), 158–70.

Sever, R. and Brugge, J.S., (2015). Signal transduction in cancer. *Cold Spring Harbor perspectives in medicine*, 5(4).

Shiovitz, S. and Korde, L.A., (2015). Genetics of breast cancer: a topic in evolution. *Annals of Oncology*, 26(7), 1291–1299.

Simionato, E., Kerner, P., Dray, N., Le Gouar, M., Ledent, V., Arendt, D. and Vervoort, M., (2008). atonal- and achaete-scute-related genes in the annelid Platynereis dumerilii: insights into the evolution of neural basic-Helix-Loop-Helix genes. *BMC Evolutionary Biology*, 8(1), 170.

Slodkowska, E.A. and Ross, J.S., (2009). MammaPrint^{TM} 70-gene signature: another milestone in personalized medical care for breast cancer patients. *Expert Review of Molecular Diagnostics*, 9(5), 417–422.

Slonim, D.K. and Yanai, I., (2009). Getting Started in Gene Expression Microarray Analysis O. G. Troyanskaya, ed. *PLoS Computational Biology*, 5(10), e1000543.

Smid, M., Rodríguez-González, F.G., Sieuwerts, A.M., Salgado, R., Prager-Van der Smissen, W.J.C., Vlugt-Daane, M. van der, van Galen, A., Nik-Zainal, S., Staaf, J., Brinkman, A.B., van de Vijver, M.J., Richardson, A.L., Fatima, A., Berentsen, K., Butler, A., Martin, S., Davies, H.R., Debets, R., Gelder, M.E.M.-V., *et al.,* (2016). Breast cancer genome and transcriptome integration implicates specific mutational signatures with immune cell infiltration. *Nature Communications*, 7, 12910.

Smith, R.J., Rao-Bhatia, A. and Kim, T.-H., (2017). Signaling and epigenetic mechanisms of intestinal stem cells and progenitors: insight into crypt homeostasis, plasticity, and niches. *Wiley Interdisciplinary Reviews: Developmental Biology*, e281.

de Sousa, E.M.F., Vermeulen, L., Richel, D. and Medema, J.P., (2011). Targeting Wnt Signaling in Colon Cancer Stem Cells. *Clinical Cancer Research*, 17(4), 647–653.

Stange, D.E., Engel, F., Longerich, T., Koo, B.K., Koch, M., Delhomme, N., Aigner, M., Toedt, G., Schirmacher, P., Lichter, P., Weitz, J. and Radlwimmer, B., (2010). Expression of an ASCL2 related stem cell signature and IGF2 in colorectal cancer liver metastases with 11p15.5 gain. *Gut*, 59(9), 1236–44.

Stinson, S., Lackner, M.R., Adai, A.T., Yu, N., Kim, H.-J., O'Brien, C., Spoerke, J., Jhunjhunwala, S., Boyd, Z., Januario, T., Newman, R.J., Yue, P., Bourgon, R., Modrusan, Z., Stern, H.M., Warming, S., de Sauvage, F.J., Amler, L., Yeh, R.-F., et al., (2011). TRPS1 Targeting by miR-221/222 Promotes the Epithelial-to-Mesenchymal Transition in Breast Cancer. *Science Signaling*, 4(177), ra41–ra41.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. and Mesirov, J.P., (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), 15545–50.

Sun, X. and Yu, Q., (2015). Intra-tumor heterogeneity of cancer cells and its implications for cancer treatment. *Acta pharmacologica Sinica*, 36(10), 1219–27.

Sundquist, T., Moravec, R., Niles, A., O'brien, M. and Riss, T., (2006). *Timing Your Apoptosis Assays*, Available from: https://promega.media/-/media/files/resources/cell-notes/cn016/timing-your-apoptosis-assays.pdf?la=en.

Suo, C., Hrydziuszko, O., Lee, D., Pramana, S., Saputra, D., Joshi, H., Calza, S. and Pawitan, Y., (2015). Integration of somatic mutation, expression and functional data reveals potential driver genes predictive of breast cancer survival. *Bioinformatics (Oxford, England)*, 31(16), 2607–13.

Tanaka, T., Kojima, K., Yokota, K., Tanaka, Y., Ooizumi, Y., Ishii, S., Nishizawa, N., Yokoi, K., Ushiku, H., Kikuchi, M., Kojo, K., Minatani, N., Katoh, H., Sato, T., Nakamura, T., Sawanobori, M., Watanabe, M. and Yamashita, K., (2019). Comprehensive Genetic Search to Clarify the Molecular Mechanism of Drug Resistance Identifies ASCL2-LEF1/TSPAN8 Axis in Colorectal Cancer. *Annals of Surgical Oncology*, 1–11.

Terry, M. B., Zhang, F. F., Kabat, G., Britton, J. A., Teitelbaum, S. L., Neugut, A. I., and Gammon, M. D., (2006). Lifetime alcohol intake and breast cancer risk. *Annals of epidemiology,* 16(3), 230-240.

The Cancer Genome Atlas Network, (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418), 61–70.

Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A. and Narechania, A., (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome research*, 13(9), 2129–41.

