The Effect of Vitamin D on Gene Expression in Colorectal Tumours and Normal Colon

Francesca M. Munro

A thesis submitted for the degree of Masters of

Biochemistry,

University of Otago,

New Zealand

Supervisors: John L. McCall and Michael A. Black

August 2015

ABSTRACT

Background

Colorectal cancer (CRC) is the third most frequently diagnosed malignancy and the fourth leading cause of death from cancer worldwide. In New Zealand, incidence rates of CRC are projected to decline in all age groups. The overall burden is projected to continue increasing as a result of population growth and an aging population. Epidemiological studies have reported an inverse association between vitamin D status and incidence of CRC. Higher serum vitamin D levels at the time of diagnosis or post-surgery have been associated with improved long-term outcome in CRC, but there is no proven causal link. It has been postulated that vitamin D levels should be maintained at >80 nmol/L for both bone and overall health. Vitamin D deficiency is common in NZ particularly in the winter months. The primary aim of this study was to examine whether a single large dose of vitamin D administered to CRC patients in the window between diagnosis and surgery could have a measurable effect on vitamin D responsive genes in the tumour and corresponding normal colon.

Methods

The study was a randomised, double-blind, placebo-controlled trial of vitamin D supplementation in patients undergoing elective surgery for CRC at Dunedin Hospital. For the primary outcome, RNA from resected normal and tumour tissue was profiled on microarray gene-chips. Array gene expression data were analysed using single gene analysis to identify differentially expressed individual genes and over-representation and gene-set analysis (GSA) to identify differences in pathway expression. Secondary outcomes including complication rate, length of hospitalisation, post-operative recovery and survival were measured.

Results

There were no baseline differences between the treatment and control groups. Pre-incision vitamin D concentrations were higher in the treatment than in the placebo group (mean 87 +/- 22 *vs* 49 +/- 19 nmol/L; p= >0.001). The complication rate and length of hospitalisation were within the expected range for this cohort. No differences in post-operative recovery, cancer recurrence or survival were observed. There were no

significant differences in single gene expression between the treatment and control groups. There were also no significant differences in gene expression of vitamin D modulated genes between the groups. Differences in pathway expression were identified between the treatment arm and placebo arm. In the normal tissue, patients randomised to receive vitamin D had down-regulation of a number of pathways compared to those randomised to the placebo group including Fatty Acid Metabolism, Drug Metabolism - Cytochrome P450, Metabolism of Xenobiotics by Cytochrome P450, Vitamin Digestion and Absorption and Negative Regulation of Growth pathways. In the tumour tissue the treated group also had down-regulation of several pathways compared to the placebo group including the Fatty Acid Metabolism, Fatty Acid beta-Oxidation and Oxidative Phosphorylation pathways. In the paired analysis the expression of the Ribosome and Translational Termination pathways were enhanced by vitamin D in the tumour tissue of a subgroup of patients and reduced in another subgroup, compared to the normal tissue. Over-representation analyses identified a number of other pathways that may be different between the study groups.

Conclusion

In this randomised controlled trial potentially significant biological differences between the vitamin D and placebo groups were identified. The Fatty Acid Metabolism and Fatty Acid beta-Oxidation pathways were down-regulated in the tumour tissue of the treated patients compared with the untreated patients. Down-regulation of fatty acid metabolism in the tumour may lead to the slowing of tumour growth. Un-metabolised butyrate (short-chain fatty acid utilised by colon epithelia) may also precipitate reduced cell proliferation and enhanced differentiation and apoptosis. Further work is required to develop our understanding of the impact of vitamin D on the biology of colorectal cancer.

ACKNOWLEDGEMENTS

This study was designed by the candidate's supervisors, John McCall and Michael Black**.** The candidate had no input into the study design.

The candidate explained the study to the patients, obtained written consent, collected and stored tissue and collated clinical data.

RNA extraction and purity assessment was performed by Mrs Victoria Phillips of the Department of Surgical Sciences of the University of Otago Medical School. RNA quality assessment was carried out by Dr Aaron Jeffs of the Department of Biochemistry of the University of Otago. RNA labeling, RNA hybridisation to the genechips and profiling and scanning of the genechips was conducted by Mr Leslie McNoe and Dr Luxmanan Selvanesan of the Department of Biochemistry of the University of Otago.

The candidate analysed the microarray gene expression data and measured the secondary outcomes of the study.

