THE IDENTIFICATION OF NON-PROTEIN CODING RNA IN THE TUMOUR VASCULATURE

by

KLARKE MICHAEL SAMPLE

A thesis submitted to the University of Birmingham for the degree of DOCTOR OF PHILOSOPHY

Immunity and Infection

College of Medical and Dental Sciences

University of Birmingham

May 2016

Abstract

This thesis describes novel contributions to scientific knowledge in the areas of non-protein coding RNA (ncRNA) and pathological angiogenesis. In the last 15 years a number of studies have sought to identify potential markers of tumour endothelium. However, as of yet, no studies have identified non-protein markers of the tumour vasculature. These 'phantoms of transcription' encompass diverse classes that have long been thought to be merely 'junk', but some have recently been attributed novel functionality.

The data used to identify non-protein markers of the tumour vasculature in this study was obtained through the RNA sequencing (RNAseq) of tumour endothelial cells (TEC) and healthy tissue associated endothelial cells (HEC). The ensuing bioinformatic analysis revealed many differentially expressed short nuclear RNA (snoRNA) and long intergenic ncRNA (lincRNA) molecules. Some of which were confirmed as being specifically expressed in (and differentially expressed between) TEC and HEC using quantitative PCR (qPCR).

One of these molecules, PCAT19, was also functionally explored *in vitro*. This study demonstrated that PCAT19 was specifically expressed in endothelial cells and expressed at lower levels in TEC than HEC (a tumour suppressor-like expression pattern). Moreover PCAT19 was shown to affect the cell cycle *in vitro* at G1 and G2/M, and cause increased levels of apoptosis. Microarray technology was used to reveal the mechanism by which affects the transcriptome of endothelial cells. Through this means a known regulator of tumour suppressors, CBX5 (Chromobox protein homolog 5), was demonstrated to

be co-expressed with PCAT19, following the knockdown and overexpression of PCAT19. PCAT19 is the first ncRNA molecule to be identified as both specific to the endothelium and to be functional in pathological angiogenesis.

Acknowledgements

I would like to extend my thanks to my supervisors Roy Bicknell and Victoria Heath for all the guidance and opportunities they have afforded me throughout my PhD. Also to Lukas and John whose hard work provided a starting point for my doctoral project. I am also especially thankful for the following contributions from these individuals:

Alan Zhuang - extraction of liver, lung and colon endothelial cells.

Joseph Wragg - extraction of kidney endothelial cells.

Nathan Li - assistance with the snoRNA qPCR.

Ivette Hernandez - analysis of acumen data.

Steve Kissane - hybridisation and scanning of the microarray chip.

I couldn't have achieved what I did without the inspiration from all of the muses in the Bicknell group. Not to mention all the help provided by members of the Bicknell group, especially Kabir, Rosie and Puja, who helped make the late nights, weekends and public holidays in the lab more enjoyable and product. I would like to espouse the contributions from my wife Sanny, whose loving support was boundless.

Abbreviations

3’ – Three Prime

5’ – Five Prime

Ap – Apoptosis

Aneu – Aneuploidy

bp – Base pair

BRCA1 – Breast cancer 1

CAGE – Cap analysis of gene expression

cDNA – Complementary DNA

CT – Cycle threshold

D1 – Duplex 1

D2 – Duplex 2

dbEST – Expressed sequence tags database

DF – Dermal fibroblasts

DMEM – Dulbecco's Modified Eagle Medium

DMSO – Dimethyl sulfoxide

dNTPs – Deoxynucleotides

EGFP – Enhanced green fluorescent protein

EST – Expressed sequence tag

FDA – Food and Drug Administration

FPKM – Fragments Per Kilobase of transcript per Million mapped reads

GB – Gigabytes

GWAS – Genome wide association study

HASMC – Human aortic smooth muscle cells

HDR – Homology directed repair

HE – Healthy endothelium

HEC - Healthy tissue associated endothelial cells

HF – High Fidelity

HKEC – Healthy kidney associated endothelial cells

HLE – Healthy liver associated endothelium

HP1 – Heterochromatin protein 1 (HP1)

HUVEC – Human umbilical vein endothelial cells

Ker – Keratinocytes

lincRNA – Long intergenic non-protein coding RNA

M199 – Medium 199

MHC – Major histocompatibility complex

miRNA – MicroRNA

MPSS – Massively parallel signature sequencing

mRNA – Messenger RNA

NCBI – National Center for Biotechnology Information

NCD – Negative control duplex

ncRNA - Non-protein coding RNA

NEB – New England BioLabs

NGS – Next generation sequencing

NSCLC – Non small cell lung cancer

PCR – Polymerase chain reaction

PEI – Polyethylenimine

Pri-miRNA - Primary miRNA

qPCR – Quantitative Polymerase Chain Reaction

RCCEC – Renal cell carcinoma associated endothelial cells

RISC – RNA-induced silencing complex

RNAseq – RNA sequencing

rRNA – Ribosomal RNA

lncRNA – Long non-coding RNAs

ORF – Open reading frame

SAGE – Serial analysis of gene expression

SCARNA – Small cajal body RNA

SEM – Standard error of the mean

siRNA – Small interfering RNA

SNORA – H/ACA box snoRNA

SNORD – Box C/D snoRNAs

snoRNA – Short nucleolar RNA

SNP – Single nucleotide polymorphism

SOC – Super Optimal broth with Catabolite repression

SOLiD – Sequencing by oligonucleotide ligation and detection

TE – Tumour endothelium

TEC – Tumour associated endothelial cells

TLE – Tumour liver associated endothelium

TP53BP1 – Tumour protein P53 binding protein 1

tRNA – Transfer RNA

UCSC – University of California, Santa Cruz

UPL – Universal probe library

UTR – Untranslated region

WNK – Whole/bulk (endothelial cell depleted) kidney tissue

WTSS – Whole transcriptome shotgun sequencing

γIFN – interferon γ

Table of Contents

Chapter 1: Introduction 1

1.1 Characteristics of healthy tissue and tumour associated endothelial cells. 3

1.2 Potential functionality for ncRNA in tumour-associated endothelial cells..... 4

1.3 Past methods for the identification of tumour endothelial cell markers 9

1.4 Whole transcriptome shotgun sequencing of the tumour endothelium..... 15

1.5 Hypotheses and Aims 21

Chapter 2: Methods 23

2.1 Generation and analysis of RNAseq data 23

2.2 Tissue Culture..... 26

2.3 RNA isolation and cDNA production 27

2.4 Universal Probe System (qPCR) 28

2.5 SYBR Green (qPCR) 30

2.6 Molecular Cloning 32

2.7 Plasmid Preparation..... 33

2.8 Gibson Assembly 34

2.9 Gene Overexpression 36

2.10 Acumen cell cycle analysis 37

2.11 siRNA transfection 37

2.12 Two Colour Microarray analysis 38

2.13 Statistical Analysis..... 39

Chapter 3: RNAseq analysis of the colon vasculature 41

3.1 Introduction..... 41

3.2 Results..... 43

3.3 Discussion 47

Chapter 4: Enrichment of snoRNA in the tumour vasculature 50

4.1 Introduction 50

4.2 Background information 51

4.3 The function of snoRNA 53

4.4 Results 55

4.5 Discussion 60

Chapter 5: RNAseq analysis of kidney cancer endothelium 66

5.1 Introduction 66

5.2 Results 69

5.3 Discussion 75

Chapter 6: The functional exploration of PCAT19 78

6.1 Introduction 78

6.2 Results 79

6.3 Discussion 88

Chapter 7: General Discussion 93

References 98

Appendix 1: Table of materials 111

Appendix 2: Renal cell carcinoma patient information 114

Appendix 3: Kidney RNAseq analysis codes 115

A3.1 Quality Checking reads from RNAseq data using FastQC 115

A3.2 Map reads to the genome using Tophat 116

A3.3 Differential expression analysis using Cufflinks 119

A3.4 Graphing with CummeRbund 123

A3.5 Example 1: Create a heatmap 123

A3.6 Example 2: Plot data phylogeny 124

A3.7 Example 3: Cluster Analysis 124

A3.8 Example 4: Gene-level plots 125

Appendix 4: DNase digest qPCR control 126

Appendix 5: Oligonucleotide Sequences 128

Appendix 6: Differential expression qPCR data 130

Appendix 7: Two Colour Microarray Analysis Codes 132

Appendix 8: Novel colon sequences: 133

A8.1 XLOC_029144 (TECA1): 133

A8.2 XLOC_032009 (TECA2): 133

A8.3 XLOC_004164 (TECA3): 137

Table of Figures

Figure 1.1: Angiogenic sprouting and blood vessel growth 4

Figure 1.2: Mature miRNA production and RISC assembly 6

Figure 1.3: The role of PTENP1 as a molecular decoy and tumour suppressor 8

Figure 1.4: The theory behind mapping reads 17

Figure 1.5: Mapping of RNAseq data to a reference genome 18

Figure 1.6: Reads that map to multiple areas of the genome 18

Figure 1.7: Endothelial cell isolation from primary tissues 20

Figure 2.1: SOLiD system of sequencing 24

Figure 2.2: Illumina sequencing system 25
Figure 2.3 Universal probe system of qPCR 29
Figure 2.4: SYBR Green qPCR 31
Figure 2.5: PCR amplification of PCAT19 33
Figure 2.6: A plasmid map of PCAT19 in pWPI 35
Figure 2.7 Gibson assembly of PCAT19 in pWPI 36
Figure 2.8 Determination of gene expression using two-colour microarray 39
Figure 3.1: Decoding colourspace data. 42
Figure 3.2: Reading basespace data and the FASTQ format 42
Figure 3.3: The expression of TECA1 compared to FLOT2 45
Figure 3.4: The expression of HNF1A-AS1 compared to FLOT2 45
Figure 3.5: The expression of TECA2 compared to FLOT2 46
Figure 3.6: The expression of TP73-AS1 compared to FLOT2 46
Figure 3.7: The expression of TECA3 compared to FLOT2 47
Figure 4.1: The impact of SNPs and errors on colourspace data 51
Figure 4.2: Excision of snoRNA from host genes 53
Figure 4.3: SNORD75 and GAS5 expression compared to FLOT2 56
Figure 4.4: SNORD76 and GAS5 expression compared to FLOT2 56
Figure 4.5: the enrichment of SNORD76 from GAS5 compared to SNORD75 57
Figure 4.6: SCARNA7 and KPNA4 expression compared to FLOT2 57
Figure 4.7: SNORA81 and EIF4A2 expression compared to FLOT2 58
Figure 4.8: SNORD32A and RPL13A expression compared to FLOT2 58
Figure 4.9: SNORD30 and SNHG1 expression compared to FLOT2 59

Figure 4.10: SNORD100 and RPS12 expression compared to FLOT2 59

Figure 4.11: Alternative splicing of GAS5 62

Figure 5.1: Phylogeny of the tumour vasculature..... 70

Figure 5.2 Cluster analysis of differentially expressed genes 72

Figure 6.1: PCAT19 contains a prostate cancer associated SNP..... 79

Figure 6.2: RNAseq gene level plot for PCAT19..... 80

Figure 6.3: PCAT19 is expressed to a lesser extent in TLE 80

Figure 6.4: PCAT19 expression correlates with cell density 81

Figure 6.5: Post-lentivirus transduction expression levels of PCAT19..... 81

Figure 6.6: The effect of PCAT19 on the cell cycle of HUVEC 82

Figure 6.7:Confirmation of PCAT19 knockdown in HUVEC..... 83

Figure 6.8: WTAP expression when PCAT19 is overexpressed..... 85

Figure 6.9: HIST1H2BK expression when PCAT19 is overexpressed..... 85

Figure 6.10: CBX5 is upregulated when PCAT19 is overexpressed..... 86

Figure 6.11: SUMF1 expression when PCAT19 is overexpressed 86

Figure 6.12: ILII expression when PCAT19 is overexpressed 87

Figure 6.13: CNN1 expression when PCAT19 is overexpressed. 87

Figure 6.14: HMOX1 expression when PCAT19 is overexpressed. 88

Figure 6.15: Predicted involvement of PCAT19 with the HP1-BRCA1 pathway ... 91

Figure 7.1: Articles pertaining to “angiogenesis and non-protein coding RNA” 93

Figure S1: No cDNA qPCR control 127

Table of Tables

Table 1.1: Anti-angiogenic therapies used in clinical trials 2

Table 1.2: Known potentially endothelial-specific genes 15

Table 3.1: Tumour and healthy colon endothelium SOLiD RNAseq data 43

Table 4.1: Differentially expressed snoRNA in the lung and colon vasculature 55

Table 5.1: Strongly tumour endothelial specific genes..... 73

Table 5.2: Genes enriched in the healthy endothelium..... 74

Table 5.3: Top significantly differentially expressed ncRNAs 74

Table 6.1: Differentially expressed genes following the knockdown of PCAT19 .. 84

Table S1: Table of materials 113

Table S2: Clinical and-pathological data for patients in the RNAseq analysis 114

Table S3: Oligonucleotide sequences..... 129

Table S4: Differential expression data (qPCR screening of different cell types). 130

Table S5: Differential expression data (qPCR analysis of sparse HUVEC)..... 130

Table S6: Differential expression data (qPCR validation of microarray data) 131

Table S7: Differential expression data (qPCR validation of PCAT19 siRNA) 131

Chapter 1: Introduction

The overall aim of the research projects detailed in this thesis was to identify non-protein coding RNA markers within endothelial cells, which compose the innermost layer of blood vessels. The acquisition of a vasculature is a fundamental component leading to the formation of solid tumours [1] and has been investigated as a possible anti-cancer target for nearly half a century [2,3]. Folkman [2] was the first to propose the concept of an anti-angiogenic therapy, which could prevent tumour growth by stopping the penetration of new blood vessels. Hence, the overarching dogma of such anti-angiogenic therapies is that a tumour will die once it has been denied the nutrients and oxygen provided by the blood vessels within it.

In 1993 Burrows and Thorpe [4] were the first to demonstrate that it was possible to achieve an anti-cancer effect by targeting tumour-associated blood vessels in mice. This proof of principle was achieved by utilising subcutaneous neuroblastoma cells expressing murine interferon γ (γIFN). The γIFN induced an inflammatory response in the tumour vasculature and caused the expression of major histocompatibility complex (MHC) class II antigens. An anti-cancer effect was then achieved through the use of ricin-A chain conjugated anti-mouse class II antibodies, which caused haemorrhaging, necrosis and subsequently tumour regression. This study raised the possibility that a therapeutic agent could target cancer should a suitably specific target be found within the tumour-associated vasculature.

Targeting tumour-associated endothelial cells is a particularly attractive strategy because they have direct and intimate contact with the blood and are therefore accessed comparatively easily by therapeutic agents [5]. The specificity of anti-angiogenic strategies is also an advantage over traditional chemotherapeutic agents. Chemotherapeutic agents target both cancerous and normal cells, which results in a narrow therapeutic window and severe side effects [6]. For these reasons, many anti-angiogenic strategies have been explored in clinical trials (Table 1.1).

Type	Drug	Target
Antibodies	AMG-386	Angiopoetin 1 and 2
	Bevacizumab	VEGF
	VEGF trap	VEGF
	IMC-1121b	VEGFR2
	IMC-18F1	VEGFR1
Receptor tyrosine kinase inhibitors	AV-299	HGF
	Apatinib	VEGFR2, RET, c-KIT, c-SRC
	Axitinib	VEGFR, PDGFR
	BIBF 1120	VEGFR, PDGFR, FGFR
	Brivanib alaninate	VEGFR2, FGFR
	Crizotinib	cMET
	Erlotinib hydrochloride	EGFR
	Enzastaurin hydrochloride	PKC-beta
	Foretinib	VEGFR, c-MET, FLT3, c-KIT
	Linifanib	VEGFR, PGDFR, FLT1, FLT3, CSF-1R, c-KIT
	MetMab	c-MET
	Pazopanib hydrochloride	VEGFR, PGDFR, c-KIT
	Sunitinib malate	VEGFR, PGDFRb, FLT3, CSF-1R, c-KIT
	Sorafenib tosylate	VEGFR, PDGFR, c-KIT, RAF
	Vandetanib	VEGFR2, EGFR
	Vatalanib	VEGFR2, EGFR
	XL184	VEGFR, c-MET, RTK, FLT3, TIE2
Integrin Inhibitors	Cilengitide	αvβ3, αvβ5 antagonist
mTOR inhibitors	Temsirolimus	mTOR
	Everolimus	mTOR
	Ridaforolimus	mTOR

Table 1.1: Anti-angiogenic therapies used in clinical trials

The anti-angiogenic therapies in this table have all been investigated in clinical trials and in a variety of tumour types [7].

1.1 Characteristics of healthy tissue and tumour associated endothelial cells.

There are multiple precipitating factors that induce transcriptional changes within tumour-associated endothelial cells. Many of these differences are caused by the microenvironmental changes that occur during the progression of healthy tissues towards that of a cancerous nature [8, 9, 10]. The blood vessels contained within solid tumours are structurally abnormal. This in turn leads to convoluted blood flow that is impeded due to an absence of the conventional hierarchical organisation between arteries, veins and capillaries [10]. The resulting hypoxia, low pH [11, 12] and low shear stress [13] fundamentally alters the RNA and protein repertoires with endothelial cells.

However transcriptional differences can be induced by more subtle changes within the endothelial cells themselves, such as the switch from quiescence to an actively angiogenic phenotype [14]. An excessively angiogenic phenotype within endothelial cells is promoted by tumours due to the excess secretion of factors including Vascular Endothelial Growth Factor (VEGF), which is a significant regulator of angiogenesis leading to capillary sprouting (Figure 1.1) [15, 16].

The process by which existing blood vessels form new blood vessels is termed angiogenesis. Angiogenesis is an essential component of normal development and wound healing. Conversely the mechanisms involved with angiogenesis can be hijacked and act as precipitating factors in the etiology of many diseases. Endothelial cells in particular contribute to the pathological processes involved with cancer through the aforementioned angiogenic

mechanisms. But whilst an ever-increasing number of molecular changes have been identified, many probably remain as yet undefined [9, 15-18].

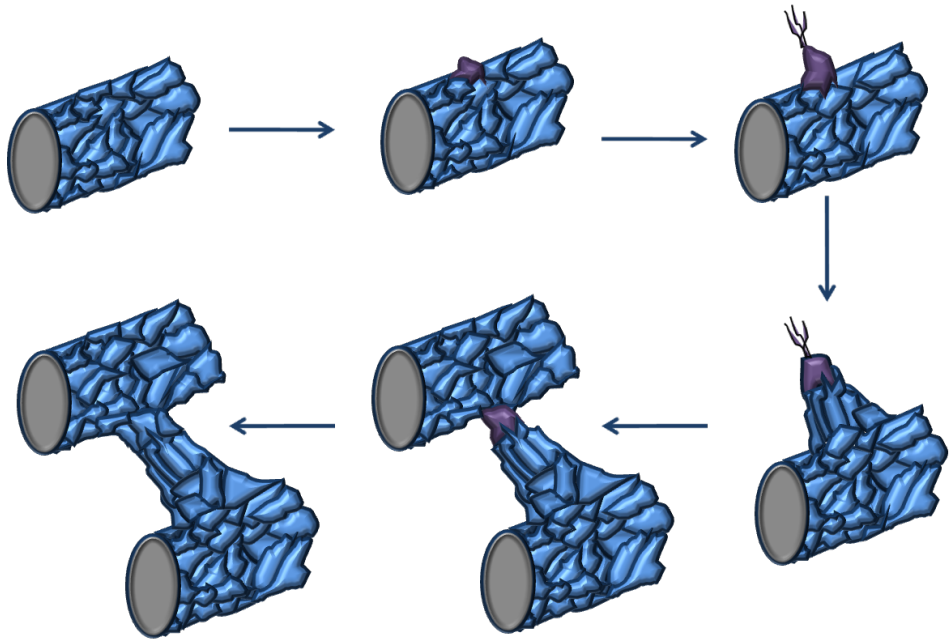


Figure 1.1: Angiogenic sprouting and blood vessel growth

Angiogenic stimuli can cause the differentiation of endothelial cells to form tip cells. Tips cells digest and migrate through the basement membrane towards the angiogenic stimuli. In the wake of these tip cells neighbouring endothelial cells proliferate and assemble to form capillary-like sprouts [19]. The newly formed sprouts extend and eventually fuse with other blood vessels (or themselves) to establish blood flow (original diagram).

1.2 Potential functionality for ncRNA in tumour-associated endothelial cells

In man, there are approximately 20,000 expressed genes are encoding within a mere 1.5% of the DNA and the rest has long been thought to be junk [20-21]. It is curious that organisms as far apart on the evolutionary scale such as C. (Caenorhabditis) *elegans* and man have similar numbers of expressed genes, and

the difference in their genomes lies in the amount of 'junk' DNA (26% in C. *elegans* versus 98.5% in man) [21]. Some schools of thought are of the opinion that therein lies one explanation for the increase in complexity. For example, we now know that a vast amount of this 'junk' DNA is transcribed and contains many regulators of the expression of the transcribed genes and transcribed pseudogenes [20]. These include microRNA (miRNA) and anti-sense RNA as well as more 'established' non-coding RNAs such as ribosomal RNA (rRNA) [22]. Currently there are about 75,000 transcripts in the reference genome that have been annotated as being non-coding RNAs, comprising of ~800 snoRNA (short nucleolar RNA), ~80,000 lncRNAs (Long non-coding RNAs), ~4,500 miRNAs, 4000 rRNAs and ~1,250 tRNAs (transfer RNA) [23].

miRNA typically ranges in size from about 21 to 23 nucleotides. Their primary function appears to be the post-transcriptional regulation of target genes and they do this by binding to complementary or partially complementary messenger RNA (mRNA) sequences, typically in the three prime untranslated region (3' UTR). The silencing activity of miRNA is achieved by utilising the RNA induced silencing complex (Figure 1.2). If the miRNA is partially complementary to the mRNA target translation will be blocked, however if their sequences are fully complementary the mRNA will be degraded [24-26].

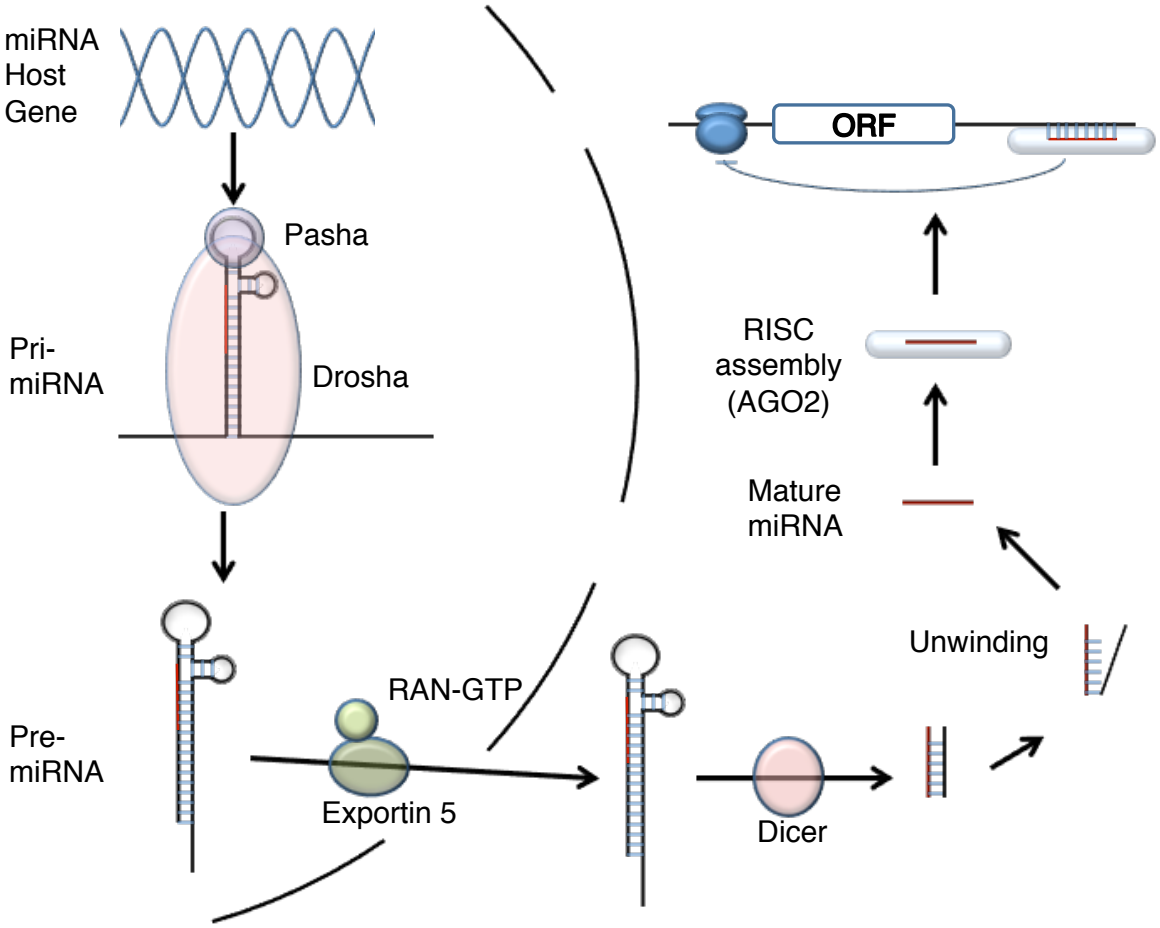


Figure 1.2: Mature miRNA production and RISC assembly

Primary miRNA (Pri-miRNA) can be expressed from a variety of transcripts including miRNA cluster host genes, the three prime untranslated regions (3' UTR) of protein coding genes and from within the introns of protein coding genes after splicing has occurred. The pre-miRNA is cleaved from the pri-miRNA by the enzyme Drosha based upon the shape of miRNAs hairpin loop. The pre-miRNA is then exported to the cytoplasm where Dicer can further process it into mature miRNA. The mature miRNA can then be used to modulate mRNA expression once Argonaute 2 (AGO2) has assimilated it into the RNA-induced silencing complex (RISC). RNA is most often targeted by miRNA downstream from the open reading fame (ORF), in the 3' UTR. If the miRNA is fully complementary to the target sequence the RISC will degrade the mRNA. If the miRNA is only partially complementary the ribosome will be inhibited to silence the expression of the transcript (original diagram).

lncRNA or ‘phantoms’ are characterised as having sequences that resemble genes, but have been traditionally thought to be biologically inconsequential. This assumption was made because of the presence of truncating mutations, frameshifts and other mutations that would not allow for translation into functional proteins. However, there is evidence that pseudogenes are transcribed into “phantom mRNA” and that these non-protein coding RNAs could have roles within the cell. This theory is supported by the high degree of conservation of nucleotide sequences within pseudogenes, showing that a selective pressure is placed on these genes. The transcription of pseudogenes can be tissue-specific and activated in cancer, suggesting that the expression of phantom mRNA could have a significant impact on angiogenesis and carcinogenesis. Nevertheless. few pseudogenes have been functionally characterised [22; 24; 27].

The pseudogene PTENP1 (Phosphatase and Tensin Homolog Pseudogene 1) is a good example of this action as its transcript acts as a decoy for mRNA of the PTEN (Phosphatase and Tensin Homolog) tumour suppressor gene and allows for the expression of PTEN to occur in the presence of miRNAs (Figure 1.3). PTENP1 is derived from a retrotranscription event, but has a mutated start codon, which stops the mRNA from being translated into a protein. Despite being a truncated form of PTEN, PTENP1 contains five conserved sites for miRNA in its 3’ untranslated region. The decoy effect of phantom mRNA appears to be essential in maintaining the activity of PTEN and reducing tumourgenicity, the knockout of PTENP1 is associated with a decrease of PTEN mRNA and protein levels, which

in turn results in accelerated cell proliferation and cancer through the Akt/PKB signaling pathway [24, 28].

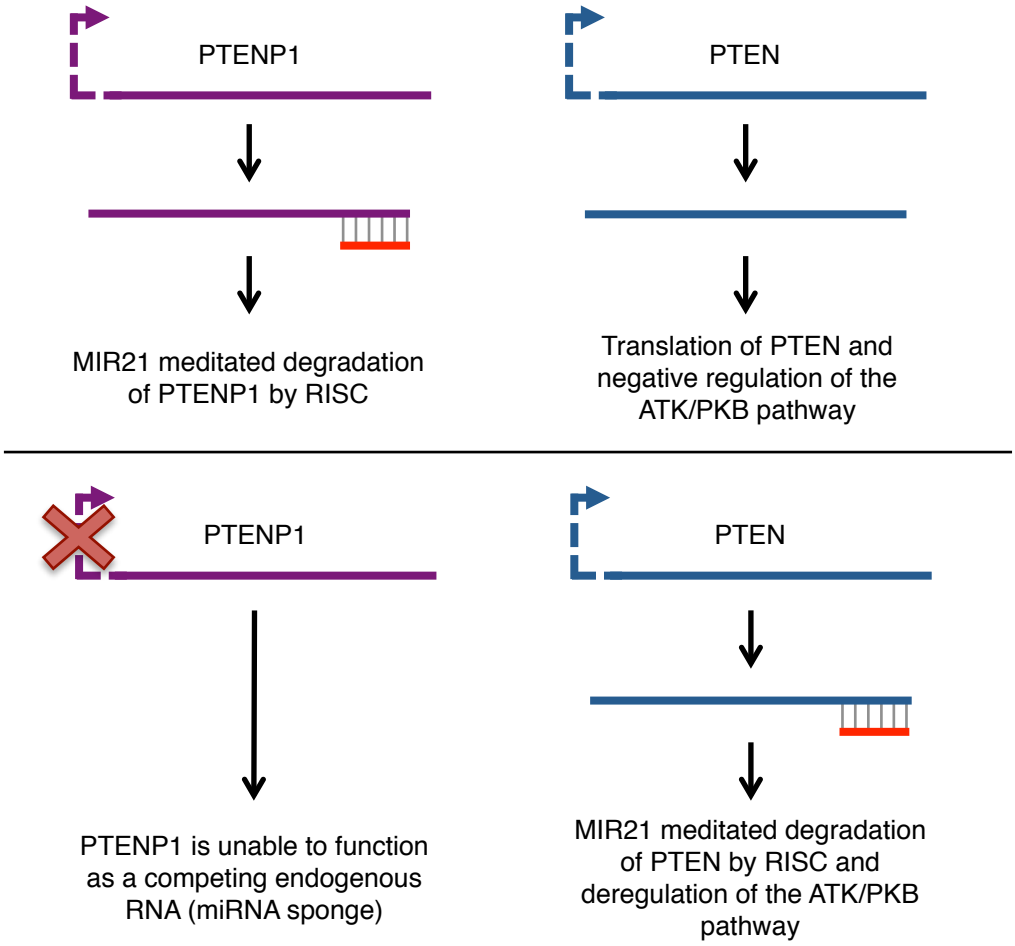


Figure 1.3: The role of PTENP1 as a molecular decoy and tumour suppressor

The pseudogene, PTENP1, shares a large part of the three prime (3') untranslated region (UTR) of the protein-coding gene, PTEN. Any miRNAs capable of degrading PTEN will also be capable of degrading PTENP1 via the RISC, providing the two transcripts share a similar target region. miRNAs such as miRNA21 will be capable of degrading both PTEN and PTENP1, but PTEN is less likely to be degraded if PTENP1 is expressed (due to miRNA binding site competition). Therefore when the expression of PTENP1 is reduced, PTEN is more likely to be degraded (original diagram).

It is important to note that phantom mRNAs do not just act as miRNA sponges; they could have many other functions. A recent article by Han *et al*. [27] highlights one of these possible functions, when the pseudogene MYLKP1 (myosin light chain kinase pseudogene 1) is overexpressed it can inhibit smMLCK (smooth muscle myosin light chain kinase) by decreasing RNA stability, however the precise enzymatic mechanism is not yet known. But it has been suggested that a 5' (five prime) UTR acting RNA-destabilizing factor could be responsible [29].

Like protein coding genes, non-coding RNAs (ncRNAs) appear to have tissue-specific profiles. It is possible that some endothelial or tumour endothelial-specific ncRNAs exist. McCall *et al*. [30] demonstrated that 6 different endothelial cell types had different miRNA expression profiles. Furthermore they demonstrated that 31 miRNAs were possibly unique to the endothelium by comparing the 166 miRNA expressed in endothelial cells to epithelial and haematologic cells. It is therefore possible that endothelial cells specifically express transcripts from the other ncRNA classes, such as lncRNA and snoRNA.

1.3 Past methods for the identification of tumour endothelial cell markers

In 2000 Huminiecki and Bicknell [31] combined two data mining approaches, which led to the identification of sixteen genes that were specifically expressed in endothelial cells, including four previously unidentified (novel) genes. A high stringency BLAST was used to screen a pool of nine human endothelial cell

libraries against the 'UniGene gene index' and 108 non-endothelial libraries derived from the 'Expressed Sequence Tags database' (dbEST).

The second approach utilised internet-based (SAGEmap exprofiler) SAGE (serial analysis of gene expression) library subtraction. However, cross-referencing between the expressed sequence tag (EST) and SAGE library analyses was required to accurately identify genes preferentially expressed in endothelial cells. Individually the two methods produced large numbers of false positives.

The study by Huminiecki and Bicknell [31] was not aimed at identifying specific markers of the tumour-associated endothelium. However, some of the publically available endothelial cell libraries included actively angiogenic cells, such as HUVEC (human umbilical vein endothelial cells). One of the four novel genes, ROBO4 (roundabout, axon guidance receptor, homolog 4), was later shown by Huminiecki *et al*. [32] to be specific to the tumour-associated endothelium and play a role in angiogenic processes, such as endothelial cell migration [33]. Furthermore, anti-angiogenic (and therefore anti-cancer) therapies against ROBO4 have been explored via *in vivo* vaccination by Zhuang *et al*. [34] and drug conjugated anti-ROBO4 antibodies by Yoshikawa *et al*. [35].