Tian, Y., Pan, Q., Shang, Y., Zhu, R., Ye, J., Liu, Y., Zhong, X., Li, S., He, Y., Chen, L., Zhao, J., Chen, W., Peng, Z. and Wang, R., (2014). MicroRNA-200 (miR-200) Cluster Regulation by Achaete Scute-like 2 (Ascl2). *Journal of Biological Chemistry*, 289(52), 36101–36115.

Tokheim, C., Papadopoulis, N., Kinzler, K.W., Vogelstein, B. and Karchin, R., (2016). Evaluating the Evaluation of Cancer Driver Genes, *Cold Spring Harbor Labs Journals,* 113(50), 14330-14335.

Tomasetti, C., Marchionni, L., Nowak, M.A., Parmigiani, G. and Vogelstein, B., (2015). Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proceedings of the National Academy of Sciences of the United States of America*, 112(1), 118–23.

Turnbull, C., Scott, R.H., Thomas, E., Jones, L., Murugaesu, N., Pretty, F.B., Halai, D., Baple, E., Craig, C., Hamblin, A. and Henderson, S., 2018. The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ*, 361, 1687.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M. and Rozen, S.G., (2012). Primer3--new capabilities and interfaces. *Nucleic acids research*, 40(15), e115.

Uzilov, A. V., Ding, W., Fink, M.Y., Antipin, Y., Brohl, A.S., Davis, C., Lau, C.Y., Pandya, C., Shah, H., Kasai, Y., Powell, J., Micchelli, M., Castellanos, R., Zhang, Z., Linderman, M., Kinoshita, Y., Zweig, M., Raustad, K., Cheung, K., *et al.,* (2016). Development and clinical application of an integrative genomic approach to personalized cancer therapy. *Genome Medicine*, 8(1), 62.

Vaklavas, C., Blume, S. W., & Grizzle, W. E., (2020). Hallmarks and Determinants of Oncogenic Translation Revealed by Ribosome Profiling in Models of Breast Cancer. *Translational Oncology,* 13(2), 452-470.

Vernieri, C., Milano, M., Brambilla, M., Mennitto, A., Maggi, C., Cona, M.S., Prisciandaro, M., Fabbroni, C., Celio, L., Mariani, G. & Bianchi, G.V., (2019). Resistance mechanisms to anti-HER2 therapies in HER2-positive breast cancer: current knowledge, new research directions and therapeutic perspectives. *Critical reviews in oncology/hematology.*

Vogel, C., & Marcotte, E. M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature reviews genetics*, 13(4), 227-232.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A. and Kinzler, K.W., (2013). Cancer genome landscapes. *Science*, 339(6127), 1546–58.

Vu, T. and Claret, F.X., (2012). Trastuzumab: updated mechanisms of action and resistance in breast cancer. *Frontiers in oncology*, 2, 62.

Wang, C., Wang, M., Arrington, J., Shan, T., Yue, F., Nie, Y., Tao, W.A. and Kuang, S., (2017). Ascl2 inhibits myogenesis by antagonizing the transcriptional activity of myogenic regulatory factors. *Development*, 144(2), 235-247.

Wang, C.Y., Shahi, P., Ting, J., Huang, W.E.I. and Phan, N.A.M.N., (2017). Systematic analysis of the achaete‑scute complex-like gene signature in clinical cancer patients. *Mol Clin Oncol.* 6(1), 7–18.

Wang, J., Cai, H., Xia, Y., Wang, S., Xing, L., Chen, C., Zhang, Y., Xu, J., Yin, P., Jiang, Y., Zhao, R., Zuo, Q. and Chen, T., (2018). Bufalin inhibits gastric cancer invasion and metastasis by down-regulating Wnt/ASCL2 expression. *Oncotarget*, 9(34), 23320–23333.

Wang, R., Li, J., Yin, C., Zhao, D. and Yin, L., (2018). Identification of differentially expressed genes and typical fusion genes associated with three subtypes of breast cancer. *Breast Cancer*, 1–12.

Wei, X., Ye, J., Shang, Y., Chen, H., Liu, S., Liu, L. and Wang, R., (2017). Ascl2 activation by YAP1/KLF5 ensures the self-renewability of colon cancer progenitor cells. *Oncotarget*, 8(65), 109301–109318.

Werner, H.M.J., Mills, G.B. and Ram, P.T., (2014). Cancer Systems Biology: a peek into the future of patient care? *Nature reviews. Clinical oncology*, 11(3), 167–76.

Wong, W.C., Kim, D., Carter, H., Diekhans, M., Ryan, M.C. and Karchin, R., (2011). CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. *Bioinformatics*, 27(15), 2147–2148.

Xu, H., Zhao, X.-L., Liu, X., Hu, X.-G., Fu, W.-J., Li, Q., Wang, Y., Ping, Y.-F., Zhang, X., Bian, X.-W. and Yao, X.-H., (2017). Elevated ASCL2 expression in breast cancer is associated with the poor prognosis of patients. *American journal of cancer research*, 7(4), 955–961.