THANKS

Firstly, I'd like to thank John McCall for giving me the opportunity to undertake further study (further thanks to John are in order and can be found below). I'd like to acknowledge the University of Otago for encouraging its staff to continue their education and having processes in place to enable this.

There are many employees of the university, Dunedin Hospital and Southern Community Laboratories who provided help and support for this project. I won't name them all individually but their involvement was very much appreciated, especially the banter ;) . These fabulous people included staff from the Department of Surgery, Wards 4A and 4B (and occasionally staff from other wards), Pre-admission Clinics, the Radiology Department, the Oncology Department and any other department that allowed me to follow their patients around to discuss the study, obtain consent and blood and administer the study treatment. The Southern Community Laboratories' phlebotomists are always so helpful (nothing's ever a problem) including at the out-of-town clinics. Staff from other hospital departments were indispensable, including the Pharmacy Department. Theatre staff were amazing (and still are), always happy to provide me with a bucket of fresh flesh. The Southern Community Laboratories' histology prep staff, pathologists and especially the pathology registrars were stunning! Thanks so much for providing tissue samples from fresh specimens which usually meant lots of mucky brown stuff. Thanks for not barfing on my shoes.

Vicky Phillips is the lab manager from heaven. Not only does she provide copious amounts of cake, but she recruits patients and extracts RNA as well. Quite a woman! I'd like to say, "Les, where's my cake?" But at least you are a master with microarrays. Cheers to you too, Lux. And Aaron. You all make working with little things look fun!

One of my fellow students became an unpaid 'go-to' for my myriad 'R' questions. Poor guy has been promoted to friend and adopted son! Thanx Tom! I rate ya!

Big ups to my various work roomies over the duration of this project. You were almost bearable…nah I love you all, especially when you're quiet Lesley! x

Arohanui to all my fantastic and supportive friends including you Bea. Xxx

Thanx Lesley B. for reading this thesis for me. Hopefully you don't find too many errors!

Projects like this one are not possible without patients consenting to be involved. I'm in awe of how much these wonderful people wanted to do something to help others even though many of these people had only just received their diagnosis. I hope your generosity combined with our efforts make for better outcomes for patients in the future.

Being grateful for one's supervisors is to be expected but mine have had to put up with a distracted, hormonal and tired student in me (we won't mention retarded- though my kids would enjoy that!). I am really grateful guys that you managed to stick with me through this and provide me with 'extra support' in various ways. Hopefully that 'extra support' pays off in my role for the department and the CTRG. Thanks John. Thanks Mik. ☺

And of course the obvious gratefulness towards one's family because they've had to put up with various inconveniences while you've been 'tripping'. Unfortunately my family has had to endure a major loss while I've been preoccupied with study. I think we've all done an amazing job of getting through what could have been a very messy time. I want to say a heartfelt thank you to you Bruce for helping to make this an amicable process and for your support during this project. Big love to all my kids, Bea, Zac, Elliot, Josi and Gianni. You're all fantabulous & I love you guys truckloads!!