A method was later developed by the Herbert *et al*. [36, 37] to reduce the false positive rate associated with their earlier study. The complementary DNA (cDNA) data mining method utilised improved statistical analysis and an assignment of ESTs. These improvements allowed the identification of a further fourteen genes that were specifically expressed in endothelial cells. This method also predicted

‘one hundred and sixty’ genes to be upregulated in endothelial cells. When combining this method with the SAGEmap xProfiler, the list was expanded to include the accurate prediction of 58 genes that were specific to endothelial cells and a further 459 upregulated genes. Finally, a list of 27 potential tumour endothelial restricted genes was predicted by subtracting bulk tumour cDNA libraries from healthy tissue libraries.

The two previously described studies were both able identify genes that were expressed specifically in the tumour-associated endothelium. However they had one overarching disadvantage in this respect. The publically available libraries were all derived from healthy endothelial cells, albeit sometimes foetal and therefore actively angiogenic.

St. Croix *et al*. [8] were able to overcome the disadvantage of these studies and identify a number of markers in the tumour-associated endothelium. The generation of two SAGE libraries from purified healthy colon associated endothelial cells and colorectal tumour-associated endothelial cells enabled this. When assembled, the SAGE tags generated roughly 32,703 unique transcripts (after the exclusion of repetitive tags). It was by comparing these transcripts to SAGE libraries from non-endothelial sources that St. Croix *et al*. [8] were able to identify 93 transcripts that were a minimum of twenty-fold higher in endothelial cells. But the key part of this study was the comparison of the healthy tissue and colorectal tumour-associated endothelial cell libraries. A minimum of ten-fold higher expression was observed in 46 tags from the colorectal tumour-associated endothelial cell library. Of the top 25 tags, eleven corresponded to known genes (6

were known markers of endothelial cells at the time) and the remaining 14 tags were derived from areas of the genome that were not yet known to be genes. Nine of these tags were confirmed to be specific to the tumour-associated endothelium by *in situ* hybridisation.

Many of the tumour-associated endothelial cell markers from this study have been the subject of further research as potential anti angiogenic therapies. Including TEM8, the anthrax toxin receptor 1. TEM8 has been targeted for the treatment of numerous tumour types *in vivo* using antibodies [38] and even using modified forms of the anthrax toxin [39, 40].

The studies conducted by Huminiecki and Bicknell [31], Herbert *et al*. [36, 37] and St. Croix *et al*. [8], demonstrate the ability for EST libraries to successfully identify endothelial cell-specific and tumour endothelial cell-specific genes. However, these libraries are generally constructed using SAGE, CAGE (cap analysis of gene expression) and MPSS (massively parallel signature sequencing). These methods utilise 'Sanger sequencing' and therefore are very labour intensive, low throughput and often prohibitively expensive. This reduces the feasibility of producing new EST data if the existing libraries are not sufficient to answer a hypothesis or to control for biological and technical variability. Conjointly, even should public libraries exist, there is no guarantee that they have been sequenced at a sufficient depth to enable the detection of biologically important transcripts that are expressed at low levels. Furthermore the tags produced by these technologies are very short and are often not able to span repetitive elements in

the genome. For this reason, it can also be extremely difficult to distinguish between splice variants [41-43].

Microarrays were utilised by Zhang *et al*. [14] and Ghilardi *et al*. [44] as a high throughput alternative method to probe the transcriptome of endothelial cells. The aim of both these studies was to reveal the transcriptional differences within endothelial cells. The former utilised quiescent and proliferative (exposed to growth factors such as VEGF) foreskin microvascular endothelial cells. Whereas the latter used endothelial cells that were isolated from a number of healthy and cancerous tissues. Ho *et al*. [45] utilised microarray technology to compare four types of endothelial cells to five non-endothelial cell types and successfully identified several previously unknown endothelial-specific genes. These studies all used microarray technology that was relatively advanced at the time. However, they were severely limited by the number of probes used and targeted 588 [14], 12,000 [44], 672 [45] genes respectively.

Ultimately the targeted nature of microarray technology means that it can provide data in a manner that is more cost effective and less labour intensive when compared to SAGE libraries. Notwithstanding this fact, SAGE libraries produce superior data. SAGE libraries are not limited by the number/variety of probes, as even the most advanced microarrays currently available can't detect areas of the genome that they are not designed to detect (such as unknown genes, repetitive regions or areas perceived to not be important) [43]. A further disadvantage of microarrays is that the data comes in the form of continuous measures (fluorescent fold change to a reference), rather than absolute values

(raw read counts). This feature makes it very difficult to compare data across experiments and determine which genes have a relevant expression pattern [42].

Endothelial cells have long been thought to be among the most transcriptionally rich cell types [46-48], and it is indeed true that many endothelial-specific genes have been discovered (Table 1.2). However, RNA sequencing (RNAseq) could reveal many more biologically interesting genes and ncRNAs that were missed by these studies. These genes could be identified not just because of the advantages of RNAseq over the SAGE and microarrays (as discussed in Section 1.4), but also by screening the transcriptome of healthy and tumour endothelium derived from different organ/tissue types. The support for this predicted diverse gene expression pattern derives from the requirement of endothelial cells to adapt to varying blood flow, pressure and microenvironments, and accommodate the needs of individual tissues [46-48].

Through a global gene expression analysis of 53 different endothelial cell types using microarrays, Chi et al [46] demonstrated that endothelial cells derived from large vessels had pervasive gene expression differences when compared to microvascular associated endothelial cells, and characteristic expression gene expression profiles in endothelium of arterial and venous sources. Moreover, Chi et al [46] identified that endothelial cells from different organs have distinct gene expression profiles, which raises the possibility that ncRNAs could be expressed specifically in endothelial cells from different tissues.

Gene	Ref.	Gene	Ref.	Gene	Ref.	Gene	Ref.
ANG	8	EMCN	36, 45	MMRN1	31, 36, 45	RAMP2	31
ANGPT2	31,45	ERG	36	MYCT1	36, 45	RAPGEF3	36
ARHGAP24	36	ESM1	45	NESH	8	RASIP1	31, 45
ARHGEF15	36	FABP4	31	NOD27	36	RHOJ	36, 45
BMP1	45	FAM124B	36	NOSTRIN	36	ROBO4	31, 36, 45
BMX	36	FGD5	31, 45	NTN4	8	S14L1	45
CALCRL	36	FZD4	36	PAK2	45	S1PR1	45
CD34	36	GIMAP6	36	PALMD	45	SDPR	45
CD93	31, 36, 45	GIMAP7	45	PCDH12	36	SEC14L1	45
CDH5	31, 36, 45	GNA11	8	PECAM1	31, 45	SELE	36
CLEC14A	45	HHIP	36, 45	PEM1	8	SERPINE1	45
COLIVA1	31	ICAM2	45	PEM2	8	SHE	45
COLIVA2	8	IFITM1	8	PEM3	8	SLC35A2	36
COLV1A1	8	IGFBP4	8	PEM5	8	SMURF2	36
COLVIA2	8	IL33	36	PEM6	8	SOX7	36
COLXVIIIA1	8	KDR	45	PEM7	8	SPARC	8
CRP2	8	LAMA4	45	PEM9	8	SPARCL1	8, 36
CTGF	8, 45	LDB2	45	PLA2G4C	36	SRPX	36
ECSCR	31, 36, 45	LPHN1	8	PLSCR4	45	TCF4	36, 45
ECSM1	31	MCAM	8, 45	PODXL	45	THBS1	45
EDG1	36	MCF2L	8	PPP1R16B	45	TMSB4X	45
EDN1	31, 45	MFNG	45	PRKACA	45	TNS2	8
EFEMP1	31, 45	MGP	8, 45	PROCR	45	VWF	8, 31, 36, 45
ELTD1	36, 45	MMP1	36	RAD54L2	36	ZNF521	45

Table 1.2: Known potentially endothelial-specific genes

The genes in this table have been predicted to be endothelial-specific using SAGE and microarray analysis. It is important to note that many of these genes may have little/no utility as anti-angiogenic targets due to being highly expressed by the healthy vasculature, or because of a non-angiogenic function (such as platelet adherence). On the other hand, the function of the gene is irrelevant when using anti-endothelial cell therapies, because the therapy is designed to kill the cell, rather than inhibit a specific function.

1.4 Whole transcriptome shotgun sequencing of the tumour endothelium

RNAseq uses technology that is known by many names: next (or 2^{nd}/3^{rd}) generation sequencing, deep sequencing and whole transcriptome shotgun

sequencing (WTSS). These names all hint at the nature of the technology being used. RNAseq has the ability to deliver data regarding the whole transcriptome and this data comes in the form of discrete counts (the frequency of a specific sequence) known as reads. Once the reads are sequenced they can be aligned to a reference to determine which genes were present in a sample (Figure 1.4, 1.5 and 1.6). RNAseq has the strengths associated with both SAGE and microarray analysis. It shares the high throughput and economical advantages of microarrays and the significant advantages of SAGE (as discussed earlier). RNAseq data also allows for information regarding the expression of individual genes to be determined easily. Namely whether the genes of interest are differentially expressed, abundant or completely absent [42, 49]. Moreover the vast number of reads produced across the full length of transcripts allows for splice variant and novel transcripts to be more easily identified, quantified and validated [43].

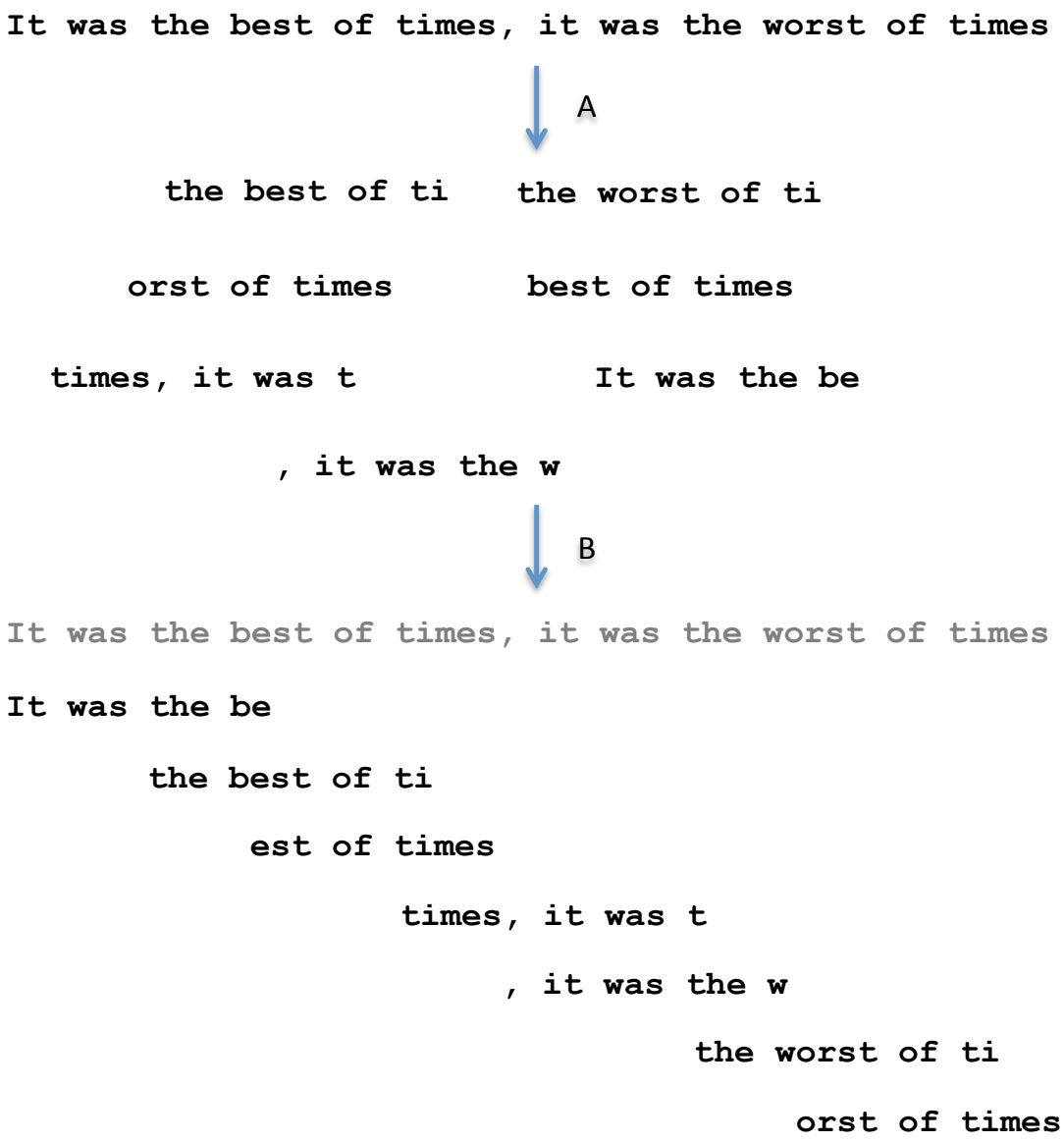


Figure 1.4: The theory behind mapping reads

The process of sequencing and mapping WTSS data is essentially akin to taking thousands of copies of 'A Tale of Two Cities' by Charles Dickens, shredding them into millions of small chunks (reads) (A) and aligning the shredded reads to a complete version to reconstruct the books (B). The process of mapping reads to a reference is somewhat easier and computationally less intensive than assembling a genome. Genome assembly necessitates the reassembly of reads by comparing the overlapping sections of the reads to reform the book without a reference (original diagram).

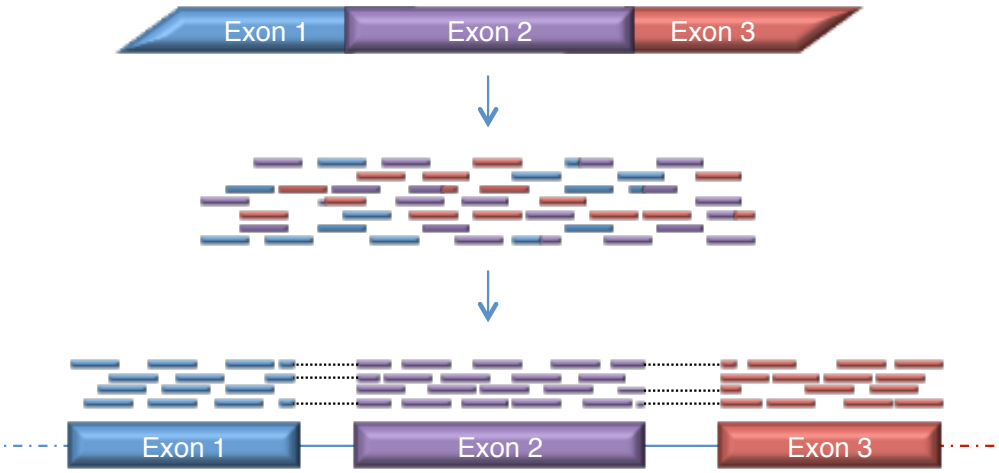


Figure 1.5: Mapping of RNAseq data to a reference genome

The alphameric files (refer to Figures 1.4 and 3.2) generated through the analysis of sequenced and fragmented cDNA are bioinformatically re-constructed and compared to a reference genome. The overlapping areas between the reads allow for individual RNA molecules to be deciphered and quantified. The overlapping regions and gaps ultimately allow for the identification of exon-intron boundaries and even the characterisation of splice variants (original diagram).

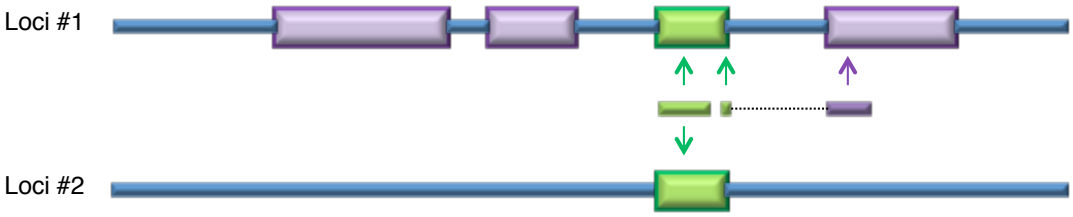


Figure 1.6: Reads that map to multiple areas of the genome

Individual reads can sometimes be difficult to assign to the genome if it contains a repetitive or non-unique sequence. It is more likely for longer reads to span these regions. But this problem can be partially negated using paired-end sequencing on short read platforms, where two reads are produced from each end of a cDNA fragment. The location of a paired read can give information regarding the location of a non-unique read and reduce the number of reads discarded (original diagram).

There are many deep sequencing platforms currently available. At the most basic level the difference between 2nd and 3rd generation platforms is the ability for the latter to detect the nucleotide sequence directly (i.e. without fluorescence) and without polymerase chain reaction (PCR) amplification. Illumina and SOLiD (sequencing by oligonucleotide ligation and detection) are the most commonly used 2nd generation platforms, whereas Ion Torrent, Pacbio and Nanopore are all 3rd generation platforms. The Illumina and SOLiD platforms were used to generate the data in this thesis, their chemistry and relative strengths and weaknesses will be discussed at length throughout Chapters 2, 3, 4 and 5. However, the chemistry and relative strengths and weaknesses of the 3rd generation platforms are beyond the scope of this thesis (these platforms produce smaller amounts/depths of very long reads that are better suited to genomics, rather than RNAseq), but the following reviews by Liu *et al*. [49], Branton *et al*. [50], Mardis [51] and Glen [52] provide a good introduction.

Ultimately the first step in achieving accurate insights into the expression profiles of endothelial cells within tumours or healthy tissues necessitates the robust and reproducible isolation of endothelial cells (Figure 1.7). But the choice of sequencing platform and experimental protocol is also critical to the success of the project. Of the aforementioned platforms SOLiD and Illumina were chosen to perform the sequencing in this thesis. This is largely due the current technological trade off between the long reads lengths required to span repeats and high sequencing depths required to identify less abundant transcripts, of which the hindmost was deemed to be more crucial.

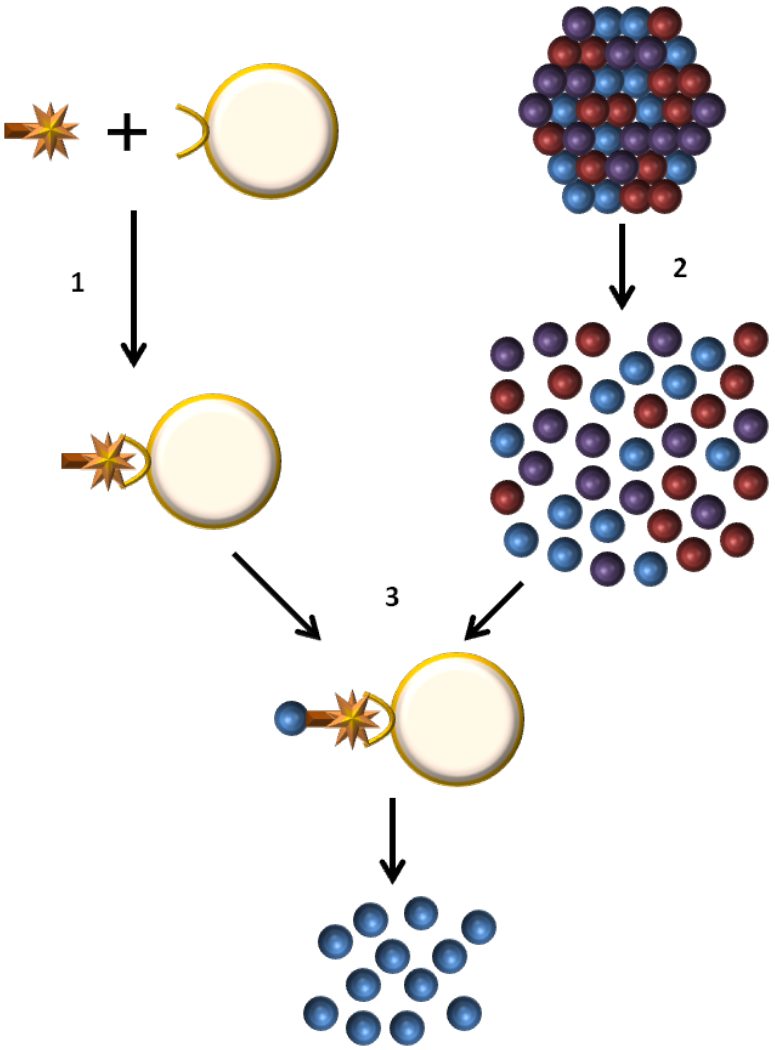


Figure 1.7: Endothelial cell isolation from primary tissues

1) Ulex lectin biotin complexes are bound to streptavadin Dynabeads.

2) A single cell suspension is obtained by dissociating primary tissue using collagenase. It's important to note that the purity of the extraction is dependent upon successfully obtaining a single cell suspension. Contamination of the single cell isolate can occur if other cells are bound to the endothelial cells..

3) The Ulex lectin binds to endothelial cell-specific glycosylation, which can be pulled out of the suspension using the magnetic dynabeads. These cells can be lysed to provide RNA for downstream molecular genetic analyses (original diagram).

1.5 Hypotheses and Aims

As mentioned earlier in this thesis, McCall *et al*. [30] identified a number of miRNAs that were expressed specifically in endothelial cells when compared to other hematologic cells. However their findings from microarrays were not validated experimentally to confirm the expression patterns and biological relevance. Moreover to date there have been no analogous reports in the scientific literature demonstrating the endothelial-specific expression of other non-coding RNA classes.

Aims:

1) To identify non-protein coding RNAs that are differentially expressed between tumour and healthy tissue associated endothelial cells. RNAseq was used to profile endothelial cells from a variety of tumour types, namely colorectal carcinoma, non-small cell lung cancer and renal cell carcinoma.

2) To identify non-protein coding transcripts which are expressed solely in endothelial cells. This was facilitated by the use of quantitative PCR (qPCR) to compare the expression of the candidates (from the bioinformatic analysis) in non-endothelial cell types. The additional sequencing of endothelial cell depleted bulk tissue was also used to help identify potential endothelial-specific candidates, through the elimination of genes that were more likely to be expressed ubiquitously.

3) The final aim of this project was to determine whether any of the non-coding RNA molecules identified during this project have relevance to angiogenesis and endothelial cell biology. In the case of PCAT19, this aim was achieved through acumen cell cycle analysis.

Chapter 2: Methods

A table of materials and their sources can be found in Appendix 1

2.1 Generation and analysis of RNAseq data

SOLiD4 RNAseq (Figure 2.1) data was analysed to identify differentially expressed ncRNA, Herbert [53] had previously generated the data. The data was derived from a single set of a patient matched colorectal cancer and healthy colon associated endothelial cells. Illumina (Figure 2.2) data derived from non-small cell lung cancer and healthy lung associated endothelial cells as described by Herbert [53] was subsequently compared to the SOLiD4 data. The analysis of these two datasets allowed for the identification of any non-protein coding RNA molecules that were differentially expressed in both lung and colon tumour associated endothelial cells.

The RNA from renal cell carcinoma associated endothelial cells (RCCEC) and healthy kidney associated endothelial cells (HKEC) were obtained (and the isolate purity confirmed by qPCR) from three patients (staging information can be found in Appendix 2) by Joseph Wragg as described by Mura *et al* [54]. Furthermore RNA was obtained from the 'endothelial cell depleted bulk healthy kidney tissue' for all three patients. The 9 sets of RNA were Poly-A selected, barcoded, pooled and paired-end deep sequenced across two lanes on an Illumina HiSeq2000 by the Beijing Genomics Institute in Hong Kong, which produced 90bp (base pair) reads from 160bp fragments. In total, this experiment produced four files (~2.8 gigabytes (GB) of data each) per condition, per patient (forward and reverse reads from two lanes). The codes for the tuxedo pipeline used to analyse the data (~100 GB in total) can be found in Appendix 3.

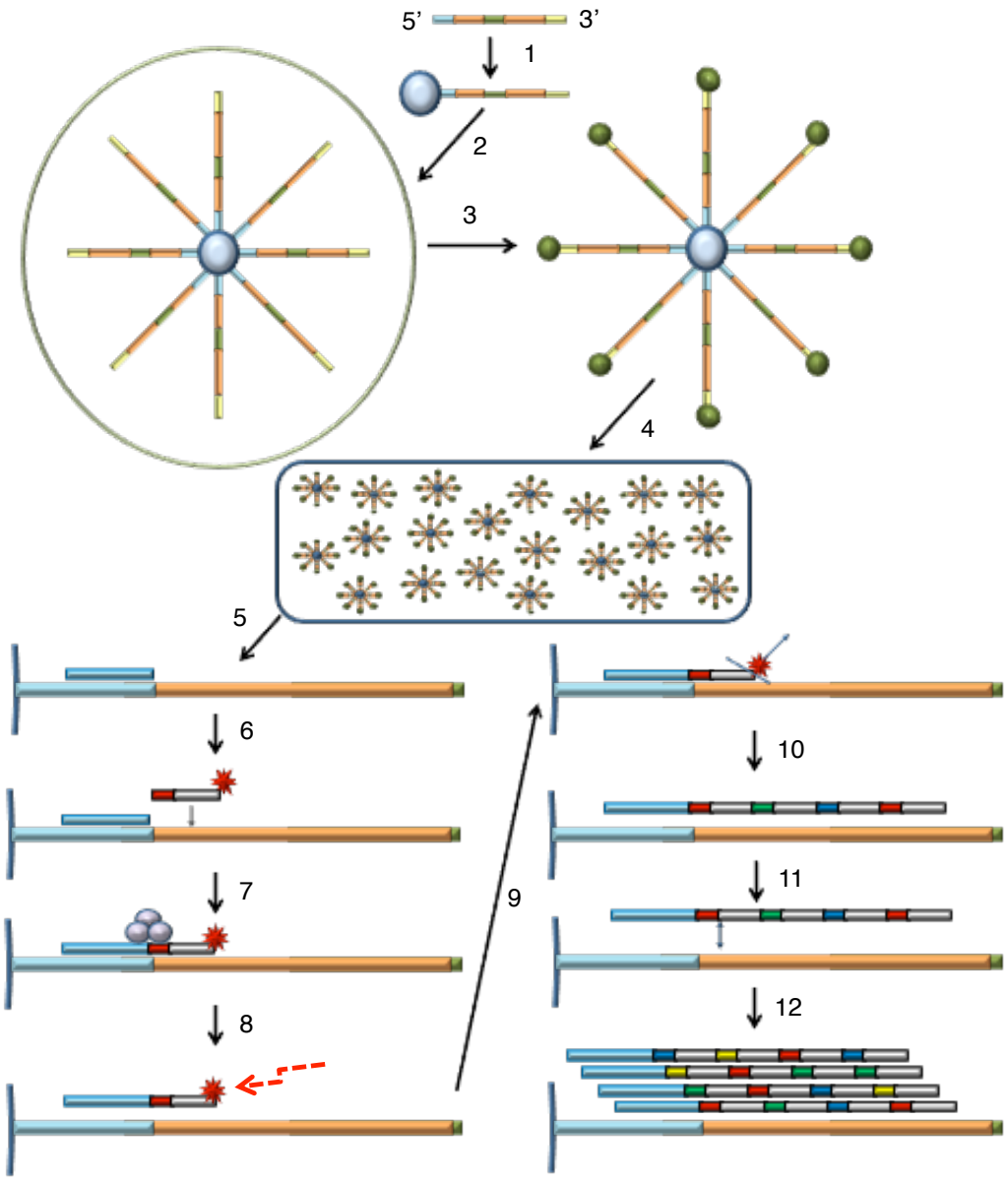


Figure 2.1: SOLiD system of sequencing

RNA is fragmented, converted into cDNA and ligated to adapter sequences (5’ adaptor - fragment 1-adaptor - fragment 2 - 3’ adaptor) to form a mate-paired library, which is used for the emulsion PCR, bead enrichment and sequencing by ligation:

1. The 5’ adaptor on the cDNA fragments is ligated to magnetic beads;

2. Clonal amplification of the cDNA construct by emulsion PCR;

3. Polystyrene beads are ligated to identify beads with successfully amplified constructs;

4. Attachment of beads to a glass slide;

5. Hybridisation of the universal primer to the cDNA fragment;

6. Hybridisation of a di-base probe (spanning two nucleotides) to the cDNA fragment;

7. Ligation of the di-base probes to the universal primer;

8. Measurement of fluorescent dye attached to the probe;

9. The fluorophore is cleaved from the probe and washed away;

10. Steps 6-9 are repeated a total of 10 times;

11. Denature the read from the cDNA;

12. Five different reads are generated by offsetting the primer by one nucleotide per read (original diagram).

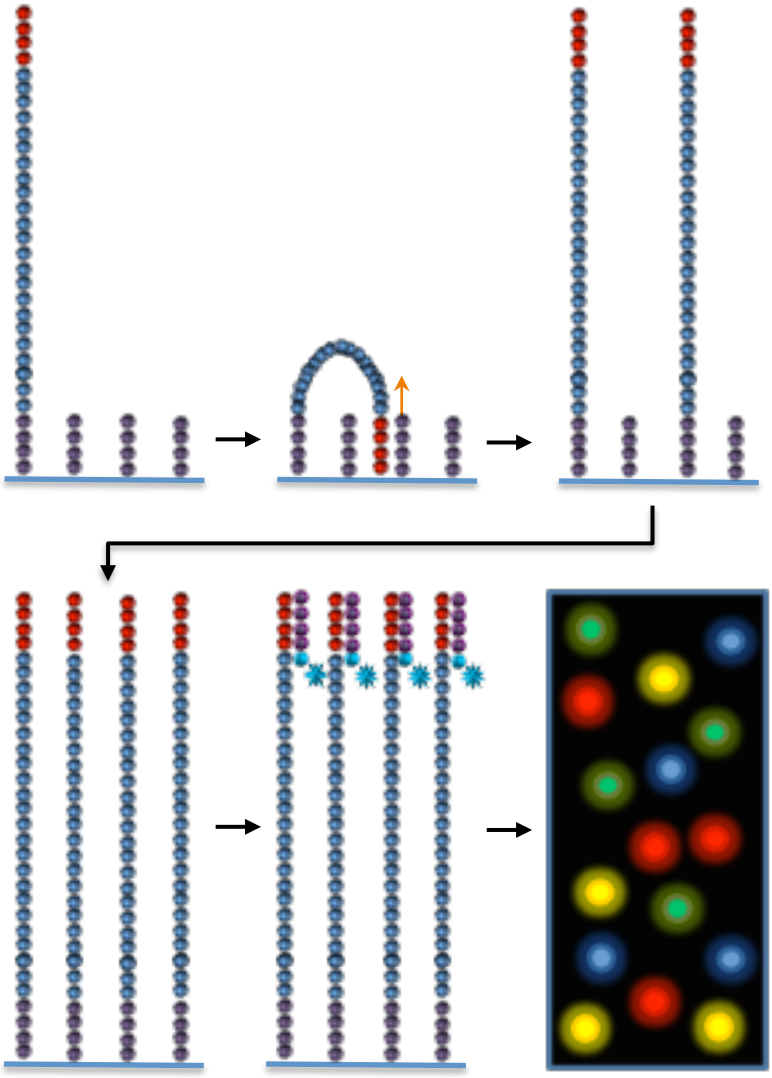


Figure 2.2: Illumina sequencing system

Following the fragmentation of RNA, cDNA is produced, size selected and attached to adaptors. The cDNA library is bound to a sequencing chip. 'Clonal clusters' are then produced by PCR based

bridge amplification. This process results in many clusters of reads for each cDNA molecule bound to the slide. Images are captured after every base (with a reversible dye terminator) is incorporated into the DNA fragment (original diagram).

2.2 Tissue Culture

Collagenase type I was used to isolated HUVEC from umbilical cords that were obtained (after informed consent) from the Birmingham Women's Hospital (local ethics number: 11T063). HUVEC were cultured *in vitro* using Medium 199 (M199), which as supplemented with 10% (v/v) foetal bovine serum, 4 mM L-glutamine, 200 units/mL of Penicillin-Streptomycin, 90 μg/mL heparin and bovine brain extract, as described by Maciag *et al*. [(55)].

Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 10% (v/v) foetal bovine serum, 4 mM L-glutamine, 200 units/mL of Penicillin-Streptomycin was used to culture the other cell types. Namely, HEK293T (human embryonic kidney 293T cells), human dermal fibroblasts (DF), human aortic smooth muscle cells (HASMC) and keratinocytes (Ker).

Both the supplemented DMEM and M199 were filtered using 0.22 μM pore bottle top filters, warmed to 37°C before use and stored at 4°C. All the cells were plated onto sterile 10 cm plates (pre-coated with 0.1% w/v sterile gelatin in PBS for HUVEC) and maintained at 37°C with 5% CO_2 in a humidified incubator. Once the cells reached confluence they were passaged at a 1:3 ratio (1:5 for the HEK293T). To split the cells they were incubated with 3 mL-1% (v/v) trypsin-EDTA after being washed with sterile PBS. Once detached the cells were centrifuged at

room temperature and 195 g for 5 minutes. The pelleted cells were resuspended in the appropriate media and divided equally between the new plates.