Xue, Z., Vis, D.J., Bruna, A., Sustic, T., van Wageningen, S., Batra, A.S., Rueda, O.M., Bosdriesz, E., Caldas, C., Wessels, L.F.A. and Bernards, R., (2018). MAP3K1 and MAP2K4 mutations are associated with sensitivity to MEK inhibitors in multiple cancer models. *Cell Research*, 28(7), 719–729.

Yan, K., Kuo, C., Li, X., Ootani, A., Su, J., Lee, J.Y., Su, N., Luo, Y., Heilshorn, S.C., Amieva, M.R. and al., et, (2015). Ascl2 Reinforces Intestinal Stem Cell Identity. *Cell Stem Cell*, 16(2), 105–106.

Yang, C., Qiu, L. and Xu, Z., (2011). Specific gene silencing using RNAi in cell culture. *Methods in molecular biology*, 793, 457–77.

Yang, L., Tang, H., Kong, Y., Xie, Xinhua, Chen, J., Song, C., Liu, X., Ye, F., Li, N., Wang, N. and Xie, Xiaoming, (2015a). LGR5 Promotes Breast Cancer Progression and Maintains Stem-Like Cells Through Activation of Wnt/β-Catenin Signaling. *Stem cells*, 33(10), 2913–2924.

Yang, S.X., Polley, E. and Lipkowitz, S., (2016). New insights on PI3K/AKT pathway alterations and clinical outcomes in breast cancer. *Cancer Treatment Reviews*, 45, 87–96.

Yang, Y., Dong, X., Xie, B., Ding, N., Chen, J., Li, Y., Zhang, Q., Qu, H. and Fang, X., (2015b). Databases and web tools for cancer genomics study. *Genomics, proteomics & bioinformatics*, 13(1), 46–50.

Yee, S.S., Lieberman, D.B., Blanchard, T., Rader, J., Zhao, J., Troxel, A.B., DeSloover, D., Fox, A.J., Daber, R.D., Kakrecha, B., Sukhadia, S., Belka, G.K., DeMichele, A.M., Chodosh, L.A., Morrissette, J.J.D. and Carpenter, E.L., (2016). A novel approach for next-generation sequencing of circulating tumor cells. *Molecular Genetics & Genomic Medicine,* 4(4), 395-406.

Yeo, S.K. and Guan, J.-L., (2017). Breast Cancer: Multiple Subtypes within a Tumor? *Trends in Cancer*, 3(11), 753-760.

Yerushalmi, R., Woods, R., Ravdin, P.M., Hayes, M.M. and Gelmon, K.A., (2010). Ki67 in breast cancer: prognostic and predictive potential. *The Lancet Oncology*, 11(2), 174–183.

Yin, S., Cheryan, V.T., Xu, L., Rishi, A.K. and Reddy, K.B., (2017). Myc mediates cancer stem-like cells and EMT changes in triple negative breast cancers cells. *PloS one*, 12(8), e0183578.

Yue, P.Y.K., Leung, E.P.Y., Mak, N.K. and Wong, R.N.S., (2010). A Simplified

Method for Quantifying Cell Migration/Wound Healing in 96-Well Plates. *Journal of Biomolecular Screening*, 15(4), 427–433.

Zardavas, D., Irrthum, A., Swanton, C. and Piccart, M., (2015). Clinical management of breast cancer heterogeneity. *Nature Reviews Clinical Oncology*, 12(7), 381–394.

Zarkou, V., Galaras, A., Giakountis, A. and Hatzis, P., (2018). Crosstalk mechanisms between the WNT signaling pathway and long non-coding RNAs. *Non-coding RNA Research,* 3(2), 42-53.

Zhang, Z., Li, H., Jiang, S., Li, R., Li, W., Chen, H. and Bo, X., (2018). A survey and evaluation of Web-based tools/databases for variant analysis of TCGA data. *Briefings in Bioinformatics*.

Zhao, W., Li, Y. and Zhang, X., (2017). Stemness-Related Markers in Cancer. *Cancer translational medicine*, 3(3), 87–95.

Zhongfeng, L., Xuan, W., Kewen, J., Xunming, J., Y., A.Z., Zhiguo, C., Liu, Z., Wang, X., Jiang, K., Ji, X., Zhang, Y.A. and Chen, Z., (2018). TNFa-induced Up-regulation of Ascl2 Affects the Differentiation and Proliferation of Neural Stem Cells. *Aging and disease*, 10.

Zhu, R., Yang, Y., Tian, Y., Bai, J., Zhang, X., Li, X., Peng, Z., He, Y., Chen, L., Pan, Q., Fang, D., Chen, W., Qian, C., Bian, X. and Wang, R., (2012). Ascl2 Knockdown Results in Tumor Growth Arrest by miRNA-302b-Related Inhibition of Colon Cancer Progenitor Cells. *PLoS ONE*, 7(2), e32170.

Zuo, Q., Wang, J., Chen, C., Zhang, Y., Feng, D.-X., Zhao, R. and Chen, T., (2018). ASCL2 expression contributes to gastric tumor migration and invasion by downregulating miR223 and inducing EMT. *Molecular medicine reports*, 18(4), 3751-3759.