TABLE OF CONTENTS

Chapter 1. INTRODUCTION……1

1.1 Colorectal Cancer……1

1.1.1 Colorectal cancer worldwide……1

1.1.2 Colorectal cancer in New Zealand……1

1.1.3 Presentation, treatment and prognosis for colorectal cancer……2

1.2 Health Effects of Vitamin D ……4

1.2.1 Vitamin D – basic biology……4

1.2.2 Mechanism of action of calcitriol……6

1.2.3 Vitamin D – role in calcium/phosphorous homeostasis and bone health……9

1.2.4 Vitamin D – other health effects……9

1.2.5 Recommendations for vitamin D for optimal health……10

1.2.6 Sunshine and vitamin D……11

1.2.7 Diet and vitamin D……14

1.2.8 Vitamin D levels in New Zealand……15

1.3 Vitamin D and Colorectal Cancer……16

1.3.1 Population studies……16

1.3.1.1 Ecological studies……16

1.3.1.2 Epidemiological studies……17

1.3.2 Human experimental studies……19

1.3.3 Animal studies……22

1.3.4 Molecular studies……23

1.3.4.1 Tumouri-genesis in colorectal cancer……23

1.3.4.2 Anti-tumoural actions of calcitriol in colorectal cancer cells……23

1.3.4.2.a Proliferation……23

1.3.4.2.b Differentiation……24

1.3.4.2.c Apoptosis……24

1.3.4.2.d Angiogenesis……25

1.3.4.2.e Detoxification……25

1.3.4.2.f Antagonism of the WNT/β-catenin pathway……25

1.3.4.3 Calcitriol target genes in colorectal cancer cells……26

1.3.4.4 Mechanisms of resistance to calcitriol action in colorectal cancer……26

1.4 Summary, primary aim and chapter outline……27

Chapter 2. TRIAL DESIGN……29
2.1 Introduction……29
2.2 Study Design……30
2.3 Participants……30
2.3.1 Consent and ethics……30
2.3.2 Inclusion criteria……31
2.3.3 Exclusion criteria……31
2.3.4 Randomisation……32
2.3.5 Blinding and concealment……32
2.3.6 Baseline measurements……32
2.3.6.1 Demographic……32
2.3.6.2 Clinical……32
2.3.6.3 Laboratory……33
2.3.6.4 Clinico-pathological……33
2.4 Intervention……33
2.4.1 Dose justification……34
2.5 Samples and Follow-up……34
2.5.1 Specimen collection and storage……34
2.5.2 Clinical follow-up……35
2.5.3 Data storage and de-identification……35
2.6 Outcome Measures……36
2.6.1 Primary outcome – differential gene expression analysis……36
2.6.2 Secondary outcomes……36
2.6.2.1 Vitamin D levels……36
2.6.2.2 Adverse events related to intervention……36
2.6.2.3 Immune function measurements……36
2.6.2.4 Post-operative recovery……37
2.6.2.5 Recurrence and survival……37
2.7 Statistical Analysis……37
2.7.1 Analysis for clinical outcomes……37
2.7.2 Differential gene expression analysis……38
2.7.3 Sample size……38
2.8 Summary……39

Chapter 3. RESULTS OF CLINICAL OUTCOMES.....40
3.1 Introduction.....40
3.2 Patient Recruitment.....40
3.3 Baseline Patient Characteristics.....41
3.3.1 Demographic.....41
3.3.2 Clinical.....42
3.3.3 Laboratory.....43
3.3.4 Clinico-pathological.....44
3.4 Outcome Measures.....44
3.4.1 Primary outcome – differential gene expression analysis.....44
3.4.2 Secondary outcomes.....45
3.4.2.1 Vitamin D and calcium levels.....45
3.4.2.2 Adverse events related to intervention.....46
3.4.2.3 Post-operative recovery.....46
3.4.2.4 Recurrence and survival.....47
3.5 Discussion.....49

Chapter 4. METHODS FOR DIFFERENTIAL GENE EXPRESSION ANALYSIS.....51
4.1 Introduction.....51
4.2 Methods for Differential Gene Expression Analysis.....51
4.2.1 Microarray methods.....51
4.2.2 Microarray quality control.....52
4.2.3 Data normalisation.....56
4.2.4 Genes with vitamin D responsive elements.....57
4.2.5 Differential expression analysis (genes).....58
4.2.5.1 Limma.....58
4.2.5.2 Multiple testing correction.....60
4.2.5.3 Normal, tumour and paired analyses.....61
4.2.6 Differential expression analysis (gene-sets).....62
4.2.6.1 Gene-sets and pathways.....62
4.2.6.2 Over-representation analysis and gene-set analysis.....63
4.2.6.3 Tools used for over-representation and gene-set analyses.....65
4.2.7 Visualisation methods.....66
4.2.7.1 Heatmaps.....66

4.2.7.2 Barcode plots……68
4.2.8 Metagene creation for summarising gene-sets……69
4.3 Summary……69

Chapter 5. RESULTS OF DIFFERENTIAL GENE EXPRESSION ANALYSIS……70
5.1 Introduction……70
5.2 Microarray quality control and data normalisation……70
5.2.1 Microarray quality control - pre-normalisation……71
5.2.2 Microarray quality control - post-normalisation……72
5.3 Normal tissue - control versus treatment……73
5.3.1 Single gene analysis……73
5.3.2 Over-representation analysis……75
5.3.3 Gene-set analysis – KEGG……76
5.3.4 Gene-set analysis – GO……80
5.4 Tumour tissue - control versus treatment……83
5.4.1 Single gene analysis……83
5.4.2 Over-representation analysis……84
5.4.3 Gene-set analysis – KEGG……85
5.4.4 Gene-set analysis – GO……92
5.5 Paired analysis - control versus treatment……98
5.5.1 Paired analysis - without group comparison – KEGG……99
5.5.2 Paired analysis - without group comparison – GO……99
5.5.3 Single gene analysis - paired analysis……100
5.5.4 Over-representation analysis - paired analysis……102
5.5.5 Gene-set analysis - paired analysis – KEGG……103
5.5.6 Gene-set analysis - paired analysis – GO……106
5.6 Discussion……110