2.3 RNA isolation and cDNA production

Alan Zhuang provided total RNA from isolated liver tumour (hepatocellular carcinoma) associated endothelium (TLE) and healthy liver associated endothelium (HLE) as described by Mura *et al*. [54]. Total RNA was obtained from HUVEC, dermal fibroblasts (DF), human aortic smooth muscle cells (HASMC) and keratinocytes (note, a low RNA yield prevented the analysis of all the genes in keratinocyte by qPCR) using a Qiagen RNeasy mini kit according to the manufacturer's instructions. It is important to note that the optional DNase I digestion was also performed (qPCR confirmation of the efficiency of this step can be found in Appendix 4). The quality and quantity of the RNA was confirmed using a NanoDrop Spectrophotometer and Steve Kissane obtained a RIN (RNA integrity number) using a RNA Bioanalyser.

The total RNA from all the cells was used to produce cDNA for the gene expression analyses. The cDNA reaction was performed as specified by the manufacturer's instructions for the High Capacity cDNA Archive kit with random primers. Each cDNA reaction contained 500 ng of RNA per 20 µL reaction mix. Following the production of cDNA a serial dilution was conducted to prepare the cDNA for the qPCR reactions (1 in 20, 1 in 40, 1 in 80, 1 in 160 and 1 in 320). However traditional PCR was conducted using undiluted cDNA.

2.4 Universal Probe System (qPCR)

Primers were designed using the web based Universal Probe Library (UPL) Assay Design Centre (primer sequences can be found in Appendix 5). Where possible all the amplicons span exon boundaries (snoRNA do not contain introns). Each universal probe qPCR (Figure 2.3) reaction mix contained: 12.5 µL 2x universal qPCR mix, 1 µL of each primer, 0.25 µL of the appropriate probe and 0.25 µL H_2O. This mix was combined with 10 µL of cDNA (concentration as specified in Section 2.3) in a different location to reduce the chances of contamination. The reactions were conducted on a Rotor-Gene RG-3000 qPCR machine using the following programme:

Denaturation: 5 minutes at 95°C

Denaturation: 15 seconds at 95°C

Annealing and elongation: 45 seconds at 60°C } 40 cycles

Standard curves were generated using the serial dilutions of cDNA that facilitated the determination of the threshold position, which in turn allowed for the cycle threshold (CT) to be determined for each test cDNA (Appendix 6). The relative abundance of gene expression was then determined by comparing the CT value to the CT value of a housekeeping gene. FLOT2 (Flotillin 2) for inter-cell-type comparisons and ACTB (Beta-actin) for same cell type comparisons.

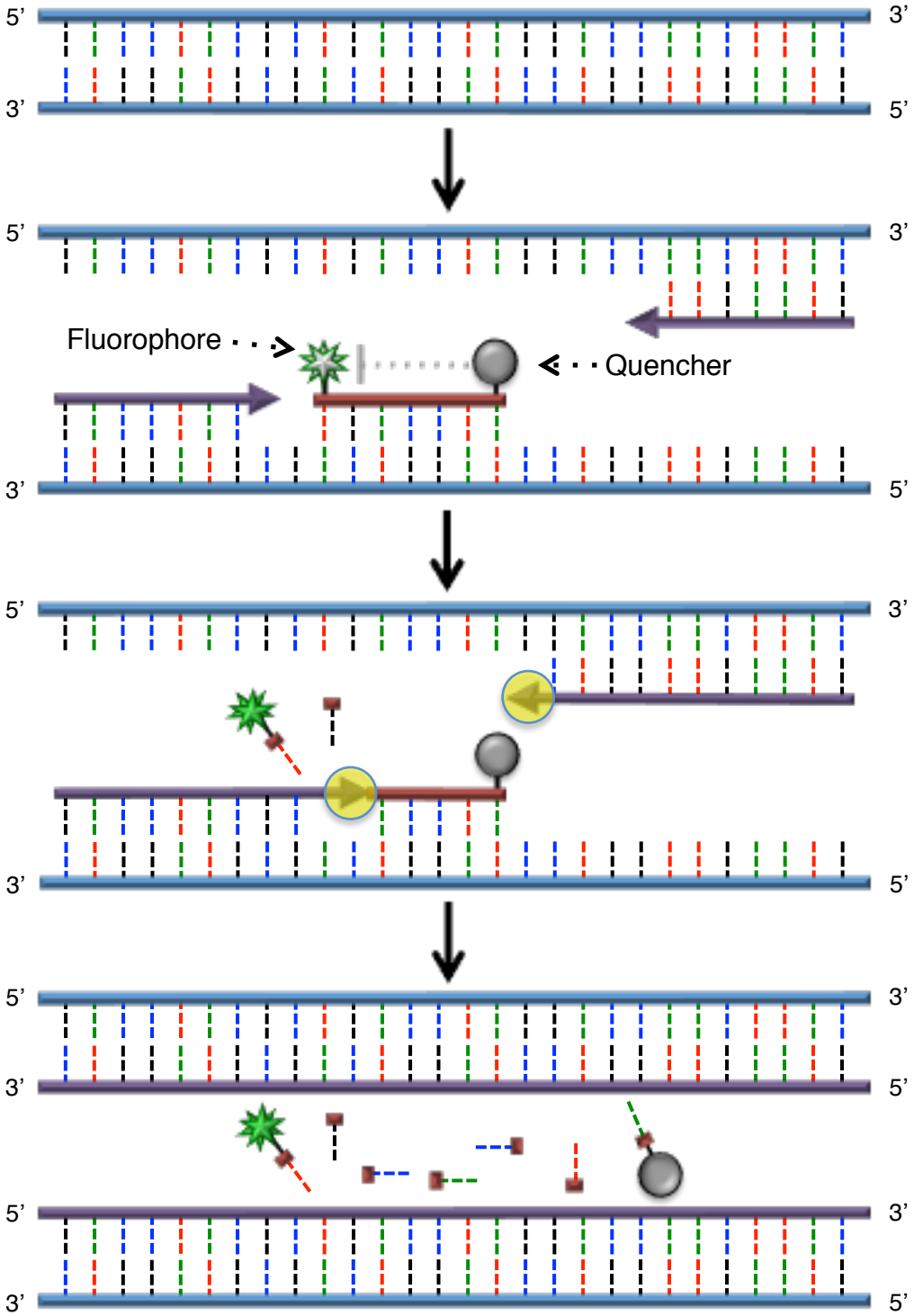


Figure 2.3 Universal probe system of qPCR

The UPL system of qPCR relies upon small probes (8-9 nucleotides) that bind to multiple areas within the genome. Much like the TaqMan qPCR system, an adjacent quencher represses fluorophore excitation. Therefore only probes within an active amplicon will contribute to the

fluorescent signal in a reaction, this occurs when a DNA polymerase destroys the probe when extending a primer, thereby separating the fluorophore from the quencher (original diagram).

2.5 SYBR Green (qPCR)

SYBR green (an asymmetrical cyanine dye) (Figure 2.4) was the method used in instances where a universal probe-binding site was not present within the target transcript (SNORD75, SNORD76 and SNORA81). SYBR green is less specific than the UPL system because of the potential for production of off target amplicons, which are fully capable of generating a fluorescent signal (unlike the UPL system, where the probe must be inside the amplicon to generate a signal). However an amplicon melt curve can be produced if a melt step is included in the qPCR programme, the presence of only one peak indicates that the reaction was specific. Each reaction mix contained: 12.5 µL 2x SYBR Green PCR mix, 0.5 µL 10 µM primers, 2 µL H_2O and 10 µL of cDNA. The following run settings were used:

Denaturation: 5 minutes at 95°C

Denaturation: 15 seconds at 95°C

Annealing and elongation: 45 seconds at 60°C } 40 cycles

Melt: 1°C incremental increases from 55 to 99°C every 5 seconds, with the exception of the first step that lasted 1 minute.

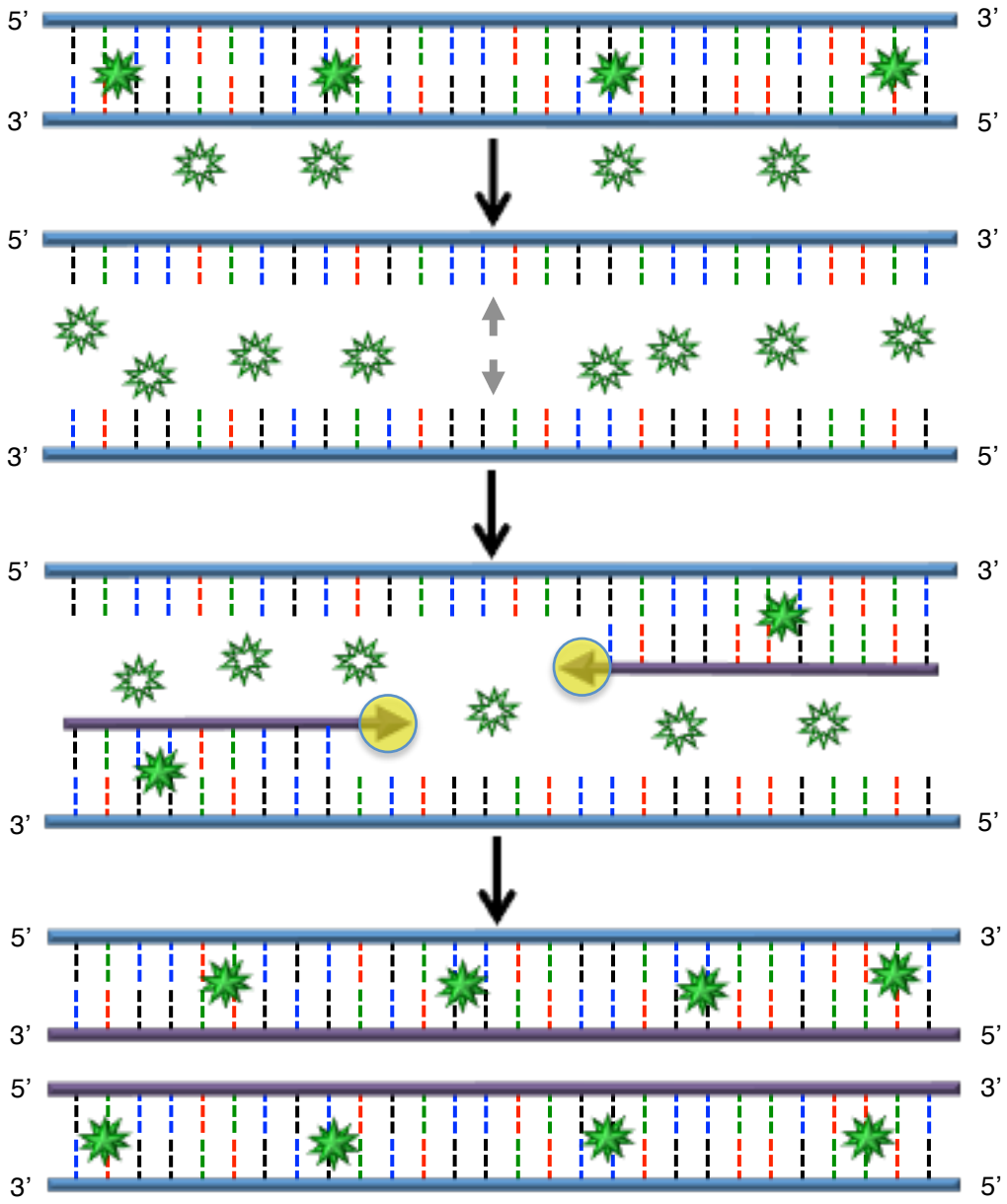


Figure 2.4: SYBR Green qPCR

SYBR green fluoresces strongly only when bound within DNA and specifically binds to double stranded DNA by intercalating between nucleotides. Therefore a fluorescent signal is directly proportional to the amount of double stranded DNA in a reaction. The fluorescent signal will be greatest during the annealing or the extension steps of qPCR and will increase as more double stranded DNA is produced by the PCR reaction (original diagram)

2.6 Molecular Cloning

PCAT19 was amplified from HUVEC cDNA (as prepared in Section 2.3) using PCR in five blocks (Figure 2.5) using a high fidelity (HF) DNA polymerase. A five block system was employed because PCAT19 contains many viral repeat regions, which inhibits the amplification of full length PCAT19 (note: even when PCAT19 was inside the plasmid, it was not possible to amplify the full length molecule out of the plasmid). 100 µL mixes were set up for each of the five blocks, containing: 2 µL 5x Phusion HF buffer, 2 µL 10 mM dNTPs (deoxynucleotides), 2 µL of forward primer, 2 µL reverse primer, 68 µL H_2O and 2 µL of HUVEC cDNA (1380ng/µL). The PCR was performed as follows:

Denaturation: 30 seconds at 98°C

Denaturation: 10 seconds at 98°C
Annealing: 30 seconds at 55°C
Elongation: 1 minute at 72°C
} 35 cycles

Elongation: 10 minutes at 72°C

It is important to note however that the parameters of the PCR reaction had to be changed for some of the blocks to improve the reaction efficiency. A 56°C annealing step was required for block one and three, and block two requires 2% DMSO (dimethyl sulfoxide). Once the blocks were amplified they were run on a 1.5% agarose gel with SYBR Safe DNA Gel stain. Finally, the appropriate band was dissected and was purified using a Fermentas GeneJET gel extraction kit according to the manufacturer's instructions.

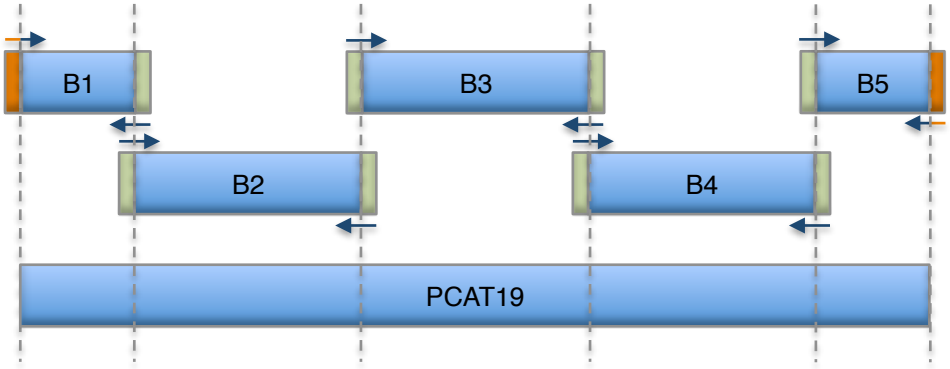


Figure 2.5: PCR amplification of PCAT19

Each of the blocks overlapped the adjacent block by 50bp (base pairs), with the exception of the 5' end of block 1 (B1) and the three-prime (3') end of block 5 (B5). The aforementioned blocks were amplified using primers that contained sequences from the lentivirus plasmid pWPI (15bp overlap) (original diagram).

2.7 Plasmid Preparation

All the plasmids required in this thesis were purified according the manufacturer's instructions from cultures of E. *coli* using a Qiagen plasmid maxi kit or a GeneJET plasmid miniprep kit for either a large or small scale, respectively. Glycerol stocks were made from the bacterial cultures for long term storage, 500 µL of the positive cultures were mixed with 500 µL of sterile 30% (v/v) glycerol and stored at -80°C.

In preparation for the Gibson assembly reaction (Section 2.8) the plasmid pWPI was linearized using the following mix: 5 µg of pWPI plasmid, 5 µL 10x New England BioLabs (NEB) restriction enzyme buffer 4, 2.5 µL PmeI and made up to 50 µL with H_2O. The digestion mix was incubated at 37°C for 1 hour. The linearized plasmids were size selected from an agarose gel (linear plasmids run

slower than in their circular supercoiled form) using a GeneJET gel extraction kit according to the manufacturer's instructions.

2.8 Gibson Assembly

To insert PCAT19 into pWPI (Figure 2.6) a Gibson assembly cloning kit was used (Figure 2.7), containing: 10 µL of PCR fragments and vector (20 ng of B1 and B5, 40 ng of B2, B3 and B4 and 37.5 ng of linearized pWPI) in water, 10 µL 2x Gibson assembly master mix. The assembly mix was incubated for 50°C for 1 hour before being placed on ice in preparation for the transformation. The chilled product was added to a vial of NEB 5-alpha competent E. *coli* cells and mixed by flicking the tube 4 times and incubated on ice for 30 minutes. The bacteria were then heat shocked at 42°C for 30 seconds and transferred to ice for 2 minutes. 950 µL of room temperature Super Optimal broth with Catabolite repression (SOC) media was then added to the bacteria and incubated for 60 minutes at 37°C with vigorous shaking at 250 rpm. 100 µL of mixture was plated onto pre-warmed Luria broth agar plates, containing 50 µg/mL ampicillin. The plasmid preparation of positive colonies was conducted as described in Section 2.7. Sanger sequencing was conducted to confirm the sequence of PCAT19 from the selected colonies by the Functional Genomics, Proteomics and Metabolomics Facility (School of Biosciences, University of Birmingham, UK).

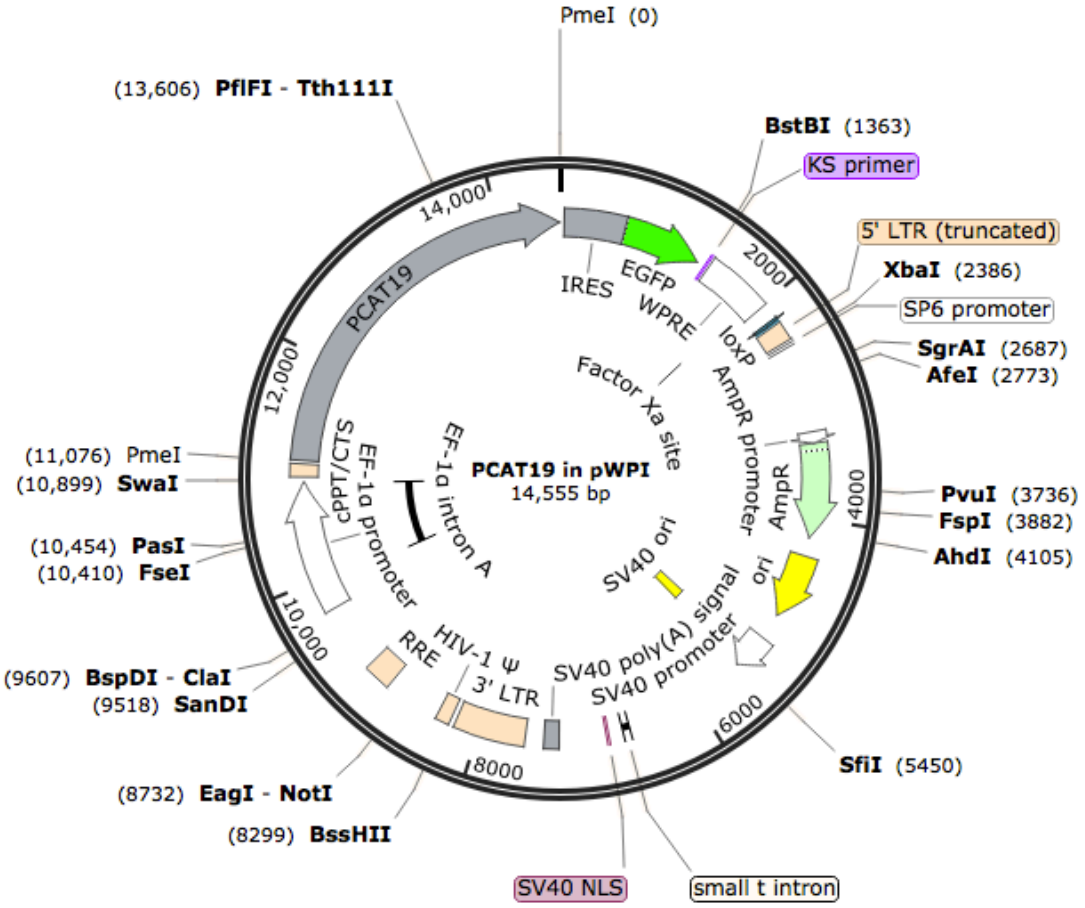


Figure 2.6: A plasmid map of PCAT19 in pWPI

PCAT19 was cloned into pWPI, a 2nd generation bicistronic lentiviral vector. This vector allows enables the simultaneous expression of PCAT19 and EGFP (enhanced green fluorescent protein). The presence of an internal ribosome entry site allows EGFP to be independently translated into protein from the same mRNA molecule as PCAT19. This feature is essential because it allows for the isolation of positively transduced cells, but also because PCAT19 does not contain an open reading frame it is not possible tag the EGFP to a protein (original diagram compiled using SnapGene).

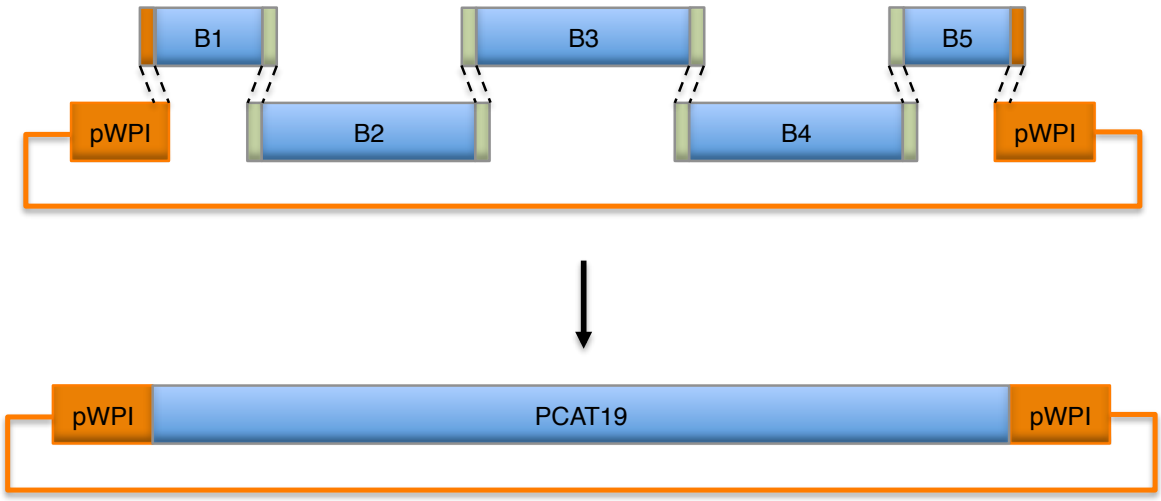


Figure 2.7 Gibson assembly of PCAT19 in pWPI

The five PCAT19 blocks were assembled and inserted into pWPI using Gibson assembly. The reaction is facilitated by 5' exonucleases that chew back the 5' ends of the blocks and linear plasmid, which creates complementary binding sites. Once the DNA fragments annealed a DNA polymerase extends the 3' ends to close any gaps at which point a DNA ligase seals the nicks (original diagram).

2.9 Gene Overexpression

The overexpression of PCAT19 in HUVEC was conducted using lentiviral transduction, which first requires the production of virus in HEK293T cells. 4.4 µg PCAT19 in pWPI (empty pWPI was used to generate lentivirus for the negative control), 3.3 µg psPAX2 and 1.3 µg pMD2G (9 µg plasmid DNA) was added to 1 mL opti-MEM, to which 36 µL 1 mg/mL PEI (polyethylenimine) was added. The mixture was vortexed gently and incubated at room temperature for 10 minutes. The PEI/plasmid mixture was added to a 10 cm plate containing $3x10^6$ HEK293T cells in supplemented DMEM (as described in Section 2.2) and returned to the incubator (37°C and 5% CO_2) for 24 hours. The media (containing the resulting

virus) was then collected from the HEK293T cells and centrifuged at 195 g for 5 minutes, supplemented with 8 µg/ml polybrene, 90 µg/ml heparin and bovine brain extract. The media was filtered with a 0.45 µm pore syringe filter before being added to a 10 cm plate containing $1x10^6$ HUVEC. After 24 hours the media was replaced with fresh M199 and cultured as previously described in Section 2.2. One week later the successfully transduced EGFP positive cells were isolated using fluorescence-activated cell sorting by the University of Birmingham Flow Cytometry Facility.

2.10 Acumen cell cycle analysis

Eight wells (on a 96 well plate) per condition were seeded with 4000 HUVEC and allowed to settle for 24 hours. The media was then gently removed and replaced with 100 µL ice-cold 85% ethanol and incubated for 30 minutes at room temperature. The ethanol was gently removed and replaced with a 100 µ solution containing: 20 µL 10% triton X-100, 20 µL 10 mg/mL RNase A, 20 µL 1 mg/mL propidium iodide and 1940 µL 1x PBS. The plate was then covered with foil and incubated at 37°C for 20 minutes before being analysed by Ivette Hernandez-Negrete using an Acumen Explorer TTP Lab Tech.

2.11 siRNA transfection

To knockdown PCAT19 in HUVEC, transfections were performed using siRNA (small interfering RNA). The siRNA were designed according to the guidelines

published by Reynolds et al. (56). To achieve this effect $1x10^6$ HUVEC were seeded onto a 10 cm plate and left to settle for 24 hours. Two independent duplexes for PCAT19 and a negative control duplex (NCD) were used at a final concentration of 30 nM in 680 µL Opti-MEM per condition. The duplexes were left to incubate at room temperature for 10 minutes before being mixed with 108 µL (800 µL total) Opti-MEM containing 0.3% (v/v) RNAiMAX lipofectamine (pre-incubated for 10 minutes at room temperature) and incubated at room temperature for an additional 10 minutes. During this incubation the media was removed from the HUVEC, which were subsequently washed with PBS twice and replaced with 3.2 mL Opti-MEM to which the transfection mix was added. The HUVEC were incubated with this transfection mix for four hours in an incubator at 37°C and 5% CO_2. Upon completion of this incubation the transfection mix was replaced with supplemented M199 (without antibiotics) and returned to the incubator for 48 hours, at which point the HUVEC were lysed and the RNA was extracted as described in Section 2.3.

2.12 Two Colour Microarray analysis

The RNA obtained from the HUVEC in Section 2.11 was used to produce two colour microarray data (Figure 2.8) by fluorescently labelling cDNA using a Quick Amp Labeling Kit (two-Colour) as per the manufacturer's instructions. The cDNA was hybridized onto an Agilent human genome microarray chip (4x44K) and scanned by Steve Kissane at the University of Birmingham Genomics and Microarray Facility. The data obtained from this experiment was analysed using

Bioconductor package Limma. The codes used to analyse the data can be found in Appendix 7.

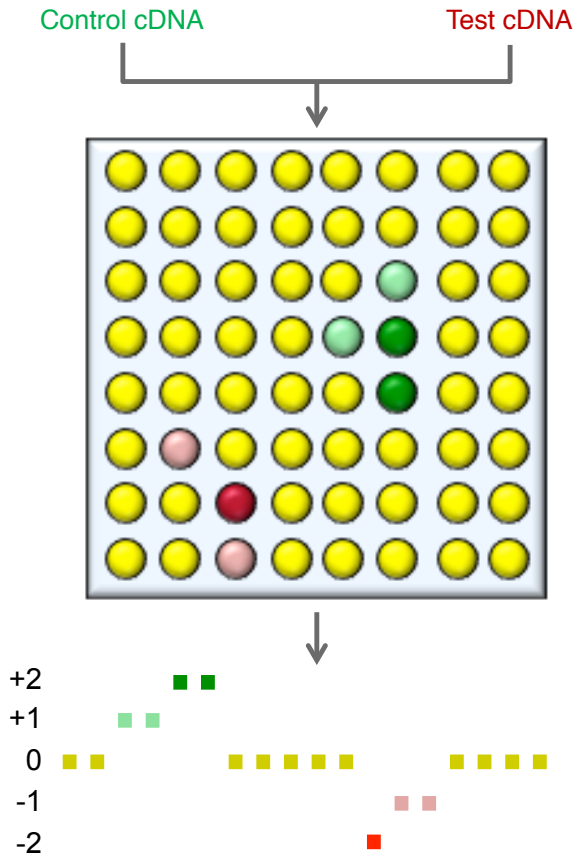


Figure 2.8 Determination of gene expression using two-colour microarray

Two colour microarrays yields data in the form of relative fluorescent intensity to a reference per probe, i.e. the ratio of green fluorescence to red fluorescence. By comparing this ratio between different conditions (tested on another section of the same chip or on an entirely different chip) the differential expression a specific gene can be determined, providing that probes for the gene is present within the microarray chip (original diagram).

2.13 Statistical Analysis

All the wet lab experiments (as opposed to dry lab/bioinformatic) contained within this thesis were performed in triplicate unless otherwise stated and confidence limits present on graphs portray the standard error of the mean (SEM). The

statistical significance of the data was determined using the Kruskal–Wallis one-way analysis of variance ("*" was used to indicate that $p = < 0.05$), this test was used because it does not assume a normal distribution of data.

The experiments requiring differential expression analysis utilised internal statistical analyses by the bioinformatics software. In the case of the RNAseq data, the significance (p value) of the differentially expressed transcripts was determined by Cuffdiff (a subprogram in Cufflinks from the Tuxedo pipeline) using Jensen-Shannon divergence, a linear statistical model that measures the similarity between read abundances for each transcript per condition. For the microarray data, the significance (p value) was determined using the Limma package, through a linear models and empirical Bayes methods and adjusted using the Benjamini-Hochberg method.

Chapter 3: RNAseq analysis of the colon vasculature

3.1 Introduction

This Chapter explores the use of RNAseq to identify non-protein coding RNAs in the vasculature of colorectal cancers. Currently the surgical resection of tumour is the 'go to' therapeutic option [57]. However, only 80% of resections demonstrate a histologically clear margin and nevertheless 50% of those patients will relapse due to the presence of micro-metastases. It is for this reason that chemotherapy and other therapeutic agents have been given as adjuvant treatments for over 50 years. But metastatic colorectal cancer still carries a poor prognosis [58]. Furthermore, insights into the growth of colorectal tumours could prove to be of importance to the health of society.

The acquisition of a dedicated vasculature is imperative for cancer progression. All solid tumours require a blood supply to exceed two millimetres in diameter and metastasize [13]. Moreover colorectal carcinomas are notable for their tendency to grow slowly and have a seemingly angiogenesis-dependent and progressive pattern of growth [8, 59]. The identification of biologically relevant ncRNAs in the vasculature of colorectal cancer could yield novel therapeutically targetable pathways, or even act as biomarkers to guide treatment.

The SOLiD platform was used to generate data, largely due to the availability of the platform 'in house'. SOLiD data is generated and analysed in 'colourspace' (Figure 3.1), rather than 'basespace' (Figure 3.2) where the reads are comprised of nucleotide sequences [52, 60, 61].

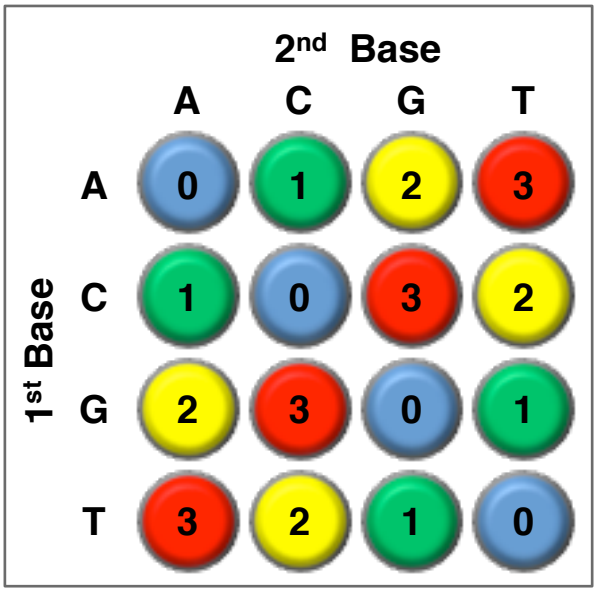


Figure 3.1: Decoding colourspace data.

SOLiD relies on a two base pair probe system yet uses a four colour system to detect the 16 combinations of bases. Each probe can bind and code for four different combinations of two nucleotides. Therefore multiple reads must be generated to determine which nucleotides are present in the sequence. Due to the use of colourspace, SOLiD platforms boast the lowest error rate (≤0.1) of all the NGS (next generation sequencing) platforms [52] (original diagram).

```
A-> @FCH8M2JADXX:1:1101:1330:2077#/1
B-> CNATGAGGCAACCAGCCAGAACGCCTGAACGCAGGCACATACTTCCTATT
    +
C-> aBP`cccefgggfhhiihihhhhifd_cfXcgdhafXae`egghh`f]eg
```

Figure 3.2: Reading basespace data and the FASTQ format

Basespace is used to store and analyse data from most NGS platforms (including Illumina) and refers to data that's stored in a text-based format. One of the most common formats is FASTQ, which stores an identification code (A) for each read and a nucleotide sequence (B) with its corresponding quality score (C) (derived during base calling, which is described in Figure 2.2). The quality scores are recorded using the 'American Standard Code for Information Interchange' (from the lowest quality to highest: !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUV

WXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~), where each symbol refers to the p value associated with the corresponding base in the nucleotide sequence (original diagram).

3.2 Results

The bioinformatic analysis of the SOLiD data from healthy and colorectal carcinoma associated endothelial cells, enabled the generation of a list of differentially expressed transcripts. Of the differentially expressed transcripts identified five were chosen to be validated using qPCR (Table 3.1). Two of the transcripts are antisense molecules to genes with well-defined functions. The other three genes are novel transcripts (these transcripts were assigned names and their sequences can be found in Appendix 8), which in the case of TECA1 (theoretically endothelial cell associated) and TECA3, are derived from genomic locations not yet known to be transcribed. Whereas TECA2 is a novel mitochondrial DNA 'readthrough' (conjoined) transcript that incorporates many smaller known transcripts.

Gene	TE	HE	log2 (FC)	Nucleotide
TECA1	78.3	0.4	6.5	XLOC_029144
HNF1A-AS1	112.0	26.0	2.1	NR_024345.1
TECA2	55639.9	35534.9	0.6	XLOC_032009
TP73-AS1	26.0	162.0	-2.6	NR_033708.1
TECA3	7.8	85.1	-3.4	XLOC_004164

Table 3.1: Tumour and healthy colon endothelium SOLiD RNAseq data

The differential expression was determined using DEGseq. The data represents the fold change of 'raw read counts'. Three transcripts (TECA1, HNK1A-AS1 and TECA2) were expressed at greater

levels in the TE (tumour endothelium), whereas two transcripts (TP73-AS1 and TECA3) were expressed at greater levels in HE (healthy endothelium)

TECA1 was highly expressed in HUVEC (Figure 3.3), which suggests that it could have a potential function in endothelial cells. However, there was no difference in expression between the TLE and HLE, which does not match the SOLiD4 differential expression data. HNF1A-AS1 (hepatic nuclear factor 1 alpha-antisense 1) was observed to be expressed higher in the SOLiD4 data, however qPCR showed the opposite, it was expressed lower in TLE compared to HLE (Figure 3.4). Likewise TECA2 displayed a similar pattern, the qPCR results (Figure 3.5) demonstrated lower expression in the TLE compared to HLE, which opposed the dynamic predicted from the NGS (next generation sequencing) data.

Tumour protein 73-antisense 1 (TP73-AS1 or KIAA0495) was expressed to a greater extent in HLE compared to the other cells types (Figure 3.6). The upregulation of TP73-AS1 in HLE compared to TLE matched the expected pattern from the SOLiD data. TP73 was the only transcript that was consistently differentially expressed across the two experiments. TECA3 did not display a significant reduction in expression between TLE and HLE (Figure 3.7), which did not mimic the predicted downregulation observed in the NGS data.

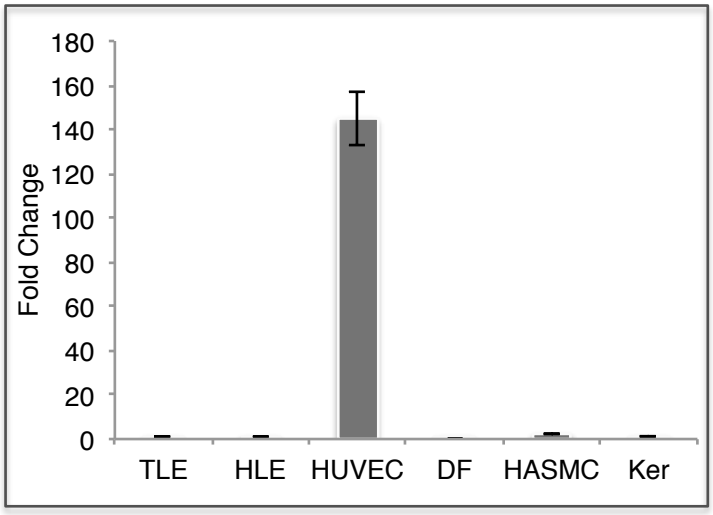


Figure 3.3: The expression of TECA1 compared to FLOT2

TECA1 was mostly highly expressed in HUVEC and 145 fold greater when compared to TLE (mean ±SEM).

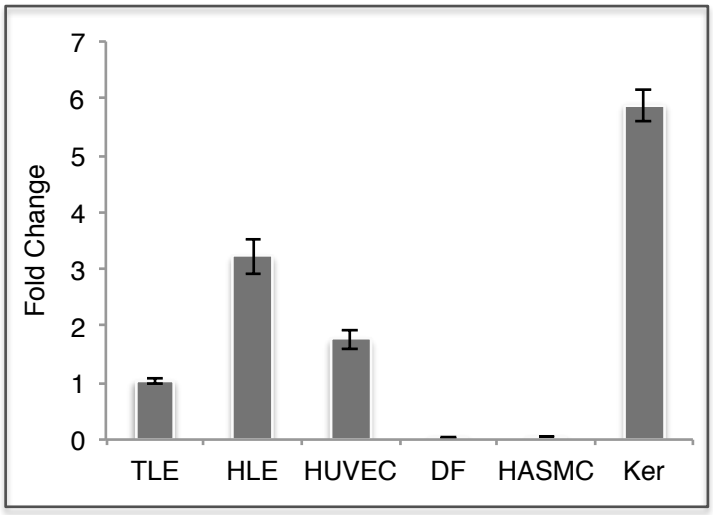


Figure 3.4: The expression of HNF1A-AS1 compared to FLOT2

HNF1A-AS was mostly highly expressed in keratinocytes and was expressed to a lesser extent in TLE compared to HLE (mean ±SEM).

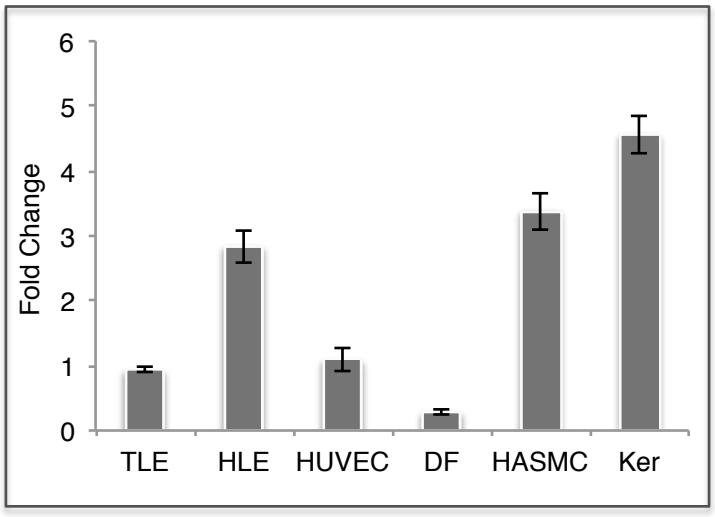


Figure 3.5: The expression of TECA2 compared to FLOT2

TECA2 demonstrated higher expression in HLE compared to TLE and HUVEC, however the expression level in TLE was similar to HASMCs and keratinocytes (mean ±SEM).

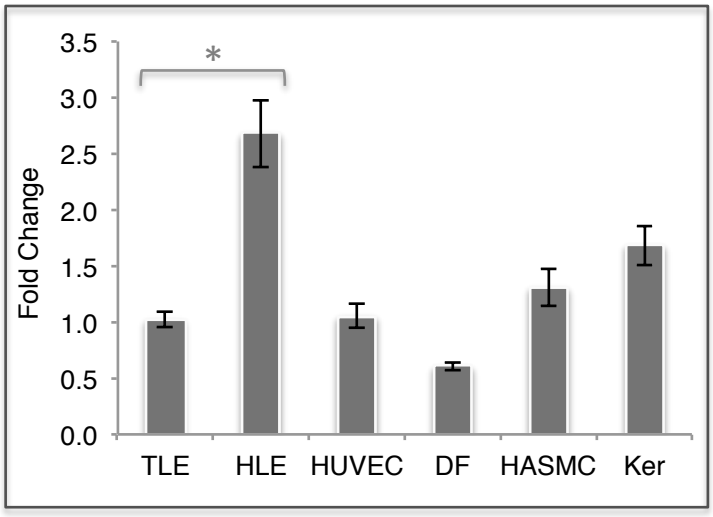


Figure 3.6: The expression of TP73-AS1 compared to FLOT2

TP73-AS1 was significantly ($p = < 0.05$) differentially expressed in between the TLE and HLE. Moreover the HLE demonstrated the highest expression levels out of the six cell types (mean ±SEM).

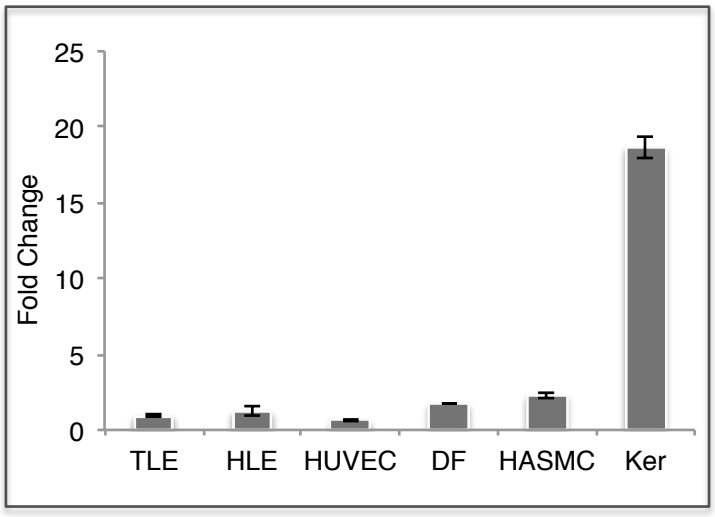


Figure 3.7: The expression of TECA3 compared to FLOT2

Of the six cell types profiled with qPCR, the keratinocytes had by far the highest expression of TECA3 (mean ±SEM).

3.3 Discussion

The use of raw read counts to display the data introduces some weaknesses to this dataset, as the counts are not normalised for the length of the transcript. This feature accounts for the massive read values for TECA2 compared to the other transcripts, as it is by far the longest transcript. Furthermore the transcript is produced by the mitochondria, which are quite abundant within cells. Notwithstanding these feature the fold change should still be accurate, yet the fold change between the TLE and HLE in the SOLiD4 and qPCR data did not match. A possible reason for this inconsistency could be due to the nature of the transcript, it is a read through RNA. Other transcripts produced in the same region could have biased the qPCR, however primers were chosen in an area not previously known to be expressed specifically for this reason. Another explanation is that

TECA2 could be an artefact produced by the mapping step of the NGS analysis, where multiple overlapping transcripts were mapped together into a single transcript incorrectly. SOLiD data is particularly susceptible to these errors because the data is generated in the form of very short reads (fewer than 50 base pairs).

Furthermore TECA2 was not the only transcript to display inconstancies between the qPCR and NGS data. TECA1 and HNF1A-AS1 were also expressed at lower levels in the qPCR data and TECA3 showed very little change. These differences could be caused by the intrinsic differences between the endothelial cells within the colon and the liver. It is possible that qPCR would show the predicted expression pattern (from the SOLiD RNAseq data) if material from colon associated endothelial cells were used. Unfortunately no specimens were available at the time. As it stands TECA1 is probably the only transcript out of the four that warrants further investigation because the increased expression of TECA1 in HUVEC compared to the other cells hints at a potential function in angiogenesis.

However, whilst it would be worthwhile conducting qPCR on endothelial cells from colorectal carcinoma and patient matched healthy colon, the qPCR might still be inconsistent with the SOLiD4 data due to the lack of technical controls in the RNAseq experiment. It is also difficult to control for biological variation when using modern transcriptomic techniques to investigate the expression patterns of cells as endothelial cells due to the need to isolate them from the surrounding tissue. Especially considering that endothelial cells only

represent a small proportion of cells. Therefore the laborious nature of the isolation and size of the resected tissue is a severe limitation. Large samples are required to obtain relatively small amounts of RNA [8]. In this vein the SOLiD4 experiment conducted by previous members of the Bicknell group did not control for variation between individuals. Whilst this is understandable, it does make the task of producing an accurate list of differentially expressed genes problematic, especially when combined with the lack of technical control (the two samples were patient matched, but were sequenced independently).

In spite of its shortcomings this dataset has yielded a very interesting result, namely the expression pattern of TP73-AS1. TP73-AS1 has been observed to be supress oligodendroglial tumours and to induce Cisplatin (a chemotherapeutic drug) resistance in glioma cells after siRNA knockdown [62]. This pattern lends support to the tumour suppressor-like expression pattern in colorectal and hepatocellular carcinoma-associated endothelial cells. It is especially interesting that the healthy endothelial cells displayed the highest expression levels of TP73-AS1 out of all the cell types. Endothelial cells within their “normal” environment are usually quiescent; therefore it is consistent to see increased levels of tumour suppressor genes. Tumour Protein p73 (TP73) is of substantial interest to cancer researchers and has been seen to be mutated and downregulated in many types of cancer [63]. Hence any potential regulation of TP73 by TP73-AS1 in endothelial cells would probably also be of considerable interest and is therefore worth future functional investigation by siRNA knockdown and *in vitro* angiogenesis assays (TP73-AS1 was not investigated further in this project due to time constraints).

Chapter 4: Enrichment of snoRNA in the tumour vasculature

4.1 Introduction

To improve the chances of successfully identifying differentially expressed genes in the tumour vasculature, the SOLiD4 data from Chapter 3 was cross-referenced with a second dataset. The second dataset was generated using Illumina sequencing, which has emerged to become the predominant NGS platform. Illumina sequencing has many advantages over the SOLiD platform. Not only is it technically simpler to conduct but also generates longer reads (> 90bp), which aids in the analysis of the data. This is incredibly important because it allows reads to span repetitive regions, which increases the number of uniquely mapped reads and therefore the genome (or transcriptome) coverage. Additionally, longer reads reduce the likelihood of SNPs (single nucleotide polymorphisms) or sequencing errors influencing read mapping [51, 52].

Furthermore Illumina can be analysed by a greater number of mapping programmes, which allows for improvements based upon the advancements of the algorithms developed by the scientific community. Additionally, SOLiD data must be analysed in colourspace (as opposed to basespace), which requires a completely separate (and far more restrictive) set of tools to map the data to a colourspace reference. If colourspace reads are translated into basespace before being analysed a single sequencing error (or a SNP) can cause changes in the sequence of every nucleotide after it (Figure 4.1) [52, 60, 61].

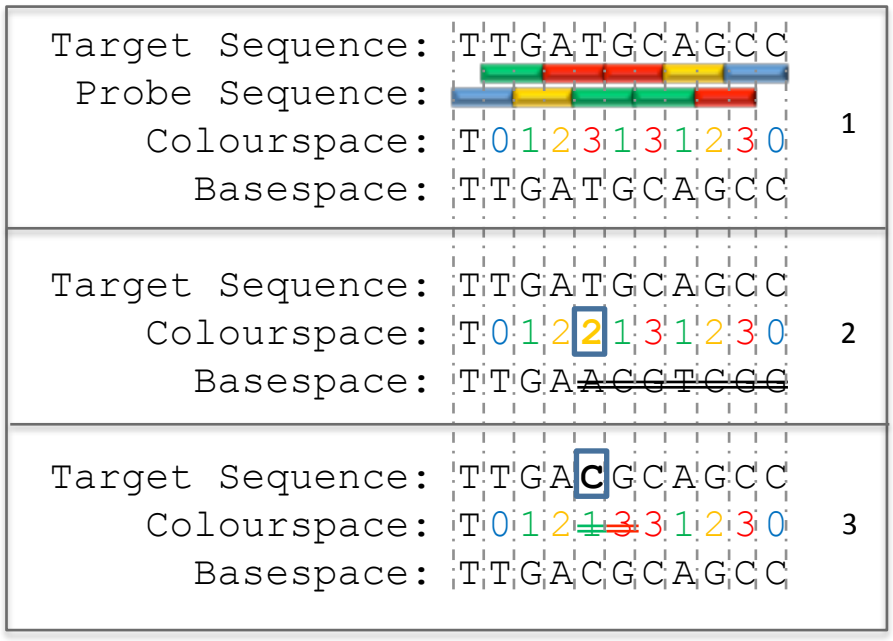


Figure 4.1: The impact of SNPs and errors on colourspace data

1) Normal read sequence - The colourspace sequence is decoded using the last base of the sequencing primer as a starting point. Every nucleotide can then be decoded sequentially using the previous nucleotide and the colour of the di-base pair probe.

2) Reads with sequencing errors - A single error will change not only one base when converted to basespace, but potentially all the nucleotides after the error.

3) Reads containing SNPs - Reads when converted to basespace will have the same sequence as it's target, but a single base pair mutation will cause a change of two numbers in the colourspace sequence (original diagram).

4.2 Background information

The Illumina data in question was derived from the endothelial cells associated with healthy lung tissue and solid lung tumour tissue. Lung cancer is the leading form of cancer mortality in the UK, it accounts for 21% of cancer deaths in women

(rising year on year) and 24% in men [64]. Although the rates vary drastically in different countries, worldwide lung cancer mortality accounts for ~18% of cancer deaths [65]. Lung cancer can be separated into two main categories: small cell or non-small cell lung cancer (NSCLC). NSCLC accounts for over 80% of the lung cancer incidence [66], it was for this reason that NSCLC was used to produce the dataset discussed in this Chapter. There is a persisting need for novel therapeutic agents to be developed as lung cancer often presents in patients at an advanced stage, at which point surgery is often not possible. Moreover, current first line treatment regimes often add just over 2 months to the mean survival rates [67]. Conjointly, the current anti-vascular therapies such as Bevacuzimab (anti-VEGF antibody) have been shown to be ineffective (and sometimes less beneficial) in patients with squamous tumours [68].

For the aforementioned reasons, the identification of differentially expressed ncRNAs between healthy and NSCLC associated endothelial cells could give productive insights into lung cancer pathology. However the experimental design shared some of the weaknesses of the colon data (as described in Chapter 3) and as a consequence of those weaknesses the two datasets were cross-referenced. This process involved the identification of the non-protein coding transcripts present in both datasets. Before determining which ncRNAs were differentially expressed and which transcripts displayed similar expression patterns in both datasets. This analysis yielded an unexpected result, the differentially expressed ncRNAs that were common to both the datasets were classified as snoRNA.

4.3 The function of snoRNA

It has long been thought that the only purpose of introns was to serve as a buffer to protect exons from mutations, frame shift mutations in particular. However functional ncRNAs have been shown in the literature to be excised from some introns [69]. One such type of these molecules are the snoRNAs (Figure 4.2). Three distinct classes of snoRNA have been identified to date. All three are involved with the posttranscriptional modification of RNA by guiding the functionality of distinct protein groups.

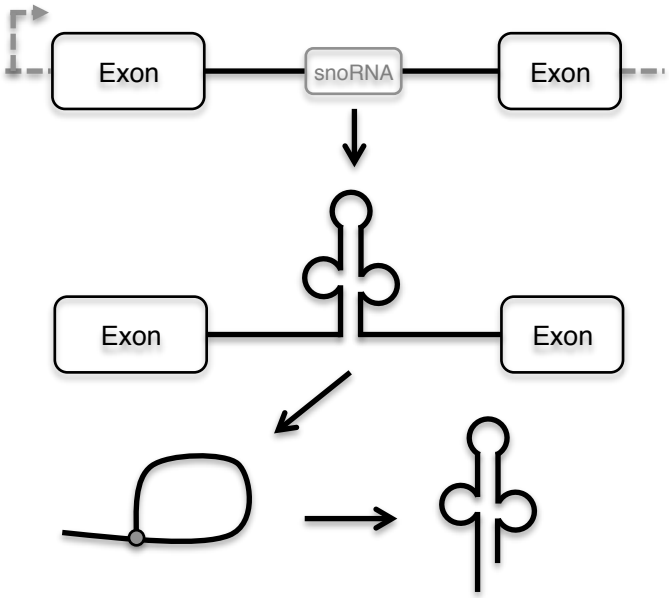


Figure 4.2: Excision of snoRNA from host genes

Following transcription of a gene, introns must be excised from pre-mRNA to produce mature mRNA. Intron lariats are produced as a waste product of splicing exons together and are usually degraded. But if the intron contains secondary structures and/or sequence motifs they can be recycled to produce functional ncRNAs such as snoRNAs. Any genes that contain such RNA products are referred to as host genes (original diagram).

Box C/D snoRNAs (SNORD) are named because of the presence of C box (UGAUGA) and D box (CUGA) motifs that facilitate the binding to proteins that guide the 2'-O-ribose methylation of RNA [70]. 2'-O-ribose methylation has been found to occur in functionally essential areas of the ribosome and splicesome, and prevents the editing of adenosine to inosine. The second group are the H/ACA box snoRNA (SNORA), which are defined by the presence of the H box (consensus ANANNA) and the ACA box (ACA). SNORA bind to a group of proteins and act as guides for pseudouridylation, which is the process of converting uridine to the 'universal base' pseudouridine (Ψ/psi). The third member of the snoRNAs was only identified as a novel class of snoRNA in 2002, the small cajal body RNA (SCARNA) are capable of both pseudouridylation and 2'-O-methylation [70]. For a review see refs. [69-72].

The snoRNA are, as their name suggests, restricted to the nucleus and SCARNA are further restricted to the cajal body, a subnuclear organelle. Of all the ncRNAs, snoRNA were among the earliest identified. It is perhaps because of their association of with the modification of ribosomal RNA that caused the functions of snoRNA to be seen as a relatively unglamorous area of research [72]. However in recent years snoRNA have been subjected to a renewed interested because of the discovery that they are associated with alternative splicing, the formation of miRNA and an ever-expanding role in human disease. In point of fact, snoRNAs have recently been identified as being of consequence in the etiology of a diverse variety of cancers [72-75]. However snoRNA have yet to be associated with the tumour vasculature.

4.4 Results

Following the analysis of RNAseq data derived from healthy and tumour endothelial cells in the lung and colon, seven snoRNA molecules were identified at greater levels in tumour endothelial cells (Table 4.1). The seven differentially expressed snoRNAs were comprised of one SNORA, one SCARNA and five SNORDs. The expression of the snoRNAs were validated and compared to the expression of their host genes using qPCR. Of the seven, SNORD75 (Figure 4.3), SNORD76 (Figures 4.4 and 4.5), SCARNA7 (Figure 4.6) and SNORA81 (Figure 4.7) were confirmed using qPCR to be expressed at higher levels in TLE than HLE. SNORD32A (Figure 4.8), SNORD30 (Figure 4.9) and SNORD100 (Figure 4.10) on the other hand did not show the anticipated expression patterns.

		Lung (Illumina HiSeq 2000)			Colon (SOLiD4)		
snoRNA	Host Gene	TE	HE	**FC**	TE	HE	**FC**
SNORD32A	RPL13A	401.7	0.0	**∞**	567.7	26.5	**3.1**
SNORD30	SNHG1	107.7	0.0	**∞**	6053.9	374.5	**2.8**
SNORD76	GAS5	71.6	0.0	**∞**	4183.9	456.1	**2.2**
SCARNA7	KPNA4	67.8	0.0	**∞**	27.5	8.1	**1.2**
SNORD75	GAS5	128.5	3.5	**3.6**	140.2	7.2	**3.0**
SNORD100	RPS12	365.4	80.4	**1.5**	11744.0	3793.6	**1.1**
SNORA81	EIF4A2	465.7	128.3	**1.3**	29.2	9.4	**1.1**

Table 4.1: Differentially expressed snoRNA in the lung and colon vasculature

The RNAseq analysis yielded seven snoRNA that appeared to be differentially expressed in both lung and colorectal cancer (TE) compared to the healthy endothelium (HE).

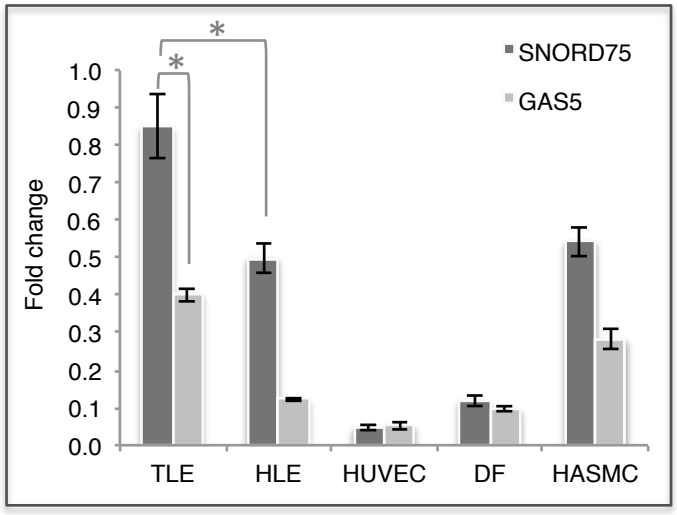


Figure 4.3: SNORD75 and GAS5 expression compared to FLOT2

SNORD75 was most highly expressed in the TLE compared to the other cell types and was expressed at significantly higher levels when compared to the HLE. Furthermore SNORD75 was significantly upregulated compared to its host gene GAS5 (mean ±SEM).

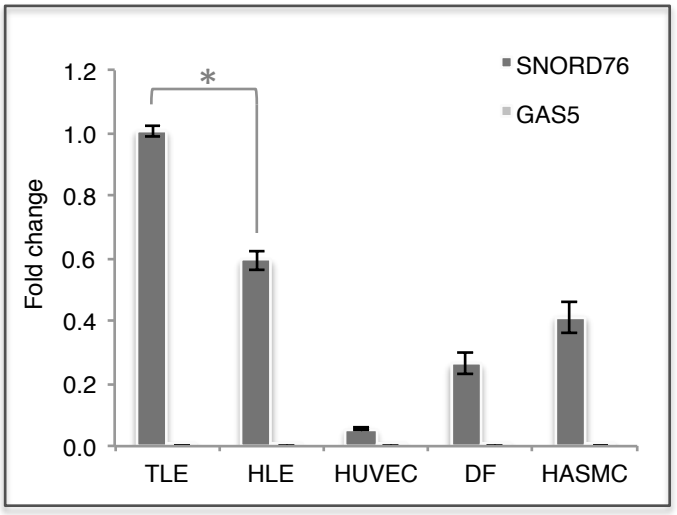


Figure 4.4: SNORD76 and GAS5 expression compared to FLOT2

SNORD76 was most highly expressed in TLE and was expressed at significantly higher levels in the TLE when compared to HLE (mean ±SEM).

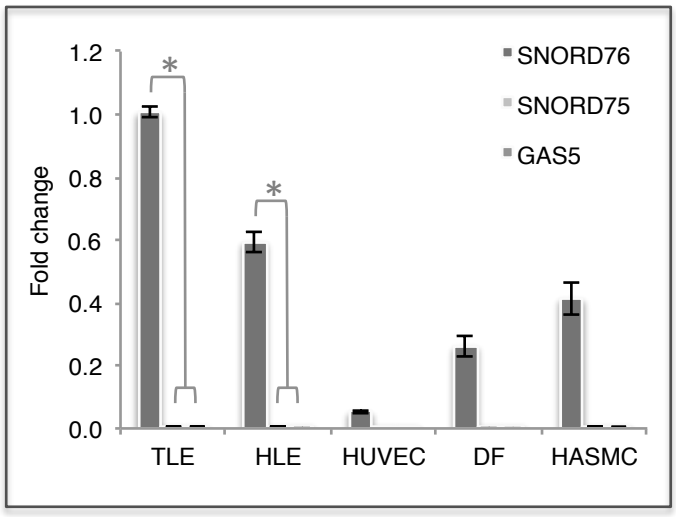


Figure 4.5: the enrichment of SNORD76 from GAS5 compared to SNORD75

The expression of SNORD76 significantly eclipses the expression of not only GAS5, but also SNORD75 (mean ±SEM).

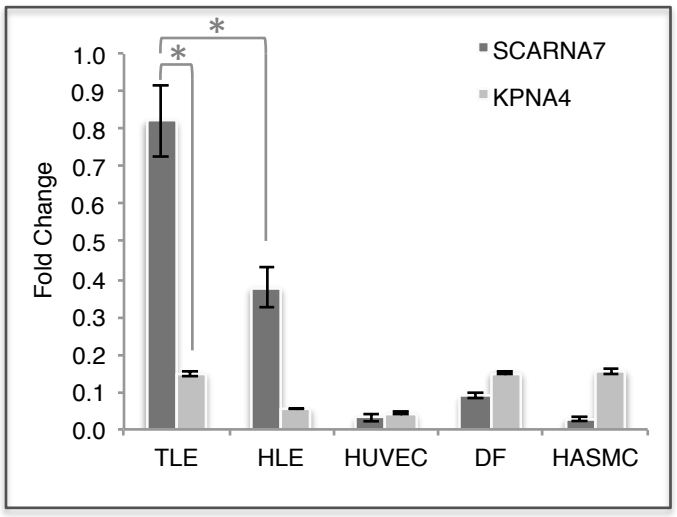


Figure 4.6: SCARNA7 and KPNA4 expression compared to FLOT2

SCARNA7 had a significant doubling of expression in TLE when compared to HLE. Furthermore the expression of SCARNA7 in TLE was many fold higher than the other three cell types (mean ±SEM).

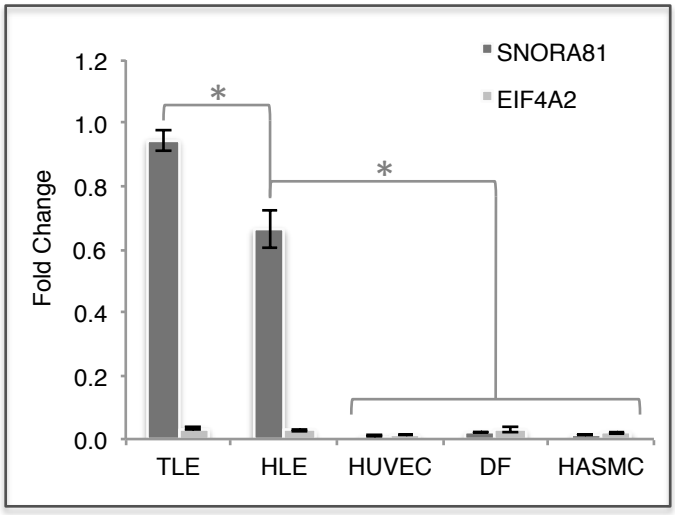


Figure 4.7: SNORA81 and EIF4A2 expression compared to FLOT2

SNORA81 displays possible endothelial-specific patterns of expression in TLE and HLE when compared to DF and HASMC. Furthermore SNORA81 was significantly differentially expressed in TLE compared to HLE (mean ±SEM).

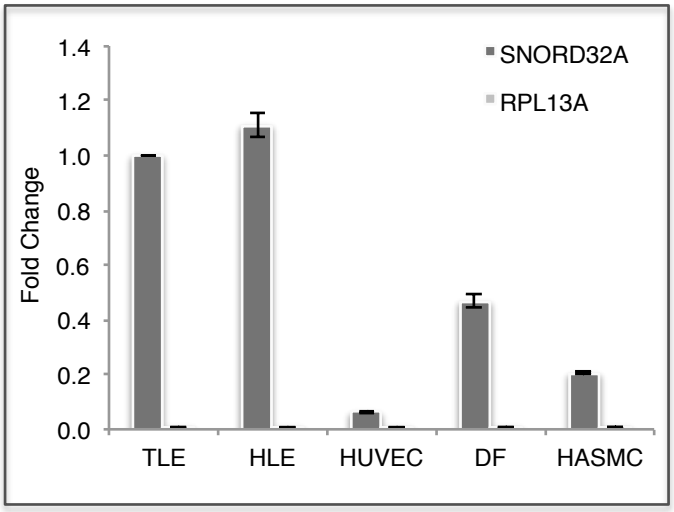


Figure 4.8: SNORD32A and RPL13A expression compared to FLOT2

SNORD32A was most highly expressed in the TLE and HLE compared to the other cell types, however the differential expression between the TLE and HLE was nominal. Furthermore there was only trace expression of the host gene RPL13A when compared to SNORD32A in all the cell types (mean ±SEM).

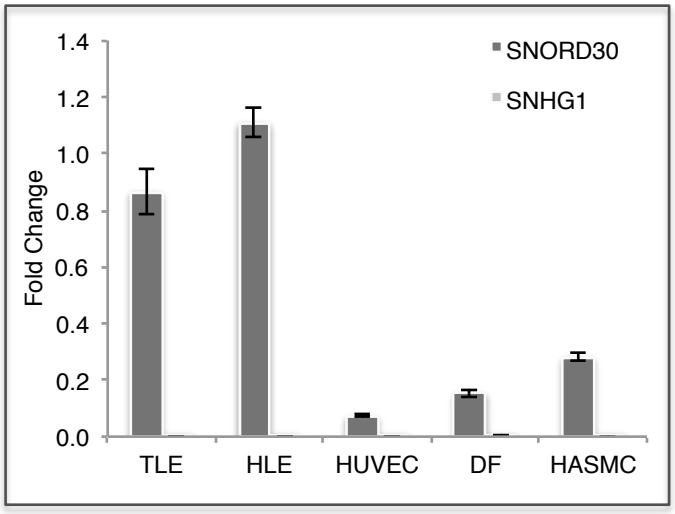


Figure 4.9: SNORD30 and SNHG1 expression compared to FLOT2

The expression of SNORD30 was slightly upregulated in HLE when compared to TLE by qPCR. SNORD30 was most highly expressed in the TLE and HLE compared to the other three cell types. The expression of SNORD30 was slightly upregulated in HLE when compared to TLE. Furthermore there was only trace expression of the host gene SNHG1 compared to SNORD30 in all the cell types (mean ±SEM).

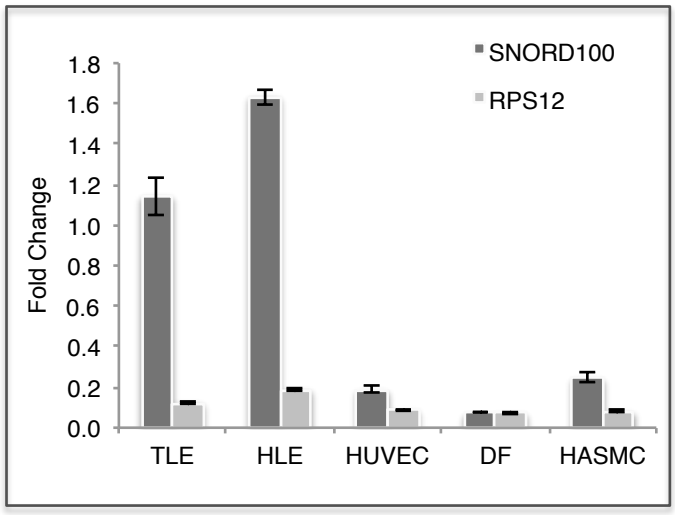


Figure 4.10: SNORD100 and RPS12 expression compared to FLOT2

SNORD100 was expressed at far higher levels in TLE and HLE compared to the other three cell types. However the expression of SNORD100 was lower in TLE than HLE (mean ±SEM).