Chapter 6. CONCLUSION……118

REFERENCES……122

APPENDIX 1. Randomisation list……142
APPENDIX 2. VDRE gene list……143

APPENDIX 3. arrayQualityMetrics Report for Microarray Expression Data……156
APPENDIX 4. arrayQualityMetrics Report for Normalised Microarray Expression Data……172
APPENDIX 5a. RNA quality – purity……181
APPENDIX 5b. RNA quality……188
APPENDIX 6. Heatmaps for top 100 genes in normal, tumour and paired tissue Analyses……190
APPENDIX 7. Proliferation metagene across the treatment and placebo groups……193

LIST OF TABLES

3.1 Baseline patient demographic characteristics……42
3.2 Baseline clinical characteristics……43
3.3 Baseline serum levels……44
3.4 Baseline patient characteristics……44
3.5 Secondary outcomes: calcidiol levels……45
3.6 Secondary outcomes: post-operative recovery……47

5.1 Linear model results summary table for normal tissue analysis……75
5.2 Significant KEGG pathways featured in GSA of normal tissue compared between study groups……78
5.3 Significant GO terms featured in GSA of normal tissue compared between study groups……82
5.4 Linear model results summary table for tumour tissue analysis……85
5.5 Significant KEGG pathways featured in GSA of tumour tissue compared between study groups……88
5.6 Significant GO terms featured in GSA of tumour tissue compared between study groups……93
5.7 Significant KEGG pathways featured in GSA of paired analysis WITHOUT comparison between study groups……100
5.8 Significant GO terms featured in GSA of paired analysis WITHOUT comparison between study groups……101
5.9 Linear model results summary table for the paired analysis……102
5.10 Significant KEGG pathways featured in GSA of the paired analysis compared between the treatment and control groups……104
5.11 Significant GO terms featured in GSA of paired analysis compared between the study groups……108

LIST OF FIGURES

1.1 Five-year survival by the American Joint Committee on Cancer (AJCC) sixth edition system stages I-IV……..4
1.2 Calcitriol synthesis……..6

3.1 CONSORT diagram……..41
3.2 Boxplot of calcidiol levels at baseline and post treatment in the study groups……..46
3.3 Kaplan-Meier curve for recurrence-free survival in the treatment and placebo groups……..48
3.4 Kaplan-Meier curve for overall survival in the treatment and placebo groups……48

5.1 Barcode plot for the Metabolism of Xenobiotics by Cytochrome P450 pathway and boxplot of 'Metabolism of Xenobiotics by Cytochrome P450 pathway' metagene expression……..79
5.2 Barcode plot for the Fatty Acid Metabolism pathway and boxplot of 'Fatty Acid Metabolism pathway' metagene expression……..81
5.3 Barcode plot for the Negative Regulation of Growth pathway and boxplot of 'Negative Regulation of Growth pathway' metagene expression……..83
5.4 Barcode plot for the Fatty Acid Metabolism pathway and boxplot of 'Fatty Acid Metabolism pathway' metagene expression……90
5.5 Heatmap of the Fatty Acid Metabolism pathway genes across the normal and tumour samples……91
5.6 Barcode plot for the Oxidative Phosphorylation pathway and boxplot of 'Oxidative Phosphorylation pathway' metagene expression……92
5.7 Barcode plot for the Fatty Acid beta-Oxidation pathway and boxplot of 'Fatty Acid beta-Oxidation pathway' metagene expression……95
5.8 Heatmap of the Fatty Acid beta-Oxidation pathway genes across the normal and tumour samples……96

5.9 Barcode plot for the Respiratory Electron Transport Chain pathway and boxplot of 'Respiratory Electron Transport Chain pathway' metagene expression……97
5.10 Barcode plot for the Ribosome pathway and boxplot of 'Ribosome pathway' metagene expression……105
5.11 Heatmap of Ribosome pathway standardised gene expression differences in the paired analysis across the treatment and control groups……106
5.12 Barcode plot for the Translational Termination pathway and boxplot of 'Translational Termination pathway' metagene expression……109
5.13 Heatmap of Translational Termination pathway standardised gene expression differences in the paired analysis across the treatment and control groups……110

LIST OF ABBREVIATIONS

+	plus
7-DHC	7-dehydro-cholesterol
A	average log intensities
ADP	adenosine diphosphate
AG	aerobic glycolysis
AJCC	American Joint Committee on Cancer Method
ASA	American Society of Anaesthesiology
ATP	adenosine triphosphate
BH	Benjamini and Hochberg
BMI	body mass index
CI	confidence interval
cm	centimetre
CRC	colorectal cancer
CT	computerised tomography
CTC	computerised tomography colonography
CTCAE	Common Terminology Criteria for Adverse Events
CYP	cytochrome P450
DE	differential expression
DEG	differentially expressed genes
DHB	District Health Board
DNA	deoxy-ribonucleic acid
DR3	direct repeat 3
DR4	direct repeat 4
ECM	extracellular matrix
EGFR	epidermal growth factor receptor
ER6	everted repeat 6
ER9	everted repeat 9
ES	enrichment score
FDR	False Discovery Rate
g	gram
G1	grade 1

G2	grade 2
G3	grade 3
GEO	Gene Expression Omnibus
GO	Gene Ontology
GSA	gene-set analysis
GSEA	gene-set enrichment analysis
HAT	histone acetyl-transferase
HDAC	histone de-acetylase
HPFS	Health Professionals Follow-Up Study
IARC	International Agency for Research on Cancer
IL1β	interleukin-1β
IL6	interleukin 6
IL8	interleukin 8
IU	international unit
kb	kilobases
KEGG	Kyoto Encyclopedia of Genes and Genomes
KS	Kolmogorov-Smirnov
LBD	ligand binding domain
ln	natural log
log	logarithm
log2	logarithm base 2
M	log intensities
MES	maximum enrichment score
mcg	microgram
mg	milligram
min	minus
miRNA	microRNA
MM	mismatch
mmol/L	millimole per litre
MOH	Ministry of Health, New Zealand
MOP	mitochondrial oxidative phosphorylation
MRI	Magnetic resonance imaging
NADH	nicotinamide adenine dinucleotide

NHI	National Health Number
NHS	Nurses' Health Study
nm	nanometre
nmol/L	nanomole per litre
NY	New York
NZ	New Zealand
^{o}	degrees latitude
^{o}C	degrees celcius
^{o}S	degrees latitude south
OR	odds ratio
p	p-value
PCA	principal component analysis
Pi	inorganic phosphate
QC	quality control
RCT	randomised controlled trial
RMA	robust multi-chip average
RNA	ribonucleic acid
RR	risk ratio
RRpMM	Rho-ROCK-p38MAPK-MSK
RXR	retinoid x receptor
SAFE	significance analysis of function and expression
sd	standard deviation
SPF	sun protection factor
TCA	tricarboxylic acid cycle
TCF/LEF	T-cell factor/lymphoid enhancer factor
TF	transcription factor
TGF-β	transforming growth factor-β
TNF-α	tumour necrosis factor-α
TSS	transcription start site
t-statistic	test statistic
UK	United Kingdom
US	United States
USA	United States of America

UV	ultra-violet
UVA	ultra-violet A
UVB	ultra-violet B
VDR	Vitamin D receptor
VDRE	vitamin D responsive element
VDS	vitamin D study
WA	Washington State
WHO	World Health Organisation

1. INTRODUCTION

1.1 Colorectal cancer

1.1.1 *Colorectal cancer worldwide*

An estimated 14.1 million cases of cancer were diagnosed and an estimated 8.2 million cancer deaths occurred worldwide in 2012, according to Globocan, a project of the International Agency for Research on Cancer (IARC) of the World Health Organisation (WHO) (IARC 2015).

Worldwide colorectal cancer is the third most frequently diagnosed malignancy and is the fourth leading cause of death from cancer, with an estimated 1,360,600 cases diagnosed and an estimated 693,900 deaths in 2012 (IARC 2015).

1.1.2 *Colorectal cancer in New Zealand*

In New Zealand (NZ) in 2011, cancer was the most common cause of death with 21,050 new cases registered, accounting for nearly a third of all deaths, and 8891 people having cancer as their underlying cause of death (MOH 2014).