4.5 Discussion

The results presented in this Chapter (when compared to Chapter 3) demonstrated an improved success rate with regard to identifying transcripts that were expressed at higher levels using both RNAseq and qPCR. The cross referencing of the two RNAseq datasets seems to have reduced the effect of some of the biological and technical variation. Although it is somewhat surprising that the snoRNAs featured so heavily in the differential expression lists following cross-referencing. The upregulation in lung, colon and liver tissues indicate that these snoRNA could potentially be non-protein coding “pan” (all-inclusive) tumour endothelial cell markers. It does, however, indicate that comparing lung to colon endothelial cells has been at the expense of observing a wider range of gene classes that could have been upregulated in the endothelium of each individual tissue.

Endothelial cells have been known to show transcriptional and functional differences between various tissues and organs [46]. A characteristic difference in the lung is the predominance of non-sprouting (intussusceptive) angiogenesis, where endothelial cells proliferate to form a large lumen. The enlarged lumen is then split by the production of transcapillary pillars (columns formed of thickened endothelial cells) [18, 19]. This process is markedly different to sprouting angiogenesis, which colorectal tumours rely upon heavily. These preferential mechanisms of angiogenesis, in conjunction with other factors could cause different gene expression profiles between the lung and colon endothelium. For

example, ROBO4 is upregulated specifically in the endothelial cells of many tissues, but is often absent in colorectal cancer associated endothelial cells [8].

The differences between the liver endothelium used for the qPCR and the tissues analysed by RNAseq is perhaps why SNORD32A, SNORD30 and SNORD100 appeared to have inconsistent levels of expression. It is possible that these three SNORDs are expressed at the predicted patterns in lung and colon endothelial cells, but not those of the liver. However, it is also possible that the differential expression for these three SNORDs is a result of unresolved technical variation. In point of fact, it is worrying that any snoRNA were detected in the RNAseq datasets. The methodology that was employed to generate the sequencing library should select specifically for the poly-A tailed RNAs from the total RNA (this dataset was originally generated to detect protein coding genes and poly-A selection can enhance the sequencing depth of mRNA by excluding classes of RNA such as rRNA). If this had been the case, the vast majority of the snoRNA would not have been sequenced because they are mostly derived from introns and therefore do not have poly-A tails (unless poly-A tails were specifically added following excision). Therefore, the fact that the snoRNA were detected indicates that there was a failure in the poly-A selection process and there is no guarantee that the selection failed to the same extent in each condition (hence the need for qPCR validation).

Perhaps the most surprising feature of this project is the high expression of all the snoRNA in the TLE (and HLE to different extents) when compared to the expression in the other three cell types by qPCR. Moreover the expression does

not seem to be caused by the parallel differential expression of the host genes, rather by a specific maintenance (or excision) of the snoRNAs in the TLE and HLE. The relationship between SNORD75 and SNORD76, and their common host gene GAS5 provides support for this dynamic (Figure 4.3, 4.4 and 4.5). SNORD76 is enriched from GAS5 in the liver endothelium to a greater extent than SNORD75. While the reasons behind this are unknown, it is possible to say that the event is deliberate and cannot be due to differential splicing of the host gene (Figure 4.11), which would cause the opposite expression pattern. Moreover the lack of inclusion of the other snoRNA from GAS5 within the differential expression list gives anecdotal evidence that changes in the expression of GAS5 is not the cause of the enhanced enrichment of SNORD76.

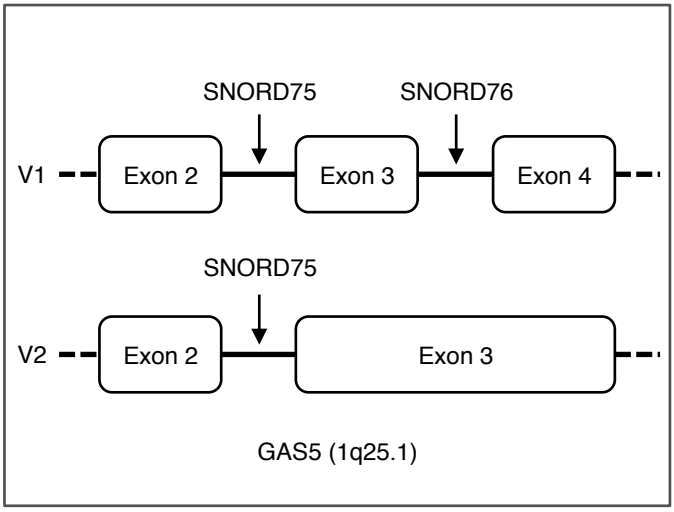


Figure 4.11: Alternative splicing of GAS5

Of the two known GAS5 splice variants, only variant 1 (V1) can produce both SNORD75 and SNORD76. Whereas variant 2 (V2) is likely to be only capable of producing SNORD75 due to the inclusion of SNORD76 within the third exon (original diagram).

The cause of the aforementioned mechanism is currently unknown, yet it is curious that it is not mimicked in HUVEC. If this phenomenon is unique to endothelial cells, it is at least only present in adult endothelial cells. The enrichment of the snoRNA is therefore probably a result of one of two causes. Firstly, the snoRNA are being deliberately excised in the liver associated endothelium (regardless of whether the snoRNAs serve a purpose or not). Secondly it could be a response to the cellular environment that is unique to the patients from which the patient matched tumour and healthy tissue was derived. One such cause of this type of response could be the anti-cancer therapies the patient was receiving at the time of surgery. The expression of snoRNA has been observed to change based upon cellular stresses such as ionising radiation [(76)], which is used in some cases as a treatment for cancer.

Regardless of the mechanism by which the snoRNA are expressed in greater quantities in liver endothelial cells; SNORD75, SNORD76, SCARNA7 and SNORA81 are expressed higher in the tumour endothelium. As stipulated earlier, both SNORD75 and SNORD76 are derived from GAS5. It is interesting that they were observed to be expressed at higher levels in the tumour endothelial cells of three tissues considering that the most documented role for GAS5 is in tumour suppression. When downregulated GAS5 has been associated with tumours of poor prognosis and is generally upregulated in quiescent cells [(78)]. However the activities of SNORD75 and SNORD76 could be independent of GAS5, indeed this idea does not contradict evidence from the scientific literature. Liao *et al*. [(73)] observed that SNORD76 was upregulated in NSCLC, which is one of the tissues

from which the endothelial cells were isolated for the NGS. Moreover GAS5 has other known functions to which SNORD75 and SNORD76 could be ascribed. One of these functions is in cellular starvation [78], which could result from the microenvironmental stresses that tumour endothelial cells are exposed to. Such microenvironmental stresses would be consistent with the qPCR for these snoRNA as the other cells were cultured and as such would have abundant nutrients.

The functions of these four snoRNAs are not yet known. However they have been predicted to guide the modification of certain genes. SNORD75 and SNORD76 could guide the 2'O-ribose methylation of different locations within 28S rRNA. SNORA81 has also been predicted to guide the pseudouridylation of 28S rRNA. SCARNA7, on the other hand is predicted to guide the 2'O-ribose methylation of the U1 spliceosomal RNA [70, 79]. It is possible that these snoRNAs are contributing to pathologically relevant changes within the proteome of tumour endothelial cells (and endothelial cells in general) through the above mechanisms. However it is also plausible that they are also functional by mechanisms as yet unknown.

Out of the four differentially expressed snoRNAs (as determined by qPCR), SNORA81 was the most enriched in the endothelium compared to the other cell types. SCARNA7 also had low levels of expression in the non-endothelial cells and had the largest differential expression between the TLE and HLE (50% lower in HLE). It is for these reasons that SNORA81 and SCARNA7 look like particularly interesting candidates for further research. Nevertheless the differential expression patterns of all four of the snoRNA imply a possible angiogenic function and are all

worth further exploration of both their own function and their downstream implications, which could be employed as anti-angiogenic targets.

Chapter 5: RNAseq analysis of kidney cancer endothelium

5.1 Introduction

Endothelial cells can be highly specialised and be markedly different in different tissues and organs [46]. This is certainly true in the kidney, where endothelial cells not only carry out their usual function, but can also form fenestrated capillary tufts in the glomerulus to filter blood and produce urine [80]. Moreover, endothelial cells within the inner medulla of the kidney are exposed to a hyperosmolar hyperkalemic environment and very low oxygen levels [46].

This functional difference makes renal cell carcinoma a prime target for a large-scale genomic study to identify organ specific markers of the tumour vasculature.

NGS has emerged as the preferred method of large scale genomic and transcriptomic characterisation. However, it is interesting to note that all too often little attention has been paid to the fundamentals associated with experimental design. Biologists have had the tendency to not treat NGS experiments as any other standard experiment with respect and close attention to controls and reproducibility [61]. This has largely been due to the precedent set by earlier SAGE studies, like the work of St. Croix *et al*. [8], where it was not feasible to have replicates due to the cost and laborious nature of the studies.

To an extent the rapid fall in the cost of NGS has precipitated more rigorous, well-designed and statistically robust experimentations. However, observational studies without biological and technical controls are common in the literature. The

Bicknell group published one such example of this: Zhuang *et al*. [81] is the only paper to date to utilise NGS technology to identify tumour endothelial cell markers. Zhuang *et al*. [81] utilised the SOLiD4 to probe the transcriptome of non-small cell lung cancer associated and adjacent healthy tissue associated endothelial cells, however it utilised the same methodology as the data presented in Chapter 3 of this thesis. As a result it suffers from the same lack of technical and biological controls. It was due to these deficiencies that the data was not used to select targets independently, but rather to confirm the selection of candidates from a more well-designed microarray experiment. Therefore, the power and advantages associated with the RNAseq data presented was restrained. RNAseq has by far more potential to identify novel targets than microarray data due to its non-targeted nature and ability to count discrete reads (as described in Chapter 1). This Chapter on the other hand adheres to a stricter canon with regard to data acquisition and analysis. Ultimately the data as presented will facilitate greater insights into the tumour endothelial transcriptome and will be sufficiently robust to stand independently, unlike the studies that have preceded it.

The availability of resected tissue that is large enough to allow for a sufficient number of endothelial cells to be isolated (and therefore RNA) is a great limitation. Especially when accounting for the laborious nature of the endothelial cell isolation. Nevertheless high quality RNA was obtained from renal cell carcinoma endothelial cells (RCCEC), healthy adjacent kidney endothelial cells (HKEC) and whole/bulk (endothelial cell depleted) kidney tissue (WNK). The three isolates from three patient-matched (tumour and healthy adjacent tissue) biological

replicates were barcoded (ligation of identifying nucleotide tags) and pooled directly after poly-A tail selection and RNA fragmentation (before library preparation). The multiplexing of the samples in this manner allows for the control of any bias that could be introduced during the reverse transcription, PCR amplification and sequencing. All 9 of the samples were then split across two Illumina HiSeq2000 lanes to enhance the read depth, without compromising the technical rigor [61].

The inclusion of WNK is an important strength of this project when compared to the previous two Chapters. Any candidates that are found to be differentially expressed between the RCCEC and HKEC can be checked against the WNK. This feature allows for the quick exclusion of genes that are present at high levels in normal kidney cells and are therefore not endothelial cell markers. Moreover the fact that the WNK is made up of multiple patient matched cell types is an improvement upon the study by St. Croix *et al*. [8], which utilised non-patient matched isolated non-endothelial cells.

The bioinformatics pipeline used to conduct the analysis of RNAseq data can give wildly different results based upon the methods and algorithms they use. The 'Tuxedo' pipeline developed by Trapnell *et al*. [82] was employed to conduct RNAseq analysis. Tuxedo is a widely used (perhaps the most used) open-source pipeline and as such it has been tested extensively. Moreover, it is a fully connective pipeline that ensures excellent compatibility between the different stages of the analysis. Tuxedo is also a prime pipeline for this project because it is capable of tolerating more than two variables in a single differential expression

analysis. Tuxedo also has a number of advantages because its aligner 'Tophat' utilises the popular 'Bowtie' mapping tool, which is an ultrafast and memory-efficient method of mapping to a Burrows-Wheeler indexed genome, the UCSC (University of California, Santa Cruz) GRCh37//hg19. Fundamentally this allows for the potential identification of novel genes and splice variants, which would not be possible using transcriptome mapping.

5.2 Results

CummeRbund was utilised to visualise the RNAseq data upon completion of the differential expression algorithms. Firstly, data from the three variables (RCCEC, HKEC an WNK) were compared to each other at the whole transcriptome level. Thereby grouping the variables based upon the similarity of their overall differential gene expression levels. Of the three variables, the gene expression levels were more similar between WNK and HKEC, than to RCCEC (Figure 5.1).

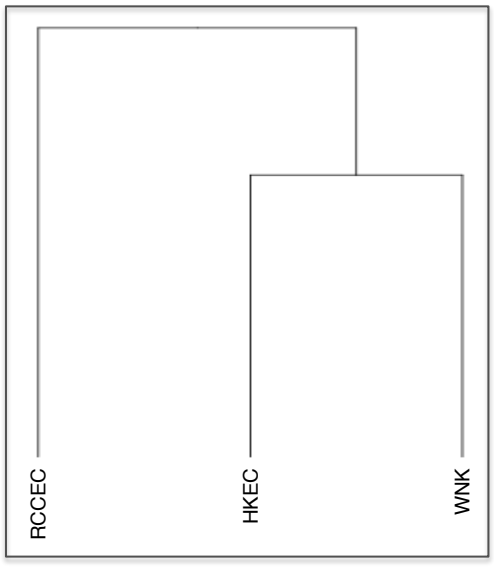


Figure 5.1: Phylogeny of the tumour vasculature

The overall expression profile of all three samples demonstrates that RCCEC, HKEC and WNK were different to each other. The assignment of RCCEC to a separate 'leaf node' (branch) indicates that the gene expression profiles were more similar between HKEC and WNK, whereas more genes were differential expressed in RCCEC. This analysis suggests that the gene expression profile of RCCEC was affected by factors such as cellular environment, rather than the gene expression profile associated with cell type.

Secondly CummeRbund was used to generate a heat map (Figure 5.2), which allowed for the visualisation of the differential expression patterns of individual genes within the dataset as a whole. Importantly the heat map confirmed the presence of discrete gene clusters, some of which demonstrated specificity to each of the three variables. Cluster analysis was conducted using CummeRbund to reveal individual genes that displayed the expression patterns of interest.

The first cluster of interest queried (100,0,0) contained genes that were expressed at high levels in the RCCEC, but were absent (or only trace expression) in both HKEC and WNK (Table 5.1). In total 24 potentially tumour endothelial-

specific genes were identified, one of which was classified as non-protein coding (LOC643733). By contrast, the second cluster probed (0,100,0) contained genes that were expressed at high levels in HKEC, but at lower levels in RCCEC and WNK (Table 5.2). This cluster contained a total of 12 genes that displayed a suppressive, healthy endothelial cell enriched expression pattern, one of which was classified as non-protein coding (H19).

The five most upregulated and downregulated non-protein coding transcripts in the RCCECs compared to the HKECs from the RNAseq data were also manually identified (Table 5.3). In which LOC643733 and H19 featured prominently along side multiple novel transcripts. Of these transcripts FLJ41200 (LOC401492) also demonstrated potential tumour endothelial cell-specificity.

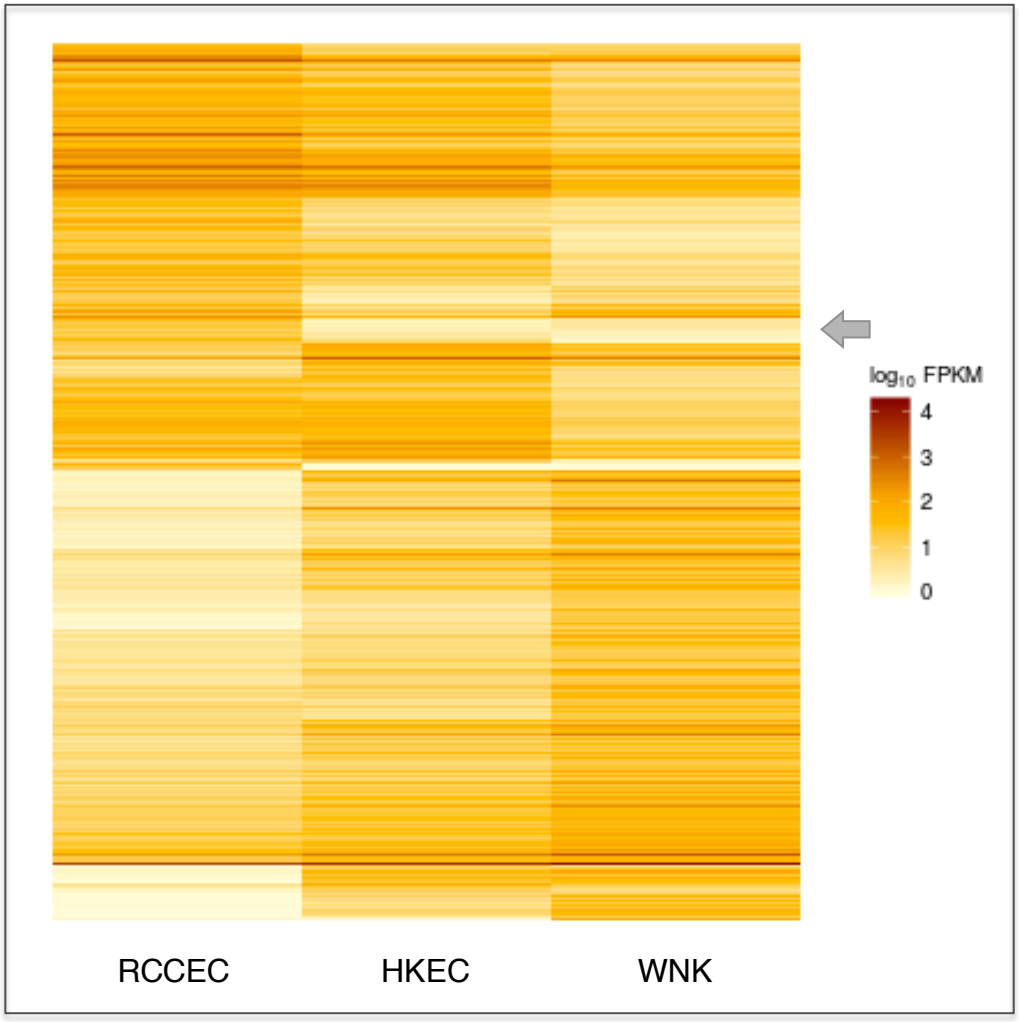


Figure 5.2 Cluster analysis of differentially expressed genes

The heat map represents the $\log_{10}$ fold change of FPKM (fragments per kilobase of transcript per million mapped reads) per differentially expressed gene (red = high expression; white = low expression). There are distinct clusters of genes that are expressed preferentially in each of the three samples. The arrow highlights a gene cluster that appears to be specific to RCCECs.

Gene	FPKM		
	RCCEC	HKEC	WNK
INSR	149.48	16.28	10.80
NDUFA4L2	102.46	12.14	13.38
INHBB	101.85	8.09	1.56
KCNE3	101.47	1.66	3.94
SERPINI1	97.76	7.54	10.04
CXCR7	95.75	6.35	6.39
TGFBI	95.67	6.35	13.24
SCARB1	90.41	5.17	3.52
MCF2L	84.94	16.97	6.75
ANGPTL4	84.25	7.05	9.30
LAMA4	74.61	8.41	3.63
MMP2	73.64	11.96	10.33
COL8A1	72.06	2.02	2.59
IL7R	70.73	3.47	7.73
KRT14,KRT17	70.12	2.05	4.27
LOC643733	68.13	14.24	3.44
C3	65.40	2.72	8.37
ANGPT2	65.30	4.36	1.48
MAGI1	62.66	9.61	11.64
TMCC3	60.45	10.59	5.97
PXDN	57.50	13.60	7.00
CHST15	56.58	8.82	4.49
NRP2	53.83	10.79	3.26
PGF	51.01	8.85	9.62

Table 5.1: Strongly tumour endothelial specific genes

This cluster of genes was significantly more highly expressed in RCCEC. The absolute expression of the genes in the three variables is displayed using ‘fragments per kilobase of transcript per million mapped reads’ (FPKM).

Gene	FPKM		
	RCCEC	HKEC	WNK
TIAM1	34.72	757.69	117.86
TFPI2	72.47	524.00	64.60
PMAIP1	119.61	516.07	118.07
PLAT	17.76	250.46	14.36
LRG1	2.92	200.51	27.28
IRF1	16.33	170.34	39.03
IL8	19.82	125.16	18.75
H19	27.94	114.92	41.53
IL6	16.65	108.71	25.77
ICAM1	10.96	96.46	11.35
GPM6A	4.03	81.04	17.10
FMO2	8.87	50.21	4.43
PCAT19	23.31	44.21	6.02

Table 5.2: Genes enriched in the healthy endothelium

This cluster of genes was expressed at significantly levels in HKEC (FPKM).

Gene	Locus	RCCEC	HKEC	log2 (FC)	p value	WNK
FLJ41200	chr9:13406378-13433075	47.80	0.26	-7.53	0.00005	0.16
XLOC_016373	chr22:17177801-17182186	22.63	1.18	-4.26	0.00005	0.81
LOC541471	chr2:111954254-112252884	45.69	6.98	-2.71	0.00050	13.51
DDX12P	chr12:9549677-9600796	30.80	5.09	-2.60	0.00005	1.32
LOC643733	chr11:104772275-104789053	68.13	14.24	-2.26	0.00005	3.44
H19	chr11:2016405-2019065	27.94	114.92	2.04	0.00005	41.53
XLOC_002346	chr1:16567639-16568319	7.73	32.47	2.07	0.00005	20.05
XLOC_008708	chr15:56365833-56365983	20.48	108.61	2.41	0.01250	259.14
LOC553103	chr5:131628997-131731307	1.93	71.19	5.20	0.00005	84.29
XLOC_020105	chr5:169615221-169626215	0.24	27.82	6.87	0.00005	41.51

Table 5.3: Top significantly differentially expressed ncRNAs

The top differentially expressed (red = high expression in RCCEC compared to HKEC; blue = low expression in RCCEC compared to HKEC) ncRNAs between RCCEC and HKEC were cross-referenced with their expression in WNK (FPKM). The locus of each transcript is zero-based.

5.3 Discussion

The cluster analysis performed in this Chapter is not only a means of identifying interesting targets for further study, but also validates the RNAseq experiment as a whole. The majority of the genes identified through the cluster analysis were protein coding. Whilst the expression of protein coding genes is not of direct interest in this study, the scientific literature associated with those genes is more developed. Therefore if a cluster analysis designed to identify tumour endothelial-specific genes actually identifies genes known to have those properties it is more likely that the read values accurately represent the transcriptome of the RCCECs, HKECs and WNK.

The majority of the genes identified by the cluster analysis when searching for tumour endothelial cells markers are indeed known to be involved with pathological angiogenesis associated tumour growth and progression. MMP2 is particularly 'eye catching' as it has an extremely well established role in both physiological and tumour angiogenesis [83, 84] and was also identified by St Croix *et al*. [8]. Additionally this list includes many other known genes that have been shown to be strong promoters of tumour angiogenesis including (but not limited to) INSR [8], CXCR7 [86], ANGPT2 [87] and PGF [88]. Moreover and reassuringly, many of these genes have been used as targets for therapies in clinical trials [88-90].

The fact that LOC643733 was shown to have comparable expression to these genes through the cluster analysis is extremely encouraging (due to its low expression in HKEC and WNK, and relatively high expression in RCCEC) and is certainly worth further study. Thus far, LOC643733 (11q22.3) has not been the

subject of any functional studies, nor has it been associated with a tumour endothelial-specific expression pattern. However, it has been annotated by NCBI (National Center for Biotechnology Information) as being a Caspase 4, Apoptosis-Related Cysteine Peptidase Pseudogene, which could be a clue to its potential function. As pseudogenes have been shown to play a role in protecting (acting as decoys) their namesakes and other genes from degradation by miRNA [91].

The strength of the tumour endothelial-specific cluster analysis also lends credibility to the endothelial cell-specific tumour suppressor analysis. The inclusion of H19 on this list is extremely interesting as it is one of the most well defined functional ncRNA in the scientific literature, and one of the earliest discovered [92-94]. H19 has been shown to be a tumour suppressor gene and mutations in H19 have been associated with Beckwith-Wiedemann syndrome and Wilms tumourigenesis, which matches its suppressor profile in this RNAseq experiment. Despite its fame very little research has been conducted to determine the role of H19 in angiogenesis and endothelial cell biology and therefore, may be worth future study.

Although it is possible to observe differences between variables using RNAseq, the data cannot directly prove the biological relevance of a gene. Therefore it is questionable whether RNAseq can be directly used definitively to prove a hypothesis [95]. But the data presented in this Chapter gives the closest snap shot of the tumour endothelial cell transcriptome that is currently available. Further analysis of this dataset combined with follow-up in the laboratory could

lead to the discovery of coding and non-coding transcripts that have angiogenic and tumourigenic significance.

Chapter 6: The functional exploration of PCAT19

6.1 Introduction

Prostate Cancer Associated Transcript 19 (PCAT19) was highlighted as a potential endothelial cell-specific suppressor (due to it's downregulation in tumour associated endothelial cells) using machine learning algorithms, specifically cluster analysis, to probe the RNAseq data described in Chapter 5 (Table 5.2). PCAT19 is classified as a long intergenic non-protein coding RNA (lincRNA), a class that up until recently has been dismissed as 'junk' [69]. Coincidentally, the Bicknell group identified PCAT19 as a gene in 2000 [31] and confirmed its restriction to endothelial cells [31, 36]. At that point in time PCAT19 was termed ECSM1 (Endothelial Cell-Specific Molecule 1) because of its endothelial-specific expression pattern. However, the nomenclature was not accepted due to the prevailing assumption that transcripts lacking a substantive open reading frame were non-functional. It is perhaps because of the stigma attached to this RNA class that deterred research into the function of PCAT19. Moreover, the nomenclature rejection and the existence the similarly named ESM1 (also Endothelial Cell-Specific Molecule 1) effectively severed the link between this early functional insight and the gene it was attributed to.

In the intervening 15 years since the discovery of PCAT19 and the writing of this thesis, it has been renamed multiple times depending on the locus it was assigned to in the human genome. LOC100505495 was the most recently approved nomenclature. The change of nomenclature from LOC100505495 to PCAT19 was prompted by the publication of two strikingly large independent

genome-wide association studies (GWAS) in 2013 [96, 97]. These studies profiled the genomes of 20,000 patients and identified PCAT19 as hosting the most highly associated SNP (Figure 6.1) with prostate cancer (including aggressive prostate cancer) and increased mortality of the patients (increased risk by 1.18). These two pieces of published information, in combination with the RNAseq cluster analysis, prompted the hypothesis that PCAT19 is functions as a suppressor endothelial cells. Of all the candidates identified in this study, PCAT19 was considered by the author to be the prime candidate with which to experimentally demonstrate the functional properties of ncRNAs in endothelial cells.

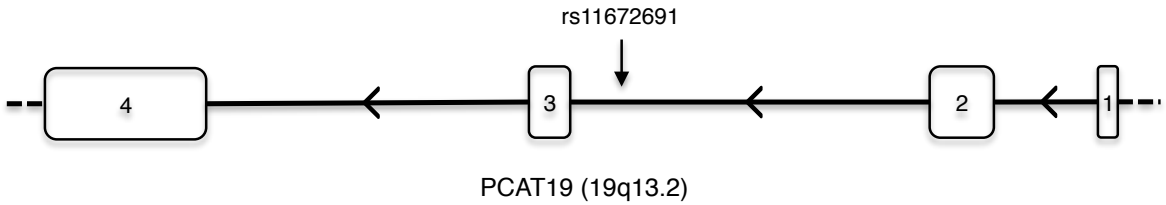


Figure 6.1: PCAT19 contains a prostate cancer associated SNP

The SNP rs11672691 (A/G substitution) is located within the second intron of PCAT19. Therefore it could potentially induce the production of different spice variants or create an alternative intragenic promoter (original diagram).

6.2 Results

A cluster analysis designed to identify potential endothelial cell-specific suppressor transcripts flagged PCAT19 as being restricted to endothelial cells and expressed at lower levels in RCCEC when compared to HKEC (Figure 6.2). The expression of PCAT19 in HKEC was roughly double that of RCCEC, which was mimicked in TLE and HLE (53% increase) when validated using qPCR (Figure 6.3). Moreover

PCAT19 was shown by qPCR to be expressed at progressively greater levels in HUVEC as the cells approached confluence (Figure 6.4).

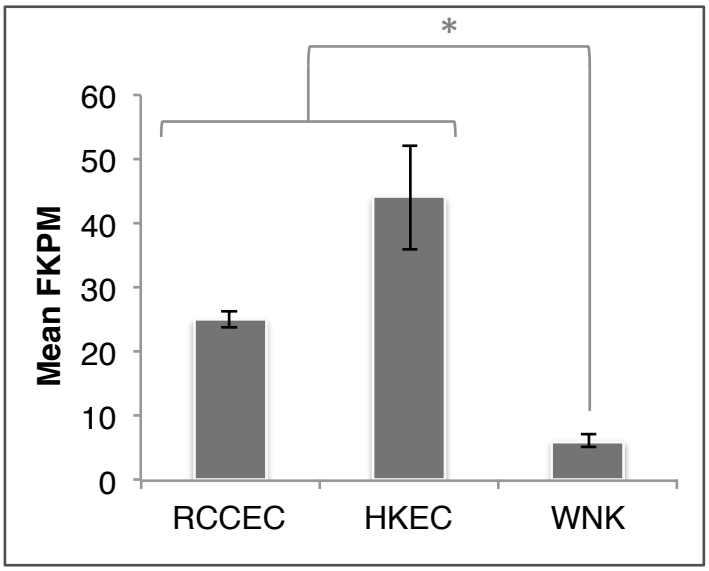


Figure 6.2: RNAseq gene level plot for PCAT19

The Illumina RNAseq data showed that PCAT19 was expressed at significantly ($p = < 0.05$) lower levels in WNK than both endothelial cell isolates, and was expressed at higher levels in HKEC than RCCEC (mean ±SEM)..

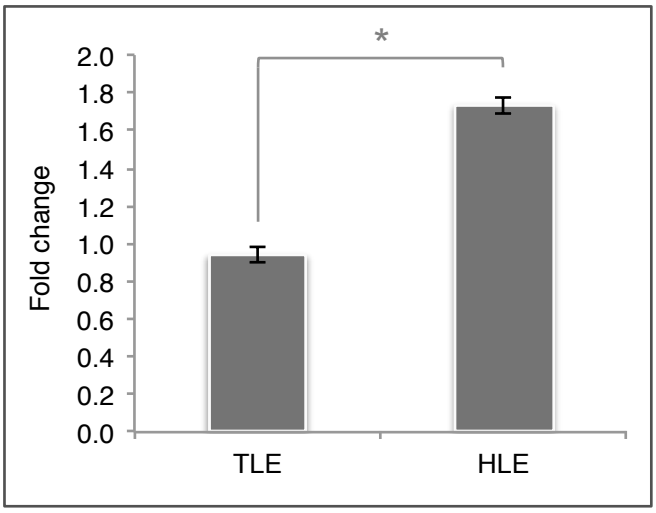


Figure 6.3: PCAT19 is expressed to a lesser extent in TLE

When normalised against FLOT2, PCAT19 was expressed at significantly ($p = < 0.05$) reduced levels in TLE compared to HLE (mean ±SEM).

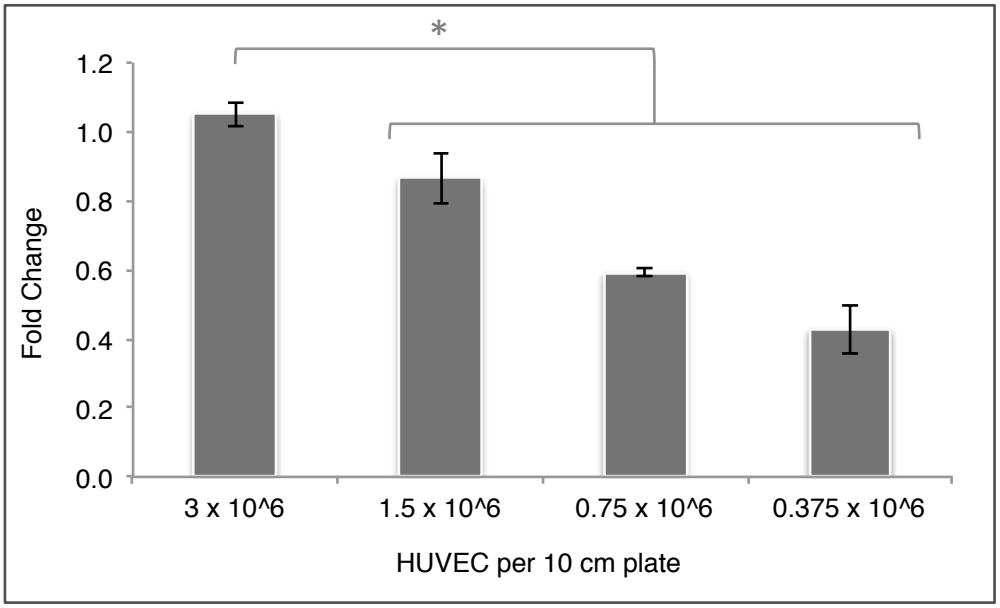


Figure 6.4: PCAT19 expression correlates with cell density

PCAT19 when normalized against ACTB (house keeping gene) was expressed at significantly greater levels in HUVEC that were cultured at higher densities (mean ±SEM).