Colorectal cancer (CRC) was one of the two most common cancers registered in NZ in 2010, accounting for 14.1% of all cancer registrations. In the same year CRC was the second most common cause of death from cancer accounting for 14.1% of cancer deaths. It was placed in the top three for the most common cause of cancer death for both men and women in all age groups bar one, <25 years of age. In women aged 75 years and over, CRC was the most commonly registered cancer leading to the most common cause of death by cancer in this age group. There was a reduction in registration rates of CRC between 2000 and 2010 for both men and women by 8.8 and 11.7% respectively. Mortality rates declined by 20.2% for men and 20.7% for women for the same period (MOH 2013).

Colorectal cancer registration and mortality rates for Māori have historically been lower than for non-Māori. In recent years this trend has changed for Māori women, their

registration and mortality rates converging with those of non-Māori women (MOH 2013). In 2012, mortality rates for Māori men and women were higher than for non-Māori with age-standardised rates of 743/100,000 and 567/100,000, respectively compared with 425/100,000 and 306/100,000 for non-Māori men and women (MOH 2015).

Patients residing in the South Canterbury and Southern District Health Board (DHB) regions exhibited significantly higher registration rates of CRC and rates of death from CRC than the NZ rates in 2008. The Waikato DHB region was the only other region to exhibit a significantly higher mortality rate than the national rate. By comparison, registration rates that were significantly lower than the national rate were recorded in the Waitemata, Counties Manukau, Tairawhiti, Hutt Valley and Capital and Coast DHBs. The Waitemata DHB region was the only region to exhibit a mortality rate that was significantly lower than the national rate (MOH 2011a).

Incidence rates of CRC are projected to decline in all age groups except in people over 75, decreasing overall by approximately one-quarter in the 45-74 age group. However, the overall burden is projected to continue increasing as a result of population growth and an aging population (MOH 2010).

1.1.3 *Presentation, treatment and prognosis of colorectal cancer*

The symptoms of CRC are mostly non-specific and include a change in bowel habit, rectal bleeding (overt or manifesting as iron deficiency anaemia), abdominal pain and weight loss. Up to one third of patients present acutely with bowel obstruction, perforation or major bleeding, and some patients with manifestations of advanced or metastatic disease (for example ascites, liver or pulmonary metastases). In some countries population based screening is offered with the aim of early diagnosis and treatment. Screening, based on faecal occult blood testing, is currently being piloted in New Zealand in the Waitemata DHB.

Patients presenting with signs and symptoms of CRC are referred for investigation that usually includes colonoscopy or computerised tomography colonography (CTC). During colonoscopy, tumours can be biopsied for histological confirmation of the diagnosis and polyps, which may be a precursor to cancer, are removed.

Patients with a confirmed diagnosis of CRC usually undergo further radiological staging with computerised tomography (CT) chest, abdomen and pelvis, and those with rectal cancer undergo magnetic resonance imaging (MRI) of the pelvis. Treatment is then planned taking into account the disease stage, patient co-morbidities and preferences. Surgery is the primary treatment for most patients with CRC, however patients with rectal cancer may also have neo-adjuvant therapy (i.e. oncological treatment prior to surgery) with radiotherapy, either alone or combined with chemotherapy.

After surgery the resected bowel and mesentery are submitted for histo-pathological evaluation to determine tumour stage (section 2.3.6.4). Clinico-pathological staging is determined upon review of histology, radiology and surgical reporting.

Tumour stage is defined according to the American Joint Cancer Committee (AJCC) system, and is used to determine both prognosis and the need for further treatment following surgery (Edge and Compton 2010). In general, patients with early stage disease confined to the bowel wall (AJCC Stage I and II) are usually treated by surgery alone, patients with lymph node metastases (AJCC Stage III) are treated with adjuvant chemotherapy after surgery, and patients with distant metastases (AJCC Stage IV) may undergo further treatment with surgery (if the metastatic disease is limited and resectable) and/or chemotherapy. Disease stage at diagnosis is the major determinant of prognosis, as shown in Figure 1.1.