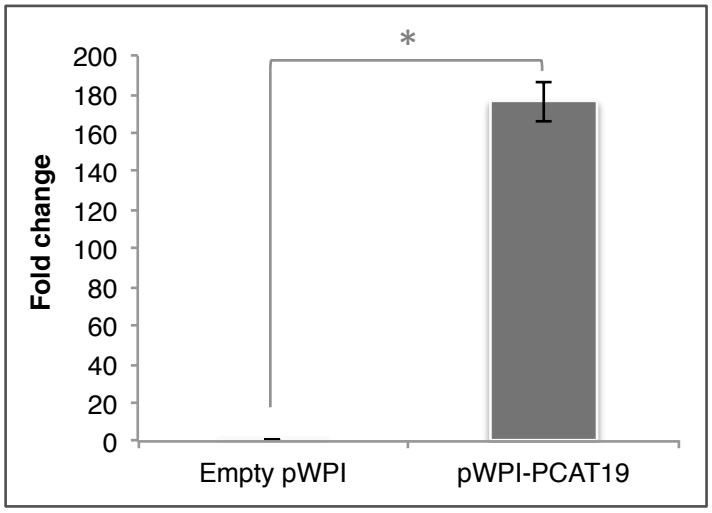


Figure 6.5: Post-lentivirus transduction expression levels of PCAT19

Virus containing empty (blank) pWPI plasmid and pWPI containing PCAT19 were exposed to HUVEC (separately). The transduction of PCAT19 under the control of the EF-1α promoter was successfully introduced to HUVEC, which resulted in greater expression levels of PCAT19 (mean ±SEM).

To explore whether the increase of PCAT19 in endothelial cells was of functional relevance *in vitro*, the expression levels of PCAT19 were increased in HUVEC using lentiviral transduction (Figure 6.5). Subsequently PI staining was conducted to facilitate cell cycle analysis using an Acumen cytometer on both the negative control (normal) HUVEC and the HUVEC transduced (overexpressing) with PCAT19 (Figure 6.6). The HUVEC containing elevated levels of PCAT19 exhibited significantly ($p = <0.05$) higher levels of dead cells and cells undergoing apoptosis and a potential decrease in the proportion of HUVEC in G1 and G2/M phases of the cell cycle.

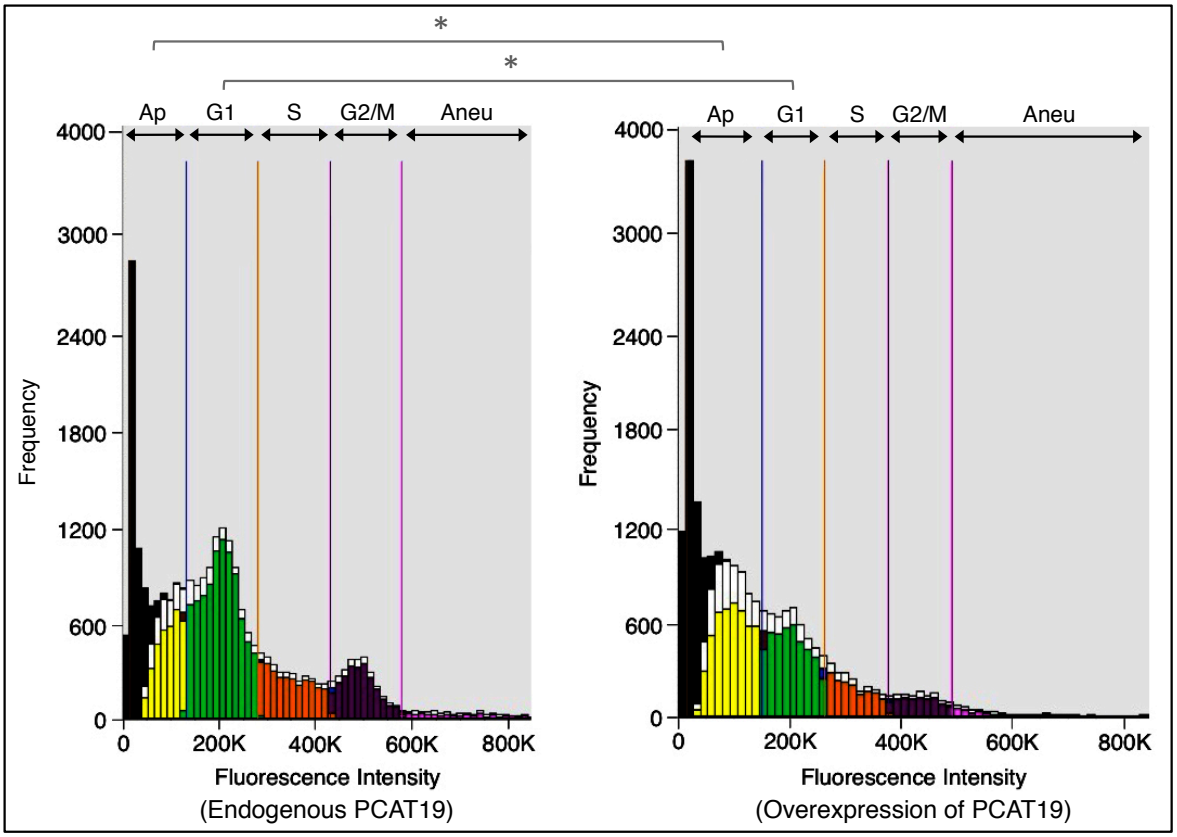


Figure 6.6: The effect of PCAT19 on the cell cycle of HUVEC

The histograms from Acumen cytometry data (8 replicates) demonstrates that HUVEC containing increased levels of PCAT19 are significantly ($p = < 0.05$) more prone to apoptosis and less likely to be in the G1 and G2/M stages of the cell cycle (mean ±SEM).

To gain further insights into the molecular basis for this interaction, siRNA knockdown of PCAT19 was performed in HUVEC (Figure 6.7), the RNA from which was used to produce fluorescently labelled cDNA and hybridised to an Agilent microarray chip. The subsequent microarray analysis provided a list of genes that could be differentially expressed as result of the PCAT19 knockdown. Knockdowns using siRNA carry possibility of causing off-target effects that could modify gene expression independently of PCAT19. It was for this reason that two duplexes (designed to target different sequences within PCAT19) were used and by cross-referencing the differentially expressed gene lists, a number genes that could interact with PCAT19 were identified (Table 6.1).

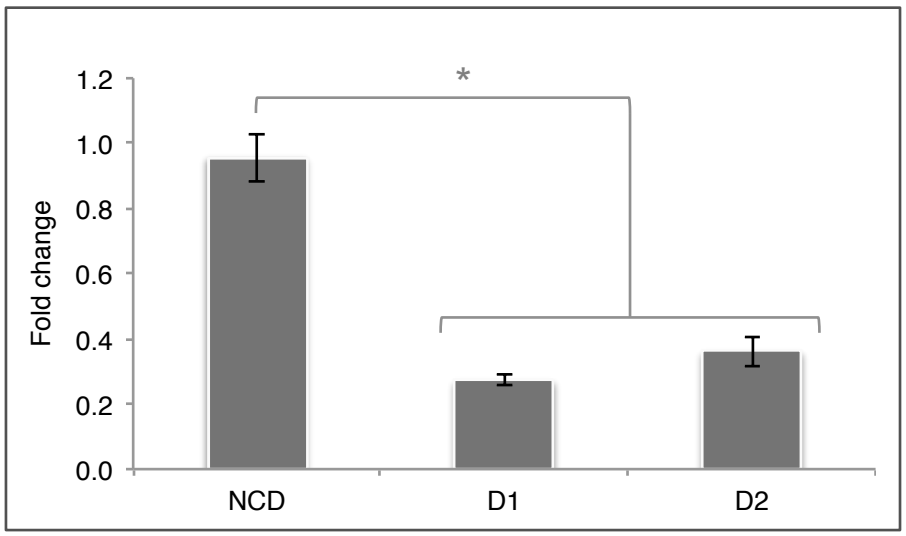


Figure 6.7:Confirmation of PCAT19 knockdown in HUVEC

HUVEC were transfected with a NCD or a duplex (D1 or D2) designed to knockdown PCAT19, both of which were determined by qPCR to be successful ($p = < 0.05$) towards this end (mean ±SEM).

Gene	Log. Fold Change		P Value
	Duplex 1	Duplex 2	
WTAP	0.74	1.00	0.000009
HIST1H2BK	1.08	0.97	0.000796
CBX5	- 0.85	- 0.86	0.000023
SUMF1	- 0.99	- 0.86	0.000022
ILII	- 1.04	- 0.94	0.000130
CNN1	- 1.69	- 0.91	0.000053
HMOX1	- 0.88	- 1.73	0.000005

Table 6.1: Differentially expressed genes following the knockdown of PCAT19

The two-colour microarray analysis (NCD versus each PCAT19 duplex) highlighted a number of differentially expressed genes. The seven genes listed in this table were chosen for their differential expression in both sets of HUVEC.

The cDNA obtained from the overexpression studies was used to perform qPCR validation of the microarray data, any gene that was upregulated after the knockdown of PCAT19 should be downregulated by the overexpression of PCAT19 (and vice versa). WTAP (Figure 6.8) and HIST1H2BK (Figure 6.9) were upregulated following knockdown of PCAT19, but rather than being downregulated following the overexpression of PCAT19 their expression levels remained relatively constant. Of the five genes observed to be downregulated in the microarray analysis: CBX5 (Figure 6.10), SUMF1 (Figure 6.11) and ILII (IL-2) (Figure 6.12) all displayed increased expression levels, but only CBX5 was significantly differentially expressed. Whereas, CNN1 (Figure 6.13) and HMOX1 (Figure 6.14) did not display the predicted change and were expressed at lower levels following the overexpression of PCAT19.

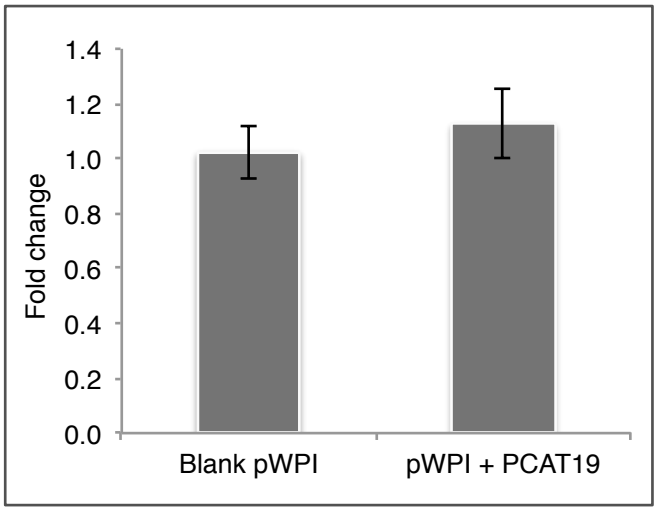


Figure 6.8: WTAP expression when PCAT19 is overexpressed

WTAP was not differentially expressed following the overexpression of PCAT19 (mean ±SEM).

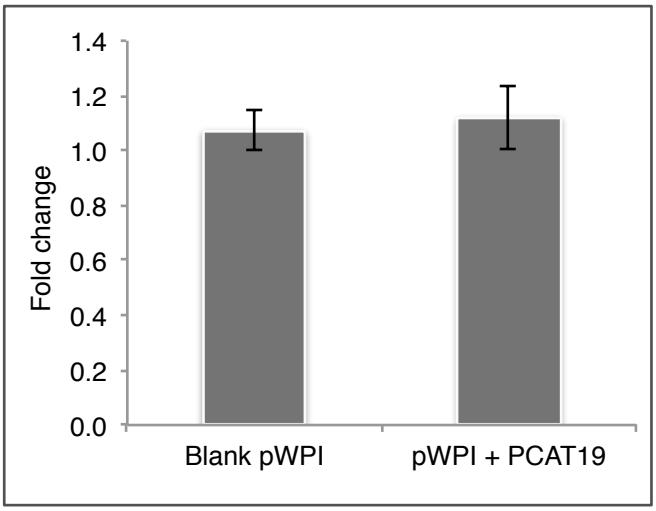


Figure 6.9: HIST1H2BK expression when PCAT19 is overexpressed

HIST1H2BK was not differentially expressed after the overexpression of PCAT19 (mean ±SEM).

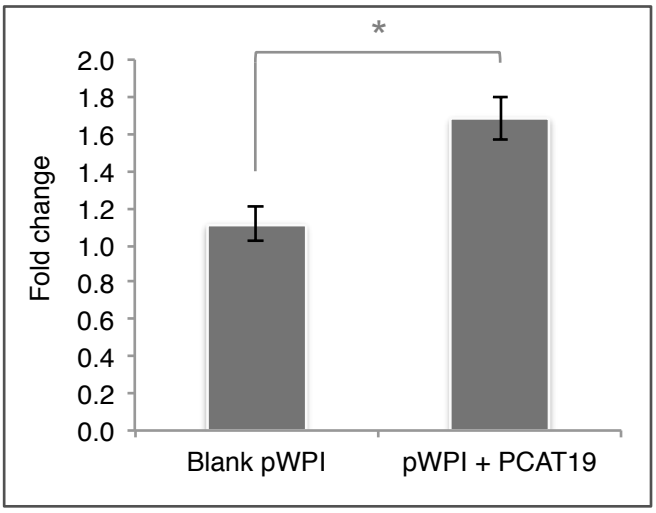


Figure 6.10: CBX5 is upregulated when PCAT19 is overexpressed

CBX5 was expressed at significantly ($p = < 0.05$) increased following the overexpression of PCAT19 (mean ±SEM).

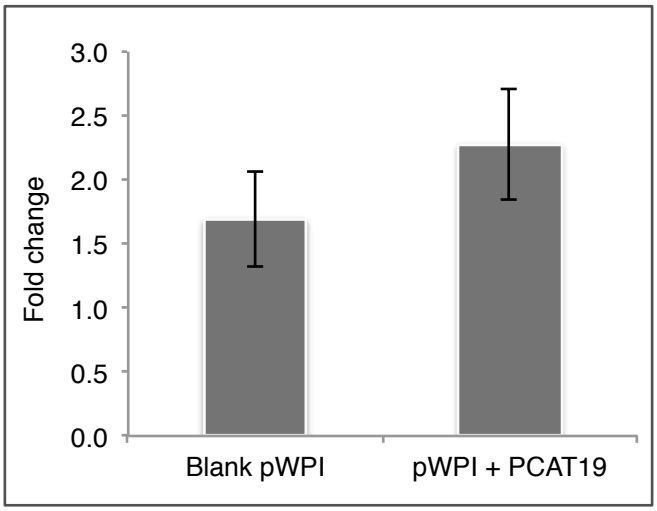


Figure 6.11: SUMF1 expression when PCAT19 is overexpressed

The expression of SUMF1 was not significantly increased following the overexpression of PCAT19 (mean ±SEM).

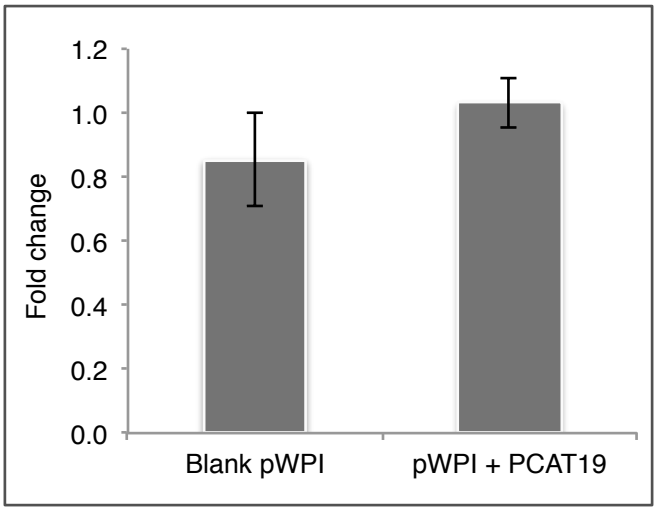


Figure 6.12: ILII expression when PCAT19 is overexpressed

The expression of ILII was not significantly increased following the overexpression of PCAT19 (mean ±SEM).

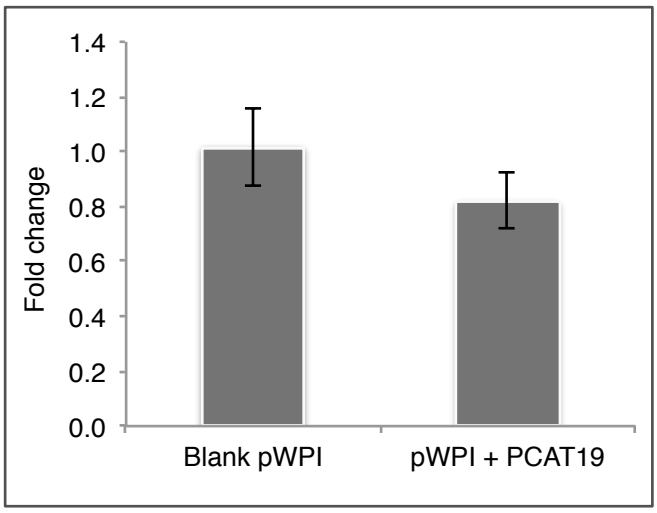


Figure 6.13: CNN1 expression when PCAT19 is overexpressed.

The expression of CNN1 was decreased following the overexpression of PCAT19 (mean ±SEM).

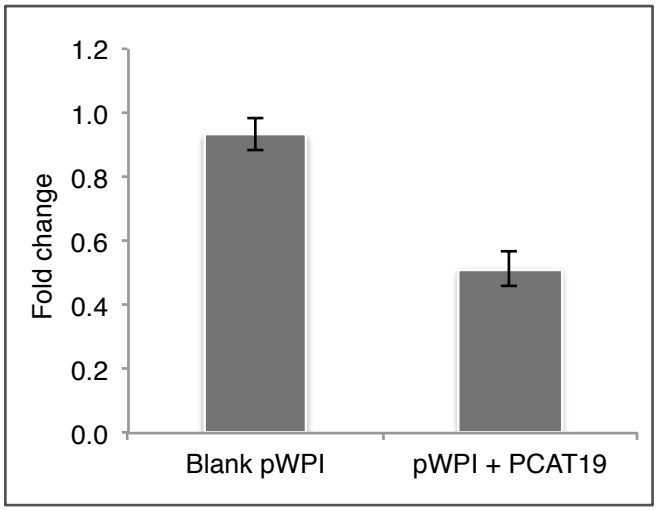


Figure 6.14: HMOX1 expression when PCAT19 is overexpressed.

The expression of HMOX1 was decreased following the overexpression of PCAT19 (mean ±SEM).

6.3 Discussion

The importance of the tumour vasculature in tumour progression provides a framework as to how a gene such as PCAT19 could have a significant correlation with aggressive prostate cancer. It is not yet known whether the SNP within PCAT19 impacts its function, but it may prime the endothelial cells to be more easily exploited by tumours. The development of vasculature is pivotal for the growth and metastasis of tumours, a tumour must acquire a blood supply to grow beyond ~2 mm in diameter [2, 92].

PCAT19 also appears to affect the vasculature in a non-hereditary manner, through differential expression between the healthy and tumour endothelium. However, it is also not clear what causes the differential expression. The hijacking of blood vessels to supply a tumour requires many growth factors, each of which causes transcriptional changes. Moreover the tortuous and disordered vasculature

network causes a host of microenvironmental changes, each of which is associated with the differential expression of genes [11, 12, 80, 98, 99]. In this work PCAT19 was shown to be responsive to changes to the confluence of endothelial cells. The tumour endothelium is prone to being leaky and having loose cell boundaries when compared to healthy endothelium [10]. A response to the loose cell boundaries between endothelial cells, which is akin to that of sparse HUVEC in which PCAT19 was expressed at lower levels, might be the cause of the differential expression of PCAT19.

The low expresion of PCAT19 in the tumour vasculature and during the HUVECs transition from a quiescent state, to a state of relatively active replication, indicates that PCAT19 could have a role in proliferation. It was for this reason that PCAT19 was overexpressed in HUVEC. The ensuing cell cycle analysis demonstrated that when PCAT19 is expressed at greater levels *in vitro*, endothelial cells are more prone to be apoptotic. Moreover PCAT19 appears to allow more cells through G2/M into G1 (indicating that PCAT19 possibly influences the blocking of the G2/M checkpoint), however further testing would be required to confirm this (which is currently being carried out by members of Bicknell and Nagy groups at the University of Birmingham, including the effect of PCAT19 knockdown at different cell densities on HUVEC cell cycle). Overall, PCAT19s has the hallmarks of a tumour suppressor because of its effect on the cell cycle and its synchronous downregulation in tumour endothelial cells. This prospect is even more exciting because of the effective restriction of PCAT19 to endothelial cells.

Nevertheless the molecular mechanism by which PCAT19 enacts the aforementioned function is not yet known. It is possible that PCAT19 acts as a molecular sponge and protects other suppressors from miRNAs or other RNA decay pathways. But it could just as easily act as an epigenetic guide. The microarray analysis and subsequent qPCR validation described in this Chapter has provided useful hints towards this end. It is possible that PCAT19 interacts with heterochromatin protein chromobox homolog 5 (CBX5), also known as heterochromatin protein 1 (HP1) alpha (α), from 12q13.13 at the RNA level due to their significant co-expression observed in this study (overexpression of PCAT19 is associated with higher CBX5 expression and knockdown of PCAT19 is associated with lower CBX5 expression). Furthermore, CBX5 is known to localise to heterochromatin during interphase and detach at the start of mitosis [100-101], the stage at which PCAT19 appears to have a functional affect according to the Acumen data.

It is not yet possible to determine whether the interaction between CBX5 and PCAT19 is direct or indirect, but RNA immunoprecipitation may provide an answer. Nevertheless, PCAT19 appears to influence the expression of CBX5, and the modulation of PCAT19 expression yielded similar functional outcomes in this study, when compared to the functional effect of CBX5 knockdown (when PCAT19 and CBX5 expression is low growth arrest is more likely, and *vice versa*) [100]. Lee et al [100] demonstrated that the depletion of CBX5 caused cell cycle arrest at the G2/M checkpoint through an interaction with BRCA1 (breast cancer 1). The aforementioned reasons raise the possibility that PCAT19 influences BRCA1

function indirectly through CBX5 (Figure 6.15). This prediction is also strengthened by the observation that PCAT19 has a tumour suppressor-like expression pattern (lower expression in healthy endothelium, than tumour endothelium).

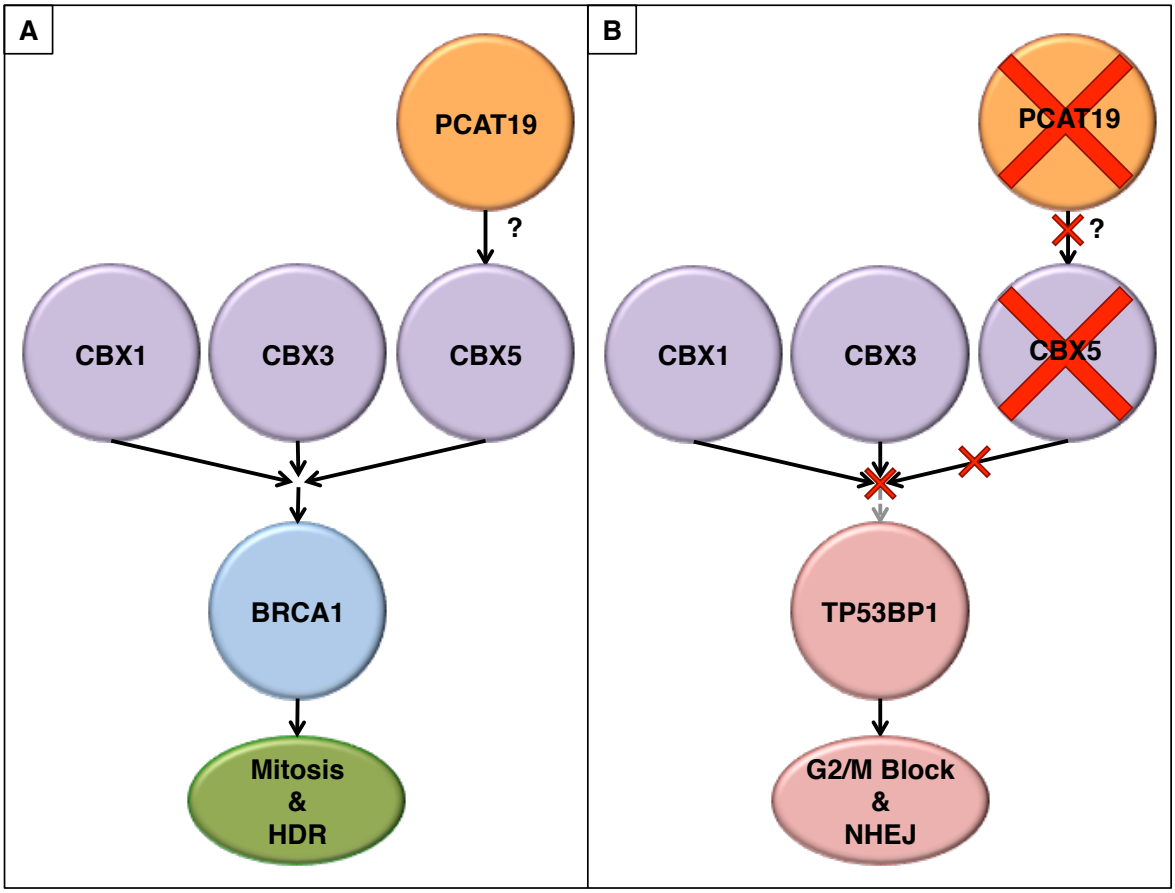


Figure 6.15: Predicted involvement of PCAT19 with the HP1-BRCA1 pathway

CBX1 (HP1-β), CBX3 (HP1-γ) and CBX5 (HP1-α) have all been shown by Lee et al [(100)] to be independently (non-redundantly) important for the recruitment of BRCA1 to sites of double strand breaks. The depletion of CBX1, CBX3, and CBX5 expression, reduced the recruitment of BRCA1 to double strand breaks and caused defects in the homology directed repair (HDR) pathway and the arrest of the cell cycle at G2/M. Intriguingly, the depletion of these proteins also resulted in increased recruitment of the TP53BP1 (Tumour Protein P53 Binding Protein 1), which is involved in the non-homologous end-joining (NHEJ) pathway. If PCAT19 influences CBX5 expression, PCAT19 could therefore indirectly influence cell cycle regulation and the selection of DNA repair

pathways. In this model, PCAT19 enables the expression of CBX5, which allows the recruitment of BRCA1 to double strand DNA breaks (A). However, lower expression of PCAT19 would result in low CBX5 expression, which would block BRCA1 recruitment, and thereby promote the recruitment of TP53B1 (B) (original diagram).

Chapter 7: General Discussion

The involvement of ncRNAs in angiogenesis is an area of increasing interest (Figure 7.1). Many ncRNA molecules have been identified in this thesis through the systematic analysis and the use of NGS technology. The computational and laboratory methods have formed the first concerted effort to demonstrate that ncRNAs are differentially expressed and potentially involved with pathological processes in tumour associated endothelial cells. The results obtained regarding PCAT19 alone effectively prove the hypothesis that was at the core of this thesis ('functional non-protein coding transcripts are expressed specifically in endothelial cells'). PCAT19 shows restricted expression in the endothelium and is differentially expressed between healthy and tumour associated endothelial cells. Furthermore the modulation of PCAT19 *in vitro* causes functional changes in the *in vitro* cell cycle of endothelial cells and could thereby interfere with angiogenesis.

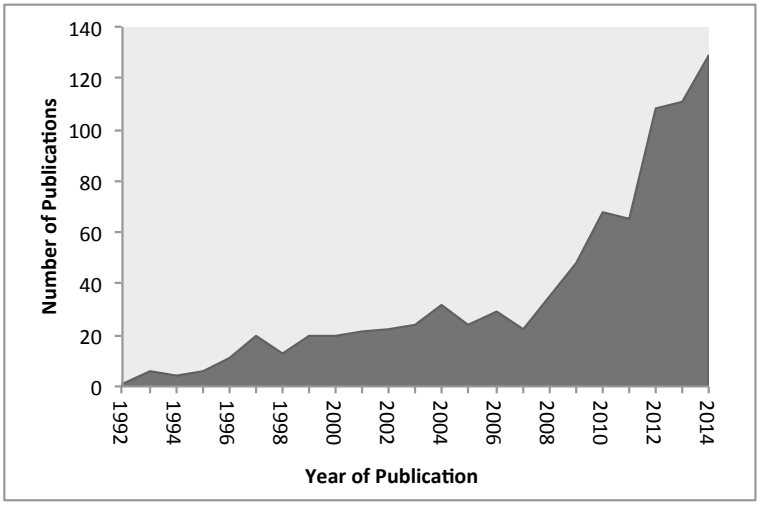


Figure 7.1: Articles pertaining to "angiogenesis and non-protein coding RNA"

In recent years there has been a spike in the number of articles in the field of angiogenesis and ncRNAs. Although it is concerning that only 46% of these articles are primary in nature, whereas 48% are secondary review articles. In the wider field of angiogenesis 21% of the published articles

are review articles and 11% in the field of non-protein coding RNA (original diagram compiled using information from Scopus [102]).

Traditionally the targeting of a protein (or protein product, e.g. glycosylation) has been seen as a prerequisite of any therapeutic intervention. In the case of tumour endothelial cell markers a desired therapeutic target would be localised in either the extracellular matrix or the plasma membrane. This protein location pattern when combined with the specific expression on the vasculature of tumours would therefore enable the systemic delivery of an anti-tumour endothelial cell agent without comprising the integrity of the healthy vasculature. One such example is vaccination against ROBO4 in mice, which limited tumour growth by stimulating the generation of anti-ROBO4 autoantibodies [34]. The tumour endothelial-specificity of ROBO4 was essential to the targeting of this protein; as the targeting of less-specific markers using similar strategies could produce significant toxicity (due to the targeting of healthy vasculature).

However, in recent years there has been a surge in oligonucleotide-based therapeutics to target RNA *in situ*, over 30 of which have progressed into clinical trials [103, 104]. Antisense DNA based oligonucleotides has been the most commonly employed strategy in these clinical trials. These antisense DNA based oligonucleotides specifically form a DNA-RNA heteroduplex with a the desired target, which in turn triggers the RNase H–dependent cleavage of the target [104]. A variety of other strategies have also reached clinical trials including siRNA to degrade VEGF in liver cancer [105], snoRNA to initiate splicing modulation in Duchene Muscular Dystrophy and morpholinos (antisense DNA molecules, which

bind to the target RNA and block interactions with the RNA other molecules) to block translation in restenosis. The individual strategies employed by all of these studies are beyond the scope of this thesis, but comprehensive reviews regarding the targeting strategies, nucleotide modifications, toxicity and pharmacology of oligonucleotide-based therapies have been published by Bennett and Swayze [104] and Burnett and Rossi [105].

The RNA based strategies do share a number of advantages over traditional therapeutics. One of the largest advantages is that once a therapeutic target it identified, it is comparatively easier (and less costly) to design and manufacture an RNA based therapeutic than a protein/drug [104, 105]. Furthermore the RNAs can be economically and reproducibly produced at a large scale for clinical use [105].

On the other hand, RNA therapies have a number of weaknesses including the accumulation of unmodified RNAs in the kidneys. Even the encapsulation of RNAs in liposomes (or other nanoparticles) can result in their accumulation in the liver [105]. Furthermore, upon entering cells the therapeutic effect is transient, and it is possible that the RNAs can trigger an innate anti-viral immune response. However, this risk can be reduced through the use of RNA modifications, such as 2-thiouridine and 5-methylcytidine [106]. RNA modifications are also necessary to prevent nuclease degradation in vivo [105].

Nevertheless, Precedent for further use of oligonucleotide-based technologies for tackling vascular pathology has in part been set by the recent FDA (Food and Drug Administration) approval of drugs including Mipomersen.

Mipomersen is an antisense oligonucleotide based treatment for the cardiovascular disease dyslipidemia (an inherited genetic defect resulting in the elevation of cholesterol). Mipomersen acts by blocking the translation (through RNAase H degradation) of Apo B-100 mRNA at its site of synthesis in the liver. Mipomersen is a synthetic 20 bp long oligonucleotide that utilises a phosphorothioate backbone (the substitution of an oxygen atom with a sulfur atom) and 2'-O-(2-methoxyethyl) terminal modification to increase the stability and binding affinity of the molecule. The reported side effects of Mipomersen are mild-to-moderate reactions at the injection-site and flu-like symptoms, which are considered to be tolerable. If these side effects were as a result of the therapy type (as opposed to targeting Apo B-100), e.g. a limited innate viral immune response, it bodes well for the use of this technology in the tumour vasculature [104, 107, 108].