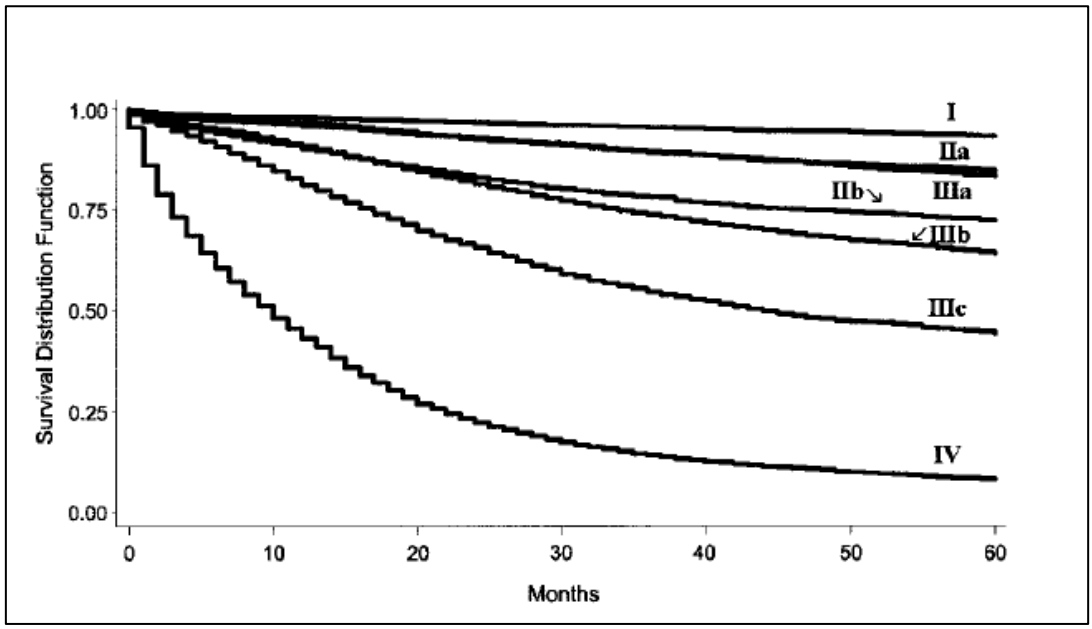


Figure 1.1. Five-year survival by the American Joint Committee on Cancer (AJCC) sixth edition system stages I–IV. According to stages defined by the AJCC sixth edition system, 5-year stage-specific survivals were 93.2% for stage I, 84.7% for stage IIa, 72.2% for stage IIb, 83.4% for stage IIIa, 64.1% for stage IIIb, 44.3% for stage IIIc, and 8.1% for stage IV (O'Connell *et al.* 2004).

Following curative treatment patients are usually offered some form of surveillance for new primary lesions and disease recurrence (MOH 2011b).

1.2 Health Effects of Vitamin D

1.2.1 *Vitamin D – basic biology*

Vitamin D comprises a group of fat-soluble seco-steroids, similar in structure to steroids. In humans the most important compound in this group is vitamin D_3, of which there are several forms (Holick 2006).

In the remainder of this thesis when generally referring to 'vitamin D_3', 'vitamin D' will be used. Calcidiol will be used when referring to plasma/serum levels of vitamin D. Calcitriol will be used when referring to the active form of vitamin D.

Although vitamin D is commonly referred to as a vitamin, in the strictest sense it is not (Feldman *et al.* 2014). A vitamin is described as an essential compound that cannot be synthesised by the body in adequate amounts (Lieberman and Bruning 1990). In contrast, upwards of 90% of an individual's vitamin D requirements can be produced in

the skin upon adequate sun exposure (Holick 2000). Nowadays, vitamin D is usually referred to as a pro-hormone given its synthesis in the skin and the diverse array of its activity (Pereira *et al.* 2012).

The hormonally active form of vitamin D in the blood is calcitriol, which is synthesised in a highly regulated multi-step process (Figure 1.2). The process begins with the photolysis, via ultra-violet B (UVB) rays, of pro-vitamin D_3 (also known as 7-dehydro-cholesterol = 7-DHC, which is produced by cholesterol metabolism) in the skin, yielding pre-vitamin D_3 (also known as pre-cholecalciferol). Pre-vitamin D_3 is a thermodynamically unstable isomer and thus spontaneously isomerises to cholecalciferol, which can also be obtained from various food sources (section 1.2.6) (Holick 2011, Pereira *et al.* 2012).

Cholecalciferol is transported to the liver via the blood, bound to vitamin D binding protein, where it is hydroxylated by vitamin-D_3-25-hydroxylases, producing the main circulating pro-hormone, calcidiol (also known as 25-hydroxy-vitamin-D_3) with an estimated half-life of between 10 to 14 days and several weeks (Brandi and Minisola 2013, Holick 2011, Hollis and Wagner 2013).