Such oligonucleotide-based therapies have been used to target ncRNAs [109] and could be developed to inhibit pathological angiogenesis and administered in conjunction with conventional chemotherapeutics (though not necessarily through the same delivery method). Especially considering all that is required is the production of a complementary oligonucleotide, which is trivial compared to the production of a specific antibody or small-molecule drugs. One concern is that the oligonucleotide-based therapeutics must enter into the cell to have the desired effect and intravenous mechanisms of delivery allow for the oligonucleotides enter cells indiscriminately, which could increase the likelihood of negative side effects. However, this would not necessarily be problematic when targeting ncRNAs such as LOC643733, which was predicted to be tumour endothelial cell-specific in this

study. The intrinsic non-toxic nature of RNA-based therapies indemnifies against toxic effects in cells that do not contain the target molecule (immunostimulation is a possibility, though not necessarily disadvantageous). Should LOC643733 be proven to promote pathological angiogenesis in RCCEC, the only effect of the oligonucleotide-based therapy would be to remove the effect of this promotion temporarily [104]. The systemic risk to healthy cells could be limited through the injection of the oligonucleotide-based therapy directly into the tumour. Moreover if technologies were developed to enable the therapeutic delivery of large RNA/DNA molecules [103], suppressor genes such as PCAT19 or H19 could be delivered into the tumour endothelium to reduce its ability to proliferate.

In conclusion, PCAT19 and the other non-protein coding transcripts identified in this study have the potential to be utilised in anti-angiogenic therapeutic interventions for solid tumours. At a minimum, the expression patterns of the transcripts identified within this thesis have contributed to the continuing understanding of cancer progression. The data in this thesis represents the first demonstration of lncRNA differential expression between tumour and healthy tissue associated endothelial cells, and the first time any ncRNAs have been shown to be specifically expressed in endothelial cells. Additionally, PCAT19 is the first endothelial cell specific ncRNA to have an experimentally demonstrated function in endothelial cell biology *in vitro*.

References

1. Algire, GH. Chalkley, HW. Vasculae reactions of normal and malignant tissues *in vivo*. I. Vascular reactions of mice to wounds and to normal and neoplastic transplants. *Journal of the National Cancer Institute*. **1945**. 6(1): 73-85.
2. Folkman, J. Tumor angiogenesis. Therapeutic implications. *New England Journal of Medicine*. **1971**. 285: 1182-1186.
3. Ribatti, D. Judah Folkman, a pioneer in the study of angiogenesis. *Angiogenesis*. **2008**. 11(1): 3-10.
4. Burrows, FJ. Thorpe, PE. Eradication of large solid tumors in mice with an immunotoxin directed against tumor vasculature. *Proceedings of the National Academy of Sciences*. **1993**. 90: 8996-9000.
5. Dougherty, GJ. Chaplin, DJ. Development of vascular disrupting agents. In: Vascular disruptive agents for the treatment of cancer. Edited by: Meyer T. New York, USA: Springer Science. **2010**. 1-30.
6. Al-Huseinm, B. Abdalla, M. Trepre, M. DeRemer, CL. Somanath, PR. Anti-angiogenic therapy for cancer: An update. *Pharmacotherapy*. **2012**. 32(12):1095-1111.
7. Bridges, EM. Harris, AL. The angiogenic process as a therapeutic target in cancer. *Biochemical Pharmacology*. **2011**. 81: 1183-1191.
8. St. Croix, B. Rago, C. Velculescu, V. Traverso, G. Romans, KE. Montgomery, E. *et al*. Genes expressed in human tumor endothelium. *Science*. **2000**. 289(5482): 1197-1202.
9. Neri, D. Bicknell, R. Tumour vascular targeting. *Nature Reviews Cancer*. **2005**. 5(6): 436-446.

10. Heath, VL. Bicknell, R. Anticancer strategies involving the vasculature. *Nature Reviews Clinical Oncology*. **2009**. 6(7): 395-404.
11. Fukumura, D. Xu, L. Chen, Y. Gohongi, T. Seed, B. Jain, RK. Hypoxia and acidosis independently up-regulate Vascular Endothelial Growth Factor transcription in brain tumors in *vivo*. *Cancer Research*. **2001**. 61: 6020-6024.
12. Helmlinger, G. Yuan, F. Dellian, M. Jain, RK. Interstitial pH and pO2 gradients in solid tumors *in vivo*: High-resolution measurements reveal a lack of correlation. *Nature Medicine*. **1997**. 3(2): 177-182.
13. Ando, J. Yamamoto, K. Vascular Mechanobiology - Endothelial cell responses to fluid shear stress. *Circulation*. **2009**. 73: 1983-1992.
14. Zhang, H-T. Gorn, M. Smith, K. Graham, AP. Lau, KKW. Bicknell, R. Transcriptional profiling of human microvascular endothelial cells in the proliferative and quiescent state using cDNA arrays. *Angiogenesis*. **1999**. 3(3): 211-219.
15. Ahmed, Z. Bicknell, R. Angiogenic signaling pathways. In: Angiogenesis protocols. 2nd ed. Edited by Martin, S. Murray, C. New York, USA: Humana Press. **2009**. 3-24.
16. Arroyo, AG. Iruela-Arispe, ML. Extracellular matrix, inflammation, and the angiogenic response. *Cardiovascular Research*. **2010**. 86 (2): 226-235.
17. Ferrara, N. Kerbel, RS. Angiogenesis as a therapeutic target. *Nature*. **2005**. 438: 967-974.
18. Risau, W. Mechanisms of angiogenesis. *Nature*. **1997**. 671: 386-390.
19. Adams, RH. Eichmann, A. Axon guidance molecules in vascular patterning. *Cold Spring Harbor Perspectives in Biology*. **2010**. 2(5): a001875-a.
20. Morris, KV. Mattick, JS. The rise of regulatory RNA. *Nature Reviews Genetics*. **2014**. 15(6): 423-437.

21. Mattick, JS. RNA regulation: a new genetics? *Nature Reviews Genetics*. **2004**. 5: 316-323.

22. Frith, MC. Wilming, LG. Forrest, A. Kawaji, H. Tan, SL. Wahlestedt, C. *et al*. Pseudo-messenger RNA: Phantoms of the transcriptome. *PLoS Genetics*. **2006**. 2(4): 504-514.

23. The RNA Central Consortium. http://www.rnacentral.org. **2015**. Last Accessed: 01/07/2015.

24. Poliseno, P. Salmena, L. Zhang, J. Carver, B. Haveman, WJ. Pandolfi, PP. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature*. **2010**. 465: 1033-1038.

25. He, L. Hannon, GJ. MicroRNAs: Small RNAs with a big role in gene regulation. *Nature Review Genetics*. **2004**. 5: 522-532.

26. Kim, VN. MicroRNA biogenesis: coordinated cropping and dicing. Nature *Reviews Molecular Cell Biology*. **2005**. 6: 376-385.

27. Han, YJ. Ma, SF. Yourek, G. Park, YD. Garcia, JG. A transcribed pseudogene of MYLK promotes cell proliferation. *FASEB Journal.* **2011**. 25(7) 2305-2312.

28. Ebert, MS. Sharp, PA. Emerging roles of natural microRNA sponges. *Current Biology*. **2010**. 20: 858-861.

29. Hirotsune, S. Yoshida, N. Chen, A. Garrett, L. Sugiyama, F. Takahashi, S. Yagami, K-I. Wynshaw-Boris, A. Yoshiki, A. An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature*. **2003**. 423(6935): 91-6.

30. McCall, MN. Kent, OA. Yu, J. Fox-Talbot, K. Zaiman, AL. Halushka, MK. MicroRNA profiling of diverse endothelial cell types. *BMC Medical Genomics*. **2011**. 4(1): 78.

31. Huminiecki, L. In Silico cloning of novel endothelial-specific genes. *Genome Research*. **2000**. 10(11): 1796-1806.

32. Huminiecki, L. Gorn, M. Suchting, S. Poulsom, R. Bicknell, R. Magic roundabout Is a new member of the roundabout receptor family that is endothelial specific and expressed at sites of active angiogenesis. *Genomics*. **2002**. 79(4): 547-552.

33. Park, KW. Morrison, CM. Sorensen, LK. Jones, CA. Rao, Y. Chien, CB. *et al*. Robo4 is a vascular-specific receptor that inhibits endothelial migration. *Developmental Biology*. **2003**. 261(1): 251-67.

34. Zhuang, X. Ahmed, F. Zhang, Y. Ferguson, HJ. Steele, JC. Steven, NM. Nagy, Z. Heath, VL. Toellner, KM. Bicknell, R. Robo4 vaccines induce antibodies that retard tumor growth. *Angiogenesis*. **2015**. 18(1): 83-95.

35. Yoshikawa, M. Mukai, Y. Okada, Y. Tsumori, Y. Tsunoda, S-I. Tsutsumi, Y. *et al*. Robo4 is an effective tumor endothelial marker for antibody-drug conjugates based on the rapid isolation of the anti-Robo4 cell-internalizing antibody. *Blood*. **2013**. 121(14): 2804-2813.

36. Herbert, JMJ. Stekel, D. Sanderson, S. Heath, VL. Bicknell, R. A novel method of differential gene expression analysis using multiple cDNA libraries applied to the identification of tumour endothelial genes. *BMC Genomics*. **2008**. 9: 1471-2164.

37. Herbert, JMJ. Stekel, D. Mura, M. Sychev, M. Bicknell, R. Bioinformatic methods for finding differentially expressed genes in cDNA Libraries, applied to the identification of tumour vascular targets. In: cDNA Libraries. Edited by Lu, C. Browse, J. Wallis, JG. cDNA Libraries. Humana Press. **2011**. 99-119.

38. Chaudhary, A. Hilton, MB. Seaman, S. Haines, DC. Stevenson, S. Lemotte, PK. *et al*. TEM8/ANTXR1 blockade inhibits pathological angiogenesis and potentiates tumoricidal responses against multiple cancer types. *Cancer Cell*. **2012**. 21(2): 212-226.

39. Phillips, DD. Fattah, RJ. Crown, D. Zhang, Y. Liu, S. Moayeri, M. *et al*. Engineering anthrax toxin variants that exclusively form octamers, and their application to targeting tumors. *Journal of Biological Chemistry*. **2013**. 288(13): 9058-9065.

40. Cryan, LM. Rogers, MS. Targeting the anthrax receptors, TEM-8 and CMG-2, for anti-angiogenic therapy. *Frontiers in Bioscience*. **2011**. 16: 1574-1588.

41. Simpson, JC. Wellreuther, R. Poustka, A. Pepperkok, R. Wiemann, S. Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. *EMBO Reports*. **2000**. 1(3): 287-292.

42. Wang, Z. Mark, G. Michael, S. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*. **2009**. 10: 57-64.

43. Mortazavi, A. Williams, BA. McCue, K. Schaeffer, L. Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*. **2008**. 5(7): 621-628.

44. Ghilardi, C. Chiorino, G. Dossi, R. Nagy, Z. Giavazzi, R. Bani, M. Identification of novel vascular markers through gene expression profiling of tumor-derived endothelium. *BMC Genomics*. **2008**. 9(1): 201.

45. Ho M, Yang E, Matcuk G, Deng D, Sampas N, Tsalenko A, *et al*. Identification of endothelial cell genes by combined database mining and microarray analysis. *Physiological Genomics*. **2003**. 13: 249-262.

46. Chi, J-T. Chang, HY. Haraldsen, G. Jahnsen, FL. Troyanskaya, OG. Chang, DS. *et al*. Endothelial cell diversity revealed by global expression profiling. *PNAS*. **2003**. 100(19): 10623-19628.

47. Conway, EM. Carmeliet, P. The diversity of endothelial cells: a challenge for therapeutic angiogenesis. *Genome Biology*. **2004**. 5(7): 207.

48. Aird, WC. Endothelial Cell Heterogeneity. *Cold Spring Harbor Perspectives in Medicine*. **2011**. 2(1): a006429.

49. Liu, L. Li, Y. Li, S. Hu, N. He, Y. Pong, R. *et al*. Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology*. **2012**: 1-11.

50. Branton, D. Deamer, DW. Marziali, A. Bayley, H. Benner, SA. Butler, T. *et al*. The potential and challenges of nanopore sequencing. *Nature Biotechnology*. **2008**. 26(10): 1146-53.

51. Mardis, ER. Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*. **2008**. 9(1): 387-402.

52. Glenn, TC. Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*. **2011**. 11(5): 759-69.

53. Herbert JHJ. Endothelial cell gene expression [PhD Thesis]. [Birmingham, UK]: University of Birmingham. **2012**. 228.

54. Mura, M. Swain, RK. Zhuang, X. Vorschmitt, H. Reynolds, G. Durant, S. *et al*. Identification and angiogenic role of the novel tumor endothelial marker CLEC14A. *Oncogene*. **2012**. 31: 293-305.

55. Maciag, T. Cerundolo, J. Ilsley, S. Kelley, PR. Forand, R. An endothelial cell growth factor from bovine hypothalamus: identification and partial characterization. *Proceedings of the National Academy of Sciences of the United States of America*. **1979**. 76(11): 5674-5678.

56. Reynolds, A. Leake, D. Boese, Q. Scaringe, S. Marshall, WS. Khvorova, A. Rational siRNA design for RNA interference. *Nature Biotechnology*. **2004**. 22(3): 326-330

57. Rodriguez-Bigas, MA. Lin, EH. Crane, CH. Adenocarcinoma of the Colon and Rectum. In: Cancer Medicine. 6th ed. Edited by Kufe, DW. Pollock, RE. Weichselbaum, RR.

Bast, RC. Gansler, TS. Holland, JF. *et al*. Hamilton, CA: BC Decker. **2003**. Available from: http://www.ncbi.nlm.nih.gov/books/NBK13270. Last Accessed: 20/01/2016.

58. Ferguson, HJ. Wragg, J. Ismail, T. Bicknell, R. Vaccination against tumour blood vessels in colorectal cancer. *European Journal of Surgical Oncology*. **2014**. 40(2): 133-136.
59. Konerding, MA. Fait, E. Gaumann, A. Dimitropoulou, C. Malkush, W. The vascularization of experimental and human primary tumors: comparative morphometric and morphologic studies. In: Angiogenesis models modulators and clinical applications. Edited by Maragoudakis, ME. New York: Plenum Press. **1998**.
60. Nature Publishing Group. Byte-ing off more than you can chew. *Nature Methods*. **2008**. 5(7): 577.
61. Auer, PL. Doerge, RW. Statistical design and analysis of RNA sequencing data. *Genetics*. **2010**. 185(2): 405-16.
62. Peng, JC. Li, KK. Lau, KM. Ng, YL. Wong, J. Chung, NY. *et al*. KIAA0495/PDAM Is frequently downregulated in oligodendroglial tumors and Its knockdown by siRNA induces cisplatin resistance in glioma cells. *Brain Pathology*. **2010**. 20(6): 1021-1032.
63. Flores, ER. Sequpta, S. Miller, JB. Newman, JJ. Bronson, R. Crowley, D. Yang, A. McKeon, F. Jacks, T. Tumor predisposition in mice mutant for p63 and p73: Evidence for broader tumor suppressor functions for the p53 family. *Cancer Cell*. **2005**. 7(4): 363-373.
64. Office for National Statistics. Births and Deaths in England and Wales, 2010 (Statistical Bulletin). Office for National Statistics: London, UK. **2011**.
65. Ferlay, J. Shin, H. Bray, F. Forman, D. Mathers, C. Parkin, D. GLOBOCAN 2008 v1.2, Cancer incidence and mortality worldwide: IARC CancerBase No. 10. International agency for research on cancer: Lyon, France. **2010**.

66. Nair, A. Klusmann, MJ. Jogeesvaran, KH. Grubnic, S. Green, SJ. Vlahos, I. Revisions to the TNM staging of non-small cell lung cancer: rationale, clinicoradiologic implications, and persistent limitations. *Radiographics*. **2011**. 31: 215–238.

67. Goldstraw, P. Ball, D. Jett, JR. Chevalier, TL. Lim, E. Nicholson, AG. Shepherd, FA. Non-small-cell lung cancer. *Lancet*. **2011**. 378: 1727-1740.

68. Johnson, DH. Fehrenbacher, L. Novotny, WF. Herbst, RS. Nemunaitis, JJ. Jablons, DM. *et al*. Randomized phase II trial comparing bevacizumab plus carboplatin and paclitaxel with carboplatin and paclitaxel alone in previously untreated locally advanced or metastatic non-small-cell lung cancer. *Journal of Clinical Oncology*. **2004**. 22: 2184-2191.

69. Esteller, M. Non-coding RNAs in human disease. *Nature Reviews Genetics*. **2011**. 12: 861-874.

70. Darzacq, X. Jady, BE. Verheggen, C. Kiss, AM. Bertrnad, E. Kiss, T. Cajal body-specific small nuclear RNAs: a novel class of 2'-O-methylation and pseudouridylation guide RNAs. *EMBO Journal*. **2002**. 21: 2746-2756.

71. Matera, AG. Terns, RM. Terns, MP. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nature Reviews Molecular Cell Biology*. **2007**. 8: 209-20.

72. Williams, GT. Farzaneh, F. Are snoRNAs and snoRNA host genes new players in cancer? *Nature Reviews Cancer*. **2012**. 12: 84-88.

73. Liao, J. Yu, L. Mei, Y. Guarnera, M. Shen, J. Li, R. Liu, Z. Jiang, F. Small nucleolar RNA signatures as biomarkers for non-small-cell lung cancer. *Molecular Cancer*. **2010**. 9: 198.

74. Su, H. Xu, T. Ganapathy, S. Shadfan, M. Long, M. Huang, TH-M. Thompson, I. Yuan, Z-M. Elevated snoRNA biogenesis is essential in breast cancer. *Oncogene*. **2014**. 22: 1348-1358.

75. Valleron, W. Laprevotte, E. Guatier, E-F. Quelen, C. Demur, C. Delabesse, E. Agirre, X. Prósper, F. Kiss, T. Brousset P. Specific small nucleolar RNA expression profiles in acute leukemia. *Leukemia*. **2012**. 26: 2052-2060.

76. Chaudhry, MA. Expression pattern of small nucleolar RNA host genes and long non-coding RNA in X-rays-treated lymphoblastoid cells. *International Journal of Molecular Sciences*. **2013**. 14(5): 9009-9110.

77. Gee, HE. Buffa, FM. Camps, C. Ramachandran, A. Leek, R. Taylor, M. *et al*. The small-nucleolar RNAs commonly used for microRNA normalisation correlate with tumour pathology and prognosis. *British Journal of Cancer*. **2011**. 104: 1168-1177.

78. Kino, T. Hurt, DE. Ichijo, T. Nader, N. Chrousos, GP. Noncoding RNA Gas5 is a growth arrest and starvation associated repressor of the glucocorticoid receptor. *Science Signaling*. **2010**. 3: 1-33.

79. Lestrade, L. Weber, MJ. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Research*. **2004**. 34, D158-162.

80. Dudley, AC. Tumor Endothelial Cells. *Cold Spring Harbour Perspectives in Medicine*. **2012**. 2: 2-18.

81. Zhuang, X. Herbert, JM. Lodhia, P. Bradford, J. Turner, AM. Newby, PM. *et al*. Identification of novel vascular targets in lung cancer. *British Journal of Cancer*. **2015**. 112 (3): 485-94.

82. Trapnell, C. Roberts, A. Goff, L. Pertea, G. Kim, D. Kelley, DR. Pimentel, H. Saizberg, SL. Rinn, JL. Pachter, L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*. **2012**. 7: 562-578.

83. Yilmaz, E. Koyuncuoglu, M. Gorken, IB. Okyay, E. Saatli, B. Ulukus, EC. Saygili, U. *Journal of Gynecologic Oncology*. **2011**. 22(2): 89-96.

84. Lurlaro, M. Loverro, G. Vacca, A. Cormio, G. Ribatti, D. Minischetti, M. Ria, R. Bruno, M. Selvaggi, L. Angiogenesis extent and expression of matrix metalloproteinase-2 and -9 correlate with upgrading and myometrial invasion in endometrial carcinoma. *European Journal of Clinical Investigation*. **1999**. 29(9): 793-801.

85. Heidegger, I. Kem, J. Ofer, P. Klocker, H. Massoner, P. Oncogenic functions of IGF1R and INSR in prostate cancer include enhanced tumor growth, cell migration and angiogenesis. *Oncotarget*. **2014**. 5(9): 2723-35.

86. Wang, J. Shiozawa, Y. Wang, J. Wang, Y. Jung, Y. Plenta, JK. Mehra, R. Loberg, R. Taichman, RS. The Role of CXCR7/RDC1 as a Chemokine Receptor for CXCL12/SDF-1 in Prostate Cancer. *Journal of Biological Chemistry*. **2007**. 283:4283-4294.

87. Kim, S. Dobi, E. Jary, M. Monnien, F. Curtit, E. Nguyen, T. *et al*. Bifractionated CPT-11 with LV5FU2 infusion (FOLFIRI-3) in combination with bevacizumab: clinical outcomes in first-line metastatic colorectal cancers according to plasma angiopoietin-2 levels. *BMC Cancer*. **2013**. 13: 611.

88. Alderton, GK. Therapy: confused? Absolutely. *Nature reviews cancer*. **2010**. 10:316-317.

89. Fingleton, B. MMPs as therapeutic targets – still a viable option? *Seminars in Cell and Developmental Biology*. **2008**. 19(1): 61-68.

90. Huang, H. Bhat, A. Woodnutt, G. Lappe, R. Targeting the ANGPT-Tie2 pathway in malignancy. *Nature Reviews Cancer*. **2010**. 10: 575-585.

91. Rigoutsos, I. Furnari, F. Gene-expression forum: Decoy for microRNAs. *Nature*. **2010**. 465: 1016-1017.

92. Ayesh, S. Matouk, I. Schneider, T. Ohana, P. Laster, M. Al-Sharef, W. Groot, ND. Hochberg, A. Possible physiological role of H19 RNA. *Molecular Carcinogenesis*. **2002**. 35(2): 63-74.

93. Matouk, I. Ohana, P. Ayesh, S. Sidi, A. Czerniak, A. Groot, ND. Hockberg, A. The oncofetal H19 RNA in human cancer, from the bench to the patient. *Cancer Therapy*. **2005**. 3: 249-266.

94. Gao, Y. Zhu, M. Wang, H. Zhao, S. Zhao, D. Yang, Y. *et al*. Association of polymorphisms in long non-coding RNA H19 with coronary artery disease risk in a Chinese population. *Mutation Research*. **2015**. 772: 15-22.

95. Sample, KM. Bicknell, R. Using RNAseq to identify anti-cancer targets in the tumour vasculature. *European Pharmaceutical Review*. **2014**. 18: 43-47.

96. Al Olama, AA. Kote-Jarai, Z. Schumacher, FR. Wiklund, F. Berndt, SI. Benlloch, S. *et al*. A meta-analysis of genome-wide association studies to identify prostate cancer susceptibility loci associated with aggressive and non-aggressive disease. *Human Molecular Genetics*. **2013.** 22: 408-415.

97. Shui, IM. Lindström, S. Kibel, AS. Berndt, SI. Campa, D. Gerke, T. *et al*. Prostate cancer (PCa) risk variants and risk of fatal PCa in the National Cancer Institute Breast and Prostate Cancer Cohort Consortium. *European Urology*. **2014**. 65: 1069-1075.

98. Kerbel, RS. Tumor angiogenesis: past, present and the near future. *Carcinogenesis*. **2000**. 21: 505-515.

99. Wragg, JW. Durant, S. McGettrick, HM. Sample, KM. Egginton, S. Bicknell, R. Shear stress regulated gene expression and angiogenesis in vascular endothelium. *Microcirculation*. **2014**. 21: 290-300.

100. Lee, YH. Juo, CY. Stark, JM. Shih, HM. Ann, DK. HP1 promotes tumor suppressor BRCA1 functions during the DNA damage response. *Nucleic Acids Research*. **2013**. 41(11): 5784-5798.

101. Chu, L. You, Y. Liu, X. Thomas, K. Jiang, H. Zhu, T. *et al*. The spatiotemporal dynamics of chromatin protein HP1α is essential for accurate chromosome segregation during cell division. *Journal of Biological Chemistry*. **2014**. 289: 26249-26262.

102. Scopus. http://www.scopus.com. **2015**. Last Accessed: 01/07/2015.

103. Poller, W. Tank, J. Skurk, C. Gast, M. Cardiovascular RNA interference therapy. *Circulation Research*. **2013**. 113: 588-602.

104. Bennett, CF. Swayze, EE. RNA targeting therapeutics: molecular mechanisms of antisense oligonucleotides as a therapeutic platform. *Pharmacology and toxicology*. **2010**. 50: 259-293.

105. Burnett, JC. Rossi, JJ. RNA-based therapeutics: current progress and future prospects. *Chemistry and Biology*. **2012**. 19 (1): 60-71.

106. McIvor, RS. Therapeutic Delivery of mRNA: The Medium Is the Message. *Molecular Therapy*. **2011**. 19(5): 822-823.

107. Tabernero, J. Shapiro, GI. LoRusso, PM. Cervantes, A. Schwartz, GK. Weiss, GJ. Paz-Ares, L. Cho, DC. Infante, JR. *et al*. First-in-Humans trial of an RNA interference therapeutic targeting VEGF and KSP in cancer patients with liver involvement. *Cancer Discovery*. **2013**. 3(4): 406-417.

108. Toth, PP. Emerging LDL therapies: Mipomersen-antisense oligonucleotide therapy in the management of hypercholesterolemia. *Journal of Clinical Lipidology*. **2013**. 7(3): S6-S10.

109. Braconi, C. Patel, T. Non-coding RNAs as therapeutic targets in hepatocellular cancer. *Current cancer drug targets*. **2014**. 12(9): 1073-1080.

Appendix 1: Table of materials

Item	Source	Protocol
Collagenase type V	Sigma, UK	Endothelial cell isolation
Streptavidin-coated dynabeads	Invitrogen, UK	Endothelial cell isolation
Biotinylated Ulex lectin	Vector labs, US	Endothelial cell isolation
Hepatocellular carcinoma	Queen Elizabeth Hospital, Birmingham	Endothelial cell isolation
Healthy liver tissue	Queen Elizabeth Hospital, Birmingham	Endothelial cell isolation
Umbilical Cords	Birmingham Women's Hospital, Birmingham	HUVEC isolation
Collagenase type Ia	Sigma, UK	HUVEC isolation
Porcine skin gelatine	Sigma, UK	Tissue culture
Medium 199	Sigma, UK	Tissue culture
Foetal calf serum	PAA, The Cell Culture Co., UK	Tissue culture
Heparin	Sigma, UK	Tissue culture
L-glutamine	Sigma, UK	Tissue culture
Penicillin-Streptomycin	Sigma, UK	Tissue culture
0.22 µm pore filters	Millipore, DE	Tissue culture
Corning 10 cm plates	Sigma, UK	Tissue culture
PBS	Sigma, UK	Tissue culture, PI staining
Trypsin-EDTA	Sigma, UK	Tissue culture
Dermal Fibroblasts	Promocell, DE	Tissue culture

Keratinocytes	TCS Cellworks , UK	Tissue culture
Human aortic smooth muscle cells	TCS Cellworks , UK	Tissue culture
HEK293T	Dr Mike Tomlinson, University of Birmingham	Tissue culture
DMEM	Sigma, UK	Tissue culture
RNeasy mini kit	Qiagen, UK	RNA isolation
2-mercaptoethanol	Sigma-Aldrich, UK	RNA isolation
Ethanol	Sigma, UK	RNA isolation, PI Staining
DNase I	Qiagen, UK	RNA isolation
NanoDrop (NDT1000) Spectrophotometer	LabTech,"UK	RNA isolation and cloning
High-capacity cDNA archive kit	Invitrogen, UK	cDNA production
Primers	Eurogentec, UK	qPCR, PCR
2x SYBR Green qPCR mix	Invitrogen, UK	qPCR
2x Universal qPCR Mastermix	Invitrogen, UK	qPCR
Nano Pure H2O	n/a	qPCR, PI staining.
Universal Probe Library set, Human	Roche Applied Science, UK	qPCR
Phusion High-Fidelity DNA Polymerase	NEB, UK	PCR
dNTP Mix	Bioline, UK	PCR
SYBR Safe DNA Gel stain	Invitrogen, UK	PCR, Plasmid preparation
Rotor-Gene RG-3000 qPCR machine	Corbett Research Ltd, Australia	qPCR

Propidium iodide (PI)	Invitrogen, UK	PI Staining
Ribonuclease A	Sigma, UK	PI Staining
Bovine serum albumin	Sigma, UK	PI Staining
Triton-X-100	BDH, US	PI Staining
Gibson Assembly Cloning Kit	NEB, UK	Molecular cloning
Plasmid Mega Kit	Qiagen, NL	Plasmid preparation
GeneJET plasmid miniprep kit	Fermentas, US	Plasmid preparation
GeneJET gel extraction kit	Fermentas, US	Plasmid preparation
PmeI	NEB, UK	Plasmid preparation
pWPI	Addgene, US	Molecular cloning, lentiviral transduction
psPAX2	Addgene, US	Lentiviral transduction
pMD2G	Addgene, US	Lentiviral transduction
Opti-MEM	Invitrogen, UK	Lentiviral transduction, siRNA knockdown
Human Embryonic Kidney 293 cells	ATCC, UK	Lentiviral transduction
Polybrene	Sigma, UK	Lentiviral transduction
siRNA Duplexes	Eurogentec, UK	siRNA knockdown
Lipofectamine RNAiMAX	Invitrogen, UK	siRNA knockdown
Whole Human Genome Microarray Kit, 4x44K	Agilent, US	Microarray analysis
Quick Amp Labeling Kit, Two-Color	Agilent, US	Microarray analysis

Table S1: Table of materials

Appendix 2: Renal cell carcinoma patient information

ID	Age	Gender	Tumour Type	Tumour Grade	Tumour Stage
1	70	Male	Clear cell (cystic)	1	pT2a Nx
2	69	Female	Clear cell	2	pT1a Nx
3	77	Female	Clear cell	2	pT2b Nx

Table S2: Clinical and-pathological data for patients in the RNAseq analysis

The tumour grade was recorded using the Fuhrman scale of I-IV, where patients with grade I tumours carry the best prognosis and grade IV the worst. The tumour staging is recorded using standard nomenclature:

p - Primary tumour (p)

T1a - limited to kidney <4 cm

T2a - limited to kidney, >7 cm but not more than 10 cm

T2b - limited to kidney, >10 cm

Nx - Regional lymph nodes cannot be evaluated.

Appendix 3: Kidney RNAseq analysis codes

A3.1 Quality Checking reads from RNAseq data using FastQC

```
# Set shell variables

%%cd

read_file_1=Kidney_NGS_Data/RCCEC3/RCCEC3_L1.fq

read_file_1P=Kidney_NGS_Data/RCCEC3/RCCEC3_L1_P.fq

read_file_2=Kidney_NGS_Data/RCCEC3/RCCEC3_L2.fq

read_file_2P=Kidney_NGS_Data/RCCEC3/RCCEC3_L2_P.fq

read_file_3=Kidney_NGS_Data/HKEC3/HKEC3_L1.fq

read_file_3P=Kidney_NGS_Data/HKEC3/HKEC3_L1_P.fq

read_file_4=Kidney_NGS_Data/HKEC3/HKEC3_L2.fq

read_file_4P=Kidney_NGS_Data/HKEC3/HKEC3_L2_P.fq

read_file_5=Kidney_NGS_Data/WNK3/WNK3_L1.fq

read_file_5P=Kidney_NGS_Data/WNK3/WNK3_L1_P.fq

read_file_6=Kidney_NGS_Data/WNK3/WNK3_L2.fq

read_file_6P=Kidney_NGS_Data/WNK3/WNK3_L2_P.fq


# Run FastQC

%%/FastQC/fastqc ${read_file_1} ${read_file_1P} ${read_file_2} ${read_file_2P}

${read_file_3} ${read_file_3P} ${read_file_4} ${read_file_4P} ${read_file_5}

${read_file_5P} ${read_file_6} ${read_file_6P}
```

A3.2 Map reads to the genome using Tophat

```
# Tophat is a splicing aware aligner

# Download the human genome

%%cd ~

%%mkdir /home/Klarke/Desktop/Genome_human

%%cd /home/Klarke/Desktop/Genome_human

%%curl -O -L

http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/chromFA.tar.gz

%%gunzip chromFA.tar.gz

%%tar -xvf chromFA.tar.gz

# To remove everything except chr1-22.fa, chrX.fa, chrY.fa and chrM.fa

%%rm chr*_*.fa

# Now concatenate (join together) all the files

cat chr*.fa>hg19.fa


# Index the genome

%%cd /home/Klarke/Desktop/Genome_human

%%bowtie2-build hg19.fa hg19


%%cd ~

%%mkdir /homeKlarke/Desktop/Tophat_Data

%%mkdir /home/Klarke/Desktop/NGS_Data
```

```
%%cd /home/Klarke/Desktop/NGS_Data

%%printf

'reference=/home/Klarke/Desktop/Genome_human/hg19\nreads_1=RCCEC3_L1\

nreads_1P=RCCEC3_L1_P\nreads_2=RCCEC3_L2\nreads_2P=RCCEC3_L2_P\

nreads_3=HKEC3_L1\nreads_3P=HKEC3_L1_P\nreads_4=HKEC3_L2\nreads_4

P=HKEC3_L2_P\nreads_5=WNK3_L1\nreads_5P=WNK12_L1_P\nreads_6=WNK

3_L2\nreads_6P=WNK3_L2_P\ntophat2 -o ${reads_1}_tophat_out ${reference}

${reads_1}.fq ${reads_1P}.fq\ntophat2 -o ${reads_2}_tophat_out ${reference}

${reads_2}.fq ${reads_2P}.fq\ntophat2 -o ${reads_3}_tophat_out ${reference}

${reads_3}.fq ${reads_3P}.fq\ntophat2 -o ${reads_4}_tophat_out ${reference}

${reads_4}.fq ${reads_4P}.fq\ntophat2 -o ${reads_5}_tophat_out ${reference}

${reads_5}.fq ${reads_5P}.fq\ntophat2 -o ${reads_6}_tophat_out ${reference}

${reads_6}.fq ${reads_6P}.fq\nmv *_tophat_out*

/home/Klarke/Desktop/Tophat_Data' > tophat2.txt


# tophat2.txt will contain the following:

reference=/home/Klarke/Desktop/Genome_human/hg19

reads_1=RCCEC3_L1

reads_1P=RCCEC3_L1_P

reads_2=RCCEC3_L2

reads_2P=RCCEC3_L2_P

reads_3=HKEC3_L1

reads_3P=HKEC3_L1_P
```

```
reads_4=HKEC3_L2

reads_4P=HKEC3_L2_P

reads_5=WNK3_L1

reads_5P=WNK3_L1_P

reads_6=WNK3_L2

reads_6P=WNK3_L2_P

tophat2 -o ${reads_1}_tophat_out ${reference} ${reads_1}.fq ${reads_1P}.fq

tophat2 -o ${reads_2}_tophat_out ${reference} ${reads_2}.fq ${reads_2P}.fq

tophat2 -o ${reads_3}_tophat_out ${reference} ${reads_3}.fq ${reads_3P}.fq

tophat2 -o ${reads_4}_tophat_out ${reference} ${reads_4}.fq ${reads_4P}.fq

tophat2 -o ${reads_5}_tophat_out ${reference} ${reads_5}.fq ${reads_5P}.fq

tophat2 -o ${reads_6}_tophat_out ${reference} ${reads_6}.fq ${reads_6P}.fq

mv *_tophat_out* /home/Klarke/Desktop/Tophat_Data


# Unprotect the file

%%chmod a+rw tophat2.txt

# Make the text file excecutable

%%chmod u+x tophat2.txt


# Map the paired ends to the genome

%%sh tophat2.txt
```

You can also include the -r option (%%tophat2 -r 300 ${reference} ${reads_1} ${reads_2}), which indicates the distance between the paired-end reads

therefore if the reads are 90bp and the fragment length is 480bp the code should be -r 300

A3.3 Differential expression analysis using Cufflinks

```
%%cd
%%reference=/home/Klarke/Desktop/Genome_human/hg19.fa
%%printf
'mapped_reads_1=RCCEC3_L1\nmapped_reads_2=RCCEC3_L2\nmapped_reads_3=HKEC3_L1\nmapped_reads_4=HKEC3_L2\nmapped_reads_5=WNK3_L1\nmapped_reads_6=WNK3_L2\nmkdir Desktop/Cufflinks_Data\ncufflinks -o cufflinks_${mapped_reads_1} Desktop/Tophat_Data/${mapped_reads_1}_tophat_out/accepted_hits.bam\ncufflinks -o cufflinks_${mapped_reads_2} Desktop/Tophat_Data/${mapped_reads_2}_tophat_out/accepted_hits.bam\ncufflinks -o cufflinks_${mapped_reads_3} Desktop/Tophat_Data/${mapped_reads_3}_tophat_out/accepted_hits.bam\ncufflinks -o cufflinks_${mapped_reads_4} Desktop/Tophat_Data/${mapped_reads_4}_tophat_out/accepted_hits.bam\ncufflinks -o cufflinks_${mapped_reads_5} Desktop/Tophat_Data/${mapped_reads_5}_tophat_out/accepted_hits.bam\ncufflinks
```

```
ks -o cufflinks_${mapped_reads_6}
Desktop/Tophat_Data/${mapped_reads_6}_tophat_out/accepted_hits.bam\nmv
*cufflinks_* Desktop/Cufflinks_Data' > cufflinks.txt

# cufflinks.txt will contain the following:
#reference=/home/Klarke/Desktop/Genome_human/hg19.fa
#mapped_reads_1=RCCEC3_L1
#mapped_reads_2=RCCEC3_L2
#mapped_reads_3=HKEC3_L1
#mapped_reads_4=HKEC3_L2
#mapped_reads_5=WNK3_L1
#mapped_reads_6=WNK3_L2
#mkdir Desktop/Cufflinks_Data
#cufflinks -o cufflinks_${mapped_reads_1}
#Desktop/Tophat_Data/${mapped_reads_1}_tophat_out/accepted_hits.bam
#cufflinks -o cufflinks_${mapped_reads_2}
#Desktop/Tophat_Data/${mapped_reads_2}_tophat_out/accepted_hits.bam
#cufflinks -o cufflinks_${mapped_reads_3}
#Desktop/Tophat_Data/${mapped_reads_3}_tophat_out/accepted_hits.bam
#cufflinks -o cufflinks_${mapped_reads_4}
#Desktop/Tophat_Data/${mapped_reads_4}_tophat_out/accepted_hits.bam
#cufflinks -o cufflinks_${mapped_reads_5}
#Desktop/Tophat_Data/${mapped_reads_5}_tophat_out/accepted_hits.bam
```

```
#cufflinks -o cufflinks_${mapped_reads_6}
#Desktop/Tophat_Data/${mapped_reads_6}_tophat_out/accepted_hits.bam
#mv *cufflinks_* Desktop/Cufflinks_Data

# Make the text file executable
%%chmod u+x cufflinks.txt

# Now we can assemble all of your data (this could take 12 hours)
%%sh cufflinks.txt
# Download the reference annotation from dropbox (up to date as of 05/2015):
%% cd Desktop/Genome_human
%% curl -O -L
https://www.dropbox.com/s/hx34okisl7p7dfe/humanG19_annotation.gtf
%%annotation=/home/Klarke/Desktop/Genome_human/humanG19_annotation.gtf
%%reference=/home/Klarke/Desktop/Genome_human/hg19.fa
# Use these codes to create a text file containing the file paths for your data
%%cd Desktop/Cufflinks_Data
%%printf
'cufflinks_RCCEC3_L1/transcripts.gtf\ncufflinks_RCCEC3_L2/transcripts.gtf\ncuffli
nks_HKEC3_L1/transcripts.gtf\ncufflinks_HKEC3_L2/transcripts.gtf\ncufflinks_WN
K3_L1/transcripts.gtf\ncufflinks_WNK3_L2/transcripts.gtf' > Cuff_assemblies.txt

# Cuff_assemblies.txt will contain the following:
```

```
#cufflinks_RCCEC3_L1/transcripts.gtf

#cufflinks_RCCEC3_L2/transcripts.gtf

#cufflinks_HKEC3_L1/transcripts.gtf

#cufflinks_HKEC3_L2/transcripts.gtf

#cufflinks_WNK3_L1/transcripts.gtf

#cufflinks_WNK3_L2/transcripts.gtf


#Merge the assembled data

%%cuffmerge -g ${annotation} -s ${reference} Cuff_assemblies.txt


#Differential analysis with gene and transcript discovery

#Separate replicates with a comma and independent conditions with a space

%%cd

%%mkdir Desktop/Final_outputs/Patient3

%%mkdir Desktop/Final_outputs/Patient3/cuffdiff_out

%%cuffdiff -o Desktop/Final_outputs/Patient3/cuffdiff_out

Desktop/Cufflinks_Data/Patient3_merged_asm/merged.gtf -L RCCEC,HKEC,WNK

Desktop/Tophat_Data/${mapped_reads_1}_tophat_out/accepted_hits.bam,Deskto

p/Tophat_Data/${mapped_reads_2}_tophat_out/accepted_hits.bam

Desktop/Tophat_Data/${mapped_reads_3}_tophat_out/accepted_hits.bam,Deskto

p/Tophat_Data/${mapped_reads_4}_tophat_out/accepted_hits.bam

Desktop/Tophat_Data/${mapped_reads_5}_tophat_out/accepted_hits.bam,Deskto

p/Tophat_Data/${mapped_reads_6}_tophat_out/accepted_hits.bam
```

A3.4 Graphing with CummeRbund

Use the language “R”

%%R

R codes start with: “>” rather than “%%”

>library(cummeRbund)

>setwd ("Desktop/Final_outputs/Patientall/cuffdiff_out")

>cuff <- readCufflinks()

A3.5 Example 1: Create a heatmap

Set the location and size of the output image

>png(filename = '/home/Klarke/heatmapb.png')

To change the size use: >png(filename = '/home/Klarke/thinheatmap.png', width = 3000, height = 11000, units = 'px')

Choose the genes you want to use and transfer the gene_ids (XLOC numbers) to a txt file (one per line).

>myIDs<-read.table("Ids.txt")

>myIDs<-as.vector(myIDs$V1)

>is.vector(myIDs)

This command should output “true”, if false there is an error in the “myIDs” codes

>myGenesIds<-getGenes(cuff,myIDs)

```
>heat<-csHeatmap(myGenesIds,labCol=T, labRow=F,
logMode=T,cluster='both',replicates=F)
>heat
# Write the data to the location entered above
>dev.off()
```

A3.6 Example 2: Plot data phylogeny

```
# These codes will diagrammatically group the datasets based upon their similarity.
> dend<-csDendro(genes(cuff))
> dend
```

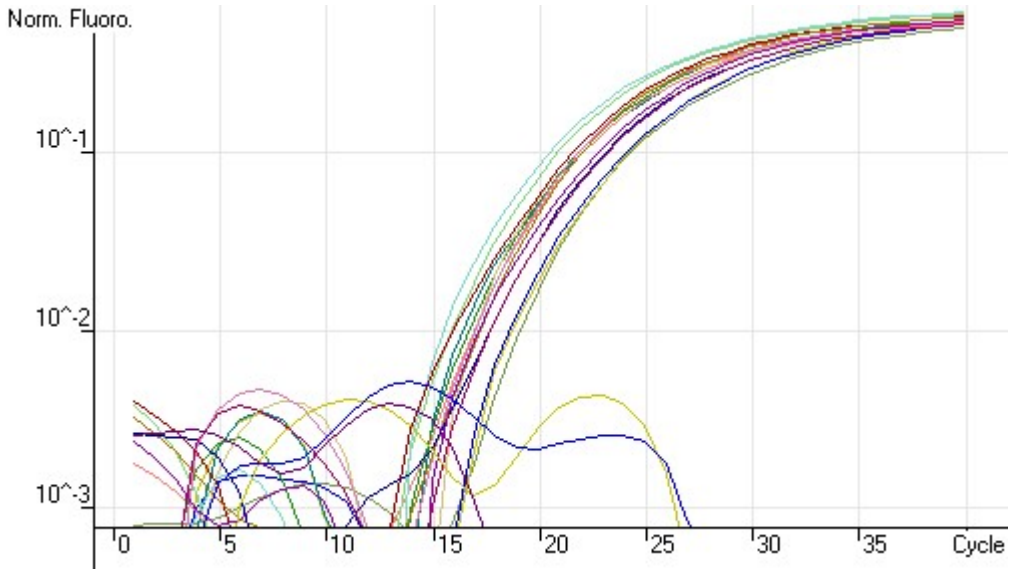
A3.7 Example 3: Cluster Analysis

```
# The following codes will list all the genes that have a similar expression pattern
to the query. The numbers in brackets represent the expression (reads) desired
(RCCEC, HKEC, WNK), they can be changed to find genes with different
expression profiles.
>myProfile<-c(0,100,0)
>mySimilar2<-findSimilar(cuff,myProfile,n=10)
>mySimilar2.expression<-
expressionPlot(mySimilar2,logMode=T,showErrorbars=F)
>mySimilar2.expression
```

A3.8 Example 4: Gene-level plots

```
# The following codes will plot the expression levels of PCAT19 (LOC100505495)
in the RNAseq data
>myGeneId<-"LOC100505495"
>myGeneId<-getGene(cuff,myGeneId)
>myGeneId
>head(fpkm(myGeneId))
>gl.rep<-
expressionPlot(myGeneId,replicates=TRUE,logMode=F,showErrorbars=F)
>gl.rep
```

Appendix 4: DNase digest qPCR control



No.	Colour	Name	Type
1		HUVEC 1:80 cDNA	Unknown
2		HUVEC 1:80 cDNA	Unknown
3		HUVEC 1:80 cDNA	Unknown
4		HUVEC 1:80 no cDNA	No Template Control
5		HUVEC 1:80 no cDNA	No Template Control
6		HUVEC 1:80 no cDNA	No Template Control
7		HLE 1:80 cDNA	Unknown
8		HLE 1:80 cDNA	Unknown
9		HLE 1:80 cDNA	Unknown
10		HLE 1:80 no cDNA	No Template Control
11		HLE 1:80 no cDNA	No Template Control
12		HLE 1:80 no cDNA	No Template Control
13		TLE 1:80 cDNA	Unknown
14		TLE 1:80 cDNA	Unknown
15		TLE 1:80 cDNA	Unknown

16		TLE 1:80 no cDNA	No Template Control
17		TLE 1:80 no cDNA	No Template Control
18		TLE 1:80 no cDNA	No Template Control
19		Fib 1:80 cDNA	Unknown
20		Fib 1:80 cDNA	Unknown
21		Fib 1:80 cDNA	Unknown
22		Fib 1:80 no cDNA	No Template Control
23		Fib 1:80 no cDNA	No Template Control
24		Fib 1:80 no cDNA	No Template Control
25		ASM 1:80 cDNA	Unknown
26		ASM 1:80 cDNA	Unknown
27		ASM 1:80 cDNA	Unknown
28		ASM 1:80 no cDNA	No Template Control
29		ASM 1:80 no cDNA	No Template Control
30		ASM 1:80 no cDNA	No Template Control
31		Negative Control	Negative Control
32		Negative Control	Negative Control
33		Negative Control	Negative Control

Figure S1: No cDNA qPCR control

During the production of cDNA from RNA, a control lacking MultiScribe (the cDNA production enzyme) was also generated. Without MultiScribe, no cDNA should be produced from the RNA and not produce fluorescence during the qPCR reaction. Through these means, it was possible to determine that the 'on column' DNA digestion successfully prevented genomic DNA contamination. This control was especially important for the reliable quantification of genes where an intron spanning qPCR amplicon was not possible, such as the snoRNAs (which are present in introns).

Appendix 5: Oligonucleotide Sequences

Target	Oligonucleotide Sequences (5'-3')
TECA1 Forward	CAACAGCTTCTCAGTGATACAGG
TECA1 Reverse	AGTACACCCTGAAAACCCACA
TECA2 Forward	TTCTGGCCACAGCACTTAAA
TECA2 Reverse	TGGTTAGGCTGGTGTTAGGG
TECA3 Forward	GCGAATGTGCATATGACTGAA
TECA3 Reverse	CTCCATTGCCCCTTTTTATG
HNF1A-AS1 Forward	CATTCCCTTCTCTGGCGTAG
HNF1A-AS1 Reverse	AAAGTGGGCAGGGGGTAA
TP73-AS1 Forward	TCCGGCTTCCCTAAAGAGAG
TP73-AS1 Reverse	GGACACAAGGGAGGGTGAG
SNORD76 Forward	GCCACAATGATGACAGTTTATTTGC
SNORD76 Reverse	AGATAATGGTGGTTAAGATCCTCAT
SNORD75 Forward	AGCCTGTGATGCTTTAAGAG
SNORD75 Reverse	TTCAGAAATCCCTTCTGTCC
GAS5 Forward	AACTTGCCTGGACCAGCTTA
GAS5 Reverse	CAAGCCGACTCTCCATACCT
SCARNA7 Forward	TTGTGGTGGCTATGGAAAGG
SCARNA7 Reverse	AGCCTCAGATGCACTCCAAT
KPNA4 Forward	CAATGGAAACCATTCAGGAGA
KPNA4 Reverse	GAGGGCCCAGACTGTGTCTA
SNORA81 Forward	ATTGCAGACACTAGGACCATGT
SNORA81 Reverse	GGTCCACCCCAGTCTTTACA
EIF4A2 Forward	TGATCTACCTACCAATCGTGAAAA

EIF4A2 Reverse	CCTTTCCTCCCAAATCGAC
PCAT19 qPCR F	GACACTGATACCAATGACATCCA
PCAT19 qPCR R	GCAGCAGAGTAGGTCAGGAAA
CBX5 qPCR F	AGAAGATGAAGGAGGGTGAAAA
CBX5 qPCR R	CCCGAGCGATATCATTG
FLOT2 qPCR F	TGTTGTGGTTCCGACTATAAACAG
FLOT2 qPCR R	GGGCTGCAACGTCATAATCT
ACTB qPCR F	CCAACCGCGAGAAGATGA
ACTB qPCR R	CCAGAGGCGTACAGGGATAG
PCAT19 Cloning 1F	ACTAGCCTCGAGAAACGTTATTTGACTGGAGTGAGG
PCAT19 Cloning 1R	TGTAATATTGGCATTGACATG
PCAT19 Cloning 2F	AATGAGAGAGACGGGAAG
PCAT19 Cloning 2R	AAGGAAAGCATATTGAAAATATAC
PCAT19 Cloning 3F	AATTGAAGTTGACTTTATGGAG
PCAT19 Cloning 3R	AGAATAGTGATTGGCCATATAG
PCAT19 Cloning 4F	TAAACATCTAGTCCAAAATTACTTG
PCAT19 Cloning 4R	CTAATTCGGCTCTTACAATC
PCAT19 Cloning 5F	TTCACCCCAACCTCCCTG
PCAT19 Cloning 5R	ATTCCTGCAGCCCGTAGTTTAAACTTCTGAAGTACAAACAT
PCAT19 siRNA D1	GGGTAATCTGGAAGAGTTT
PCAT19 siRNA D2	CAATGGAGGAAGAGGGTAA

Table S3: Oligonucleotide sequences

Appendix 6: Differential expression qPCR data

Mean fold change relative to FLOT2 and normalised to TLE (SEM)							
Gene	TLE	NLE	HUVEC	DF	ASMC	Ker	Figure
TECA1	1.03 (0.07)	1.09 (0.14)	144.82 (11.88)	0.10 (0.00)	2.32 (0.21)	1.32 (0.07)	3.3
HNF1A-AS1	1.04 (0.06)	3.22 (0.31)	1.76 (0.17)	0.02 (0.00)	0.06 (0.00)	5.88 (0.28)	3.4
TECA2	0.93 (0.04)	2.84 (0.27)	1.09 (0.17)	0.29 (0.04)	3.38 (0.29)	4.56 (0.29)	3.5
TP73-AS1	1.03 (0.07	2.68 (0.30)	1.06 (0.11)	0.61 (0.04)	1.31 (0.17)	1.69 (0.18)	3.6
TECA3	0.93 (0.08)	1.22 (0.27)	0.63 (0.05)	1.73 (0.06)	2.27 (0.22)	18.60 (0.68)	3.7
SNORD75	0.85 (0.08)	0.50 (0.04)	0.05 (0.01)	0.12 (0.01)	0.54 (0.04)	n/a	4.3
GAS5	0.40 (0.02)	0.12 (0.01)	0.05 (0.01)	0.10 (0.01)	0.28 (0.03)	n/a	
SNORD76	1.00 (0.02)	0.59 (0.03)	0.05 (0.01)	0.26 (0.03)	0.41 (0.05)	n/a	4.4
GAS5	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	n/a	
SNORD76	1.0 (0.02)	0.59 (0.03)	0.05 (0.01)	0.26 (0.03)	0.41 (0.05)	n/a	4.5
SNORD75	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	n/a	
GAS5	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	n/a	
SCARNA7	0.82 (0.09)	0.38 (0.05)	0.03 (0.01)	0.09 (0.01)	0.03 (0.01)	n/a	4.6
KPNA4	0.15 (0.01)	0.05 (0.00)	0.04 (0.00)	0.15 (0.00)	0.15 (0.00)	n/a	
SNORA81	0.95 (0.03)	0.67 (0.06)	0.01 (0.00)	0.02 (0.00)	0.01 (0.00)	n/a	4.7
EIF4A2	0.03 (0.01)	0.03 (0.00)	0.01 (0.00)	0.03 (0.01)	0.02 (0.00)	n/a	
SNORD32	1.00 (0.00)	1.11 (0.04)	0.06 (0.00)	0.47 (0.02)	0.21 (0.01)	n/a	4.8
RPL13A	0.01 (0.00)	0.01 (0.00)	0.00 (0.00)	0.01 (0.00)	0.01 (0.00)	n/a	
SNORD30	0.87 (0.08)	1.11 (0.05)	0.07 (0.01)	0.15 (0.01)	0.28 (0.01)	n/a	4.9
SNHG1	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	n/a	
SNORD100	1.14 (0.10)	1.63 (0.04)	0.19 (0.01)	0.07 (0.00)	0.25 (0.02)	n/a	4.10
RPS12	0.12 (0.00)	0.19 (0.01)	0.09 (0.00)	0.07 (0.01)	0.08 (0.01)	n/a	
PCAT19	0.94 (0.04)	1.73 (0.05)	1.16 (0.06)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	6.3

Table S4: Differential expression data (qPCR screening of different cell types)

The mean fold changes of all the transcripts tested in the qPCR cell panel are displayed in this table (n=3).

Mean fold change relative to ACTB and normalised to Confluent HUVEC (SEM)					
Gene	Confluent HUVEC (3x10^6)	Sparse HUVEC (1.5x10^6)	Sparse HUVEC (7.5x10^5)	Sparse HUVEC (3.75x10^5)	Figure
PCAT19	1.05 (0.03)	0.86 (0.07)	0.59 (0.01)	0.43 (0.07)	6.4

Table S5: Differential expression data (qPCR analysis of sparse HUVEC)

The mean fold change for PCAT19 at each of the cell densities is displayed in this table (n=3).

Mean fold change relative to ACTB and normalised to blank pWPI (SEM)			
Gene	HUVEC with blank pWPI	HUVEC with pWPI and PCAT19	Figure
PCAT19	1.00 (0.12)	176.38 (10.40)	6.5
WTAP	1.02 (0.10)	1.13 (1.13)	6.8
HIST1H2BK	1.07 (0.07)	1.12 (0.12)	6.9
CBX5	1.11 (0.09)	1.68 (0.11)	6.10
SUMF1	1.69 (0.37)	2.27 (0.43)	6.11
ILII	0.85 (0.15)	1.03 (0.08)	6.12
CNN1	1.02 (0.14)	0.82 (1.10)	6.13
HMOX1	0.93 (0.05)	0.51 (0.05)	6.14

Table S6: Differential expression data (qPCR validation of microarray data)

The mean fold changes for all of the genes validated using qPCR after the PCAT19 knockdown (microarray data) are displayed in this table (n=3).

Mean fold change relative to ACTB and normalised to NCD (SEM)				
Gene	NCD	D1	D2	Figure
PCAT19	0.96 (0.07)	0.27 (0.01)	0.36 (0.05)	6.7

Table S7: Differential expression data (qPCR validation of PCAT19 siRNA)

The mean fold changes for the qPCR validation of the PCAT19 knockdown are displayed in this table (n=3).

Appendix 7: Two Colour Microarray Analysis Codes

```
%% R

> library(limma)

> targets <- readTargets("~/Desktop/ECSM1 Microarray analysis/targets.txt")

> RG <- read.maimages(targets, path="~/Desktop/ECSM1 Microarray analysis", source="agilent.median")

> RG <- backgroundCorrect(RG, method="normexp", offset=16)

> MA <- normalizeWithinArrays(RG, method="loess")

> MA.avg <- avereps(MA, ID=MA$genes$ProbeName)

> design <- modelMatrix(targets, ref="wt")

> fit <- lmFit(MA.avg, design)

> fit2 <- eBayes(fit)

> output <- topTable(fit2, adjust="BH", coef="kd", number=44000)

> write.table(output, file="~/Desktop/ECSM1 Microarray analysis/Output.txt", sep="\t", quote=FALSE)

# Type 'q()' to quit R.
```

Appendix 8: Novel colon sequences:

A8.1 XLOC_029144 (TECA1):

GGTTTGCTTCTGCTCTTGAAGATGTGAACAGCTTCTAAGCATTCATTTTCTCTG

ACCCATACAACAGCTTCTCAGTGATACAGGGTTTAATTTAAACACATACAATGT

CCACCCCCAAACCTTCTGCCCACATCTACAAGTTTTATTTATTTTGTGGGTTTT

CAGGGTGACTAAGTTTTTCCCTACATTGAAAAGAGAAGTTGCCAAAAGGTGCA

CAGGAAATCATTTTTTTAAGTGAATATGATAATATGGGTCCGTGCTTAATACAA

CTGAGACATATTTGTTCTCTGTTTTTTTAGA GTCACCTCTTAAAGTCC

A8.2 XLOC_032009 (TECA2):

TCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGT

ATTTTCGTCTGGGGGGTGTGCACGCGATAGCATTGCGAGACGCTGGAGCCG

GAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTGCCTCATTCTATTATTTA

TCGCACCTACGTTCAATATTACAGGCGAACATACCTACTAAAGTGTGTTAATTA

ATTAATGCTTGTAGGACATAATAATAACAATTGAATGTCTGCACAGCCGCTTTC

CACACAGACATCATAACAAAAAATTTCCACCAAACCCCCCCCTCCCCCCGCTT

CTGGCCACAGCACTTAAACACATCTCTGCCAAACCCCAAAAACAAAGAACCCT

AACACCAGCCTAACCAGATTTCAAATTTTATCTTTAGGCGGTATGCACTTTTAA

CAGTCACCCCCCAACTAACACATTATTTTCCCCTCCCACTCCCATACTACTAAT

CTCATCAATACAACCCCCGCCCATCCTACCCAGCACACACACACCGCTGCTA

ACCCCATACCCCGAACCAACCAAACCCCAAAGACACCCCCCACAGTTTATGTA

GCTTACCTCCTCAAAGCAATACACTGAAAATGTTTAGACGGGCTCACATCACC

CCATAAACAAATAGGTTTGGTCCTAGCCTTTCTATTAGCTCTTAGTAAGATTAC

ACATGCAAGCATCCCCGTTCCAGTGAGTTCACCCTCTAAATCACCACGATCAA

AAGGGACAAGCATCAAGCACGCAGCAATGCAGCTCAAAACGCTTAGCCTAGC

CACACCCCCACGGGAAACAGCAGTGATTAACCTTTAGCAATAAACGAAAGTTT

AACTAAGCTATACTAACCCCAGGGTTGGTCAATTTCGTGCCAGCCACCGCGG

TCACACGATTAACCCAAGTCAATAGAAGCCGGCGTAAAGAGTGTTTTAGATCA

CCCCCTCCCCAATAAAGCTAAAACTCACCTGAGTTGTAAAAAACTCCAGTTGA

CACAAAATAGACTACGAAAGTGGCTTTAACATATCTGAACACACAATAGCTAA

GACCCAAACTGGGATTAGATACCCCACTATGCTTAGCCCTAAACCTCAACAGT

TAAATCAACAAAACTGCTCGCCAGAACACTACGAGCCACAGCTTAAAACTCAA

AGGACCTGGCGGTGCTTCATATCCCTCTAGAGGAGCCTGTTCTGTAATCGATA

AACCCCGATCAACCTCACCACCTCTTGCTCAGCCTATATACCGCCATCTTCAG

CAAACCCTGATGAAGGCTACAAAGTAAGCGCAAGTACCCACGTAAAGACGTT

AGGTCAAGGTGTAGCCCATGAGGTGGCAAGAAATGGGCTACATTTTCTACCC

CAGAAAACTACGATAGCCCTTATGAAACTTAAGGGTCGAAGGTGGATTTAGCA

GTAAACTGAGAGTAGAGTGCTTAGTTGAACAGGGCCCTGAAGCGCGTACACA

CCGCCCGTCACCCTCCTCAAGTATACTTCAAAGGACATTTAACTAAAACCCCT

ACGCATTTATATAGAGGAGACAAGTCGTAACATGGTAAGTGTACTGGAAAGTG

CACTTGGACGAACCAGAGTGTAGCTTAACACAAAGCACCCAACTTACACTTAG

GAGATTTCAACTTAACTTGACCGCTCTGAGCTAAACCTAGCCCCAAACCCACT

CCACCTTACTACCAGACAACCTTAGCCAAACCATTTACCCAAATAAAGTATAG

GCGATAGAAATTGAAACCTGGCGCAATAGATATAGTACCGCAAGGGAAAGAT

GAAAAATTATAACCAAGCATAATATAGCAAGGACTAACCCCTATACCTTCTGC

ATAATGAATTAACTAGAAATAACTTTGCAAGGAGAGCCAAAGCTAAGACCCCC

GAAACCAGACGAGCTACCTAAGAACAGCTAAAAGAGCACACCCGTCTATGTA

GCAAAATAGTGGGAAGATTTATAGGTAGAGGCGACAAACCTACCGAGCCTGG

TGATAGCTGGTTGTCCAAGATAGAATCTTAGTTCAACTTTAAATTTGCCCACAG

AACCCTCTAAATCCCCTTGTAAATTTAACTGTTAGTCCAAAGAGGAACAGCTCT

TTGGACACTAGGAAAAAACCTTGTAGAGAGAGTAAAAAATTTAACACCCATAG

TAGGCCTAAAAGCAGCCACCAATTAAGAAAGCGTTCAAGCTCAACACCCACTA

CCTAAAAAATCCCAAACATATAACTGAACTCCTCACACCCAATTGGACCAATCT

ATCACCCTATAGAAGAACTAATGTTAGTATAAGTAACATGAAAACATTCTCCTC

CGCATAAGCCTGCGTCAGATCAAAACACTGAACTGACAATTAACAGCCCAATA

TCTACAATCAACCAACAAGTCATTATTACCCTCACTGTCAACCCAACACAGGC

ATGCTCATAAGGAAAGGTTAAAAAAAGTAAAAGGAACTCGGCAAACCTTACCC

CGCCTGTTTACCAAAAACATCACCTCTAGCATCACCAGTATTAGAGGCACCGC

CTGCCCAGTGACACATGTTTAACGGCCGCGGTACCCTAACCGTGCAAAGGTA

GCATAATCACTTGTTCCTTAAATAGGGACCTGTATGAATGGCTCCACGAGGGT

TCAGCTGTCTCTTACTTTTAACCAGTGAAATTGACCTGCCCGTGAAGAGGCGG

GCATGACACAGCAAGACGAGAAGACCCTATGGAGCTTTAATTTATTAATGCAA

ACAGTACCTAACAAACCCACAGGTCCTAAACTACCAAACCTGCATTAAAAATT

TCGGTTGGGGCGACCTCGGAGCAGAACCCAACCTCCGAGCAGTACATGCTAA

GACTTCACCAGTCAAAGCGAACTACTATACTCAATTGATCCAATAACTTGACC

AACGGAACAAGTTACCCTAGGGATAACAGCGCAATCCTATTCTAGAGTCCATA

TCAACAATAGGGTTTACGACCTCGATGTTGGATCAGGACATCCCGATGGTGC

AGCCGCTATTAAAGGTTCGTTTGTTCAACGATTAAAGTCCTACGTGATCTGAG

TTCAGACCGGAGTAATCCAGGTCGGTTTCTATCTACTTCAAATTCCTCCC

TGTACGAAAGGACAAGAGAAATAAGGCCTACTTCACAAAGCGCCTTCCCCCG

TAAATGATATCATCTCAACTTAGTATTATACCCACACCCACCCAAGAACAGGG

TTTGTTAAGATGGCAGAGCCCGGTAATCGCATAAAACTTAAAACTTTACAGTC

AGAGGTTCAATTCCTCTTCTTAACAACATACCCATGGCCAACCTCCTACTCCT

CATTGTACCCATTCTAATCGCAATGGCATTCCTAATGCTTACCGAACGAAAAAT

TCTAGGCTATATACAACTACGCAAAGGCCCCAACGTTGTAGGCCCCTACGGG

CTACTACAACCCTTCGCTGACGCCATAAAACTCTTCACCAAAGAGCCCCTAAA

ACCCGCCACATCTACCATCACCCTCTACATCACCGCCCCGACCTTAGCTCTCA

CCATCGCTCTTCTACTATGAACCCCCCTCCCCATACCCAACCCCCTGGTCAAC

CTCAACCTAGGCCTCCTATTTATTCTAGCCACCTCTAGCCTAGCCGTTTACTC

AATCCTCTGATCAGGGTGAGCATCAAACTCAAACTACGCCCTGATCGGCGCA

CTGCGAGCAGTAGCCCAAACAATCTCATATGAAGTCACCCTAGCCATCATTCT

ACTATCAACATTACTAATAAGTGGCTCCTTTAACCTCTCCACCCTTATCACAAC

ACAAGAACACCTCTGATTACTCCTGCCATCATGACCCTTGGCCATAATATGAT

TTATCTCCACACTAGCAGAGACCAACCGAACCCCCTTCGACCTTGCCGAAGG

GGAGTCCGAACTAGTCTCAGGCTTCAACATCGAATACGCCGCAGGCCCCTTC

GCCCTATTCTTCATAGCCGAATACACAAACATTATTATAATAAACACCCTCACC

ACTACAATCTTCCTAGGAACAACATATGACGCACTCTCCCCTGAACTCTACAC

AACATATTTTGTCACCAAGACCCTACTTCTAACCTCCCTGTTCTTATGAATTCG

AACAGCATACCCCCGATTCCGCTACGACCAACTCATACACCTCCTATGAAAAA

ACTTCCTACCACTCACCCTAGCATTACTTATATGATATGTCTCCATACCCATTA

CAATCTCCAGCATTCCCCCTCAAA

A8.3 XLOC_004164 (TECA3):

ACGCAAGTGGGGTGAAAAAAAAGGATACGCGAATGTGCATATGACTGAATAG

GGAGGAAGGTCAGGGCTAGAAAGGAGGCTACATAAAAAGGGGCAATGGAGA

GTGCACAGGAAAGACACAGGA