The last step in the generation of the hormonally active form of vitamin D, calcitriol, occurs in the kidney via another hydroxylase, 1-α-hydroxy-vitamin-D_3-hydroxylase encoded by *CYP27B1* (Holick 2011). *CYP27B1* is also expressed in the brain, colon, prostate, endothelial and immune cells, suggesting an extra-renal production of calcitriol with autocrine/paracrine activity (Pereira *et al.* 2012). Calcitriol has a half-life of 3-12 hours (Hollis and Wagner 2013, Smith and Goodman 1971) and is rendered less active by a further hydroxylation step via the ubiquitous enzyme, 1,25-dihydroxyvitamin-D-24-hydroxylase encoded by *CYP24A1* (Figure 1.2) (Holick 2011).

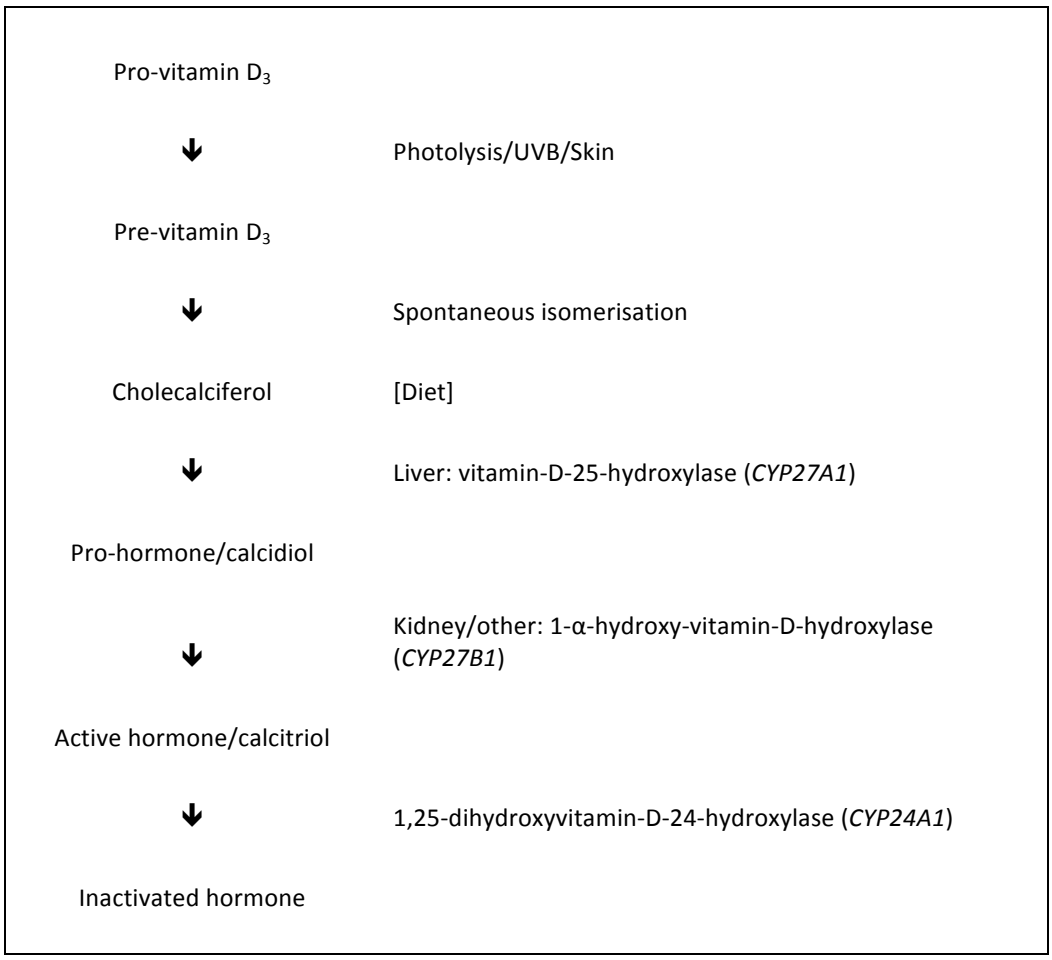


Figure 1.2. Calcitriol synthesis. Pro-vitamin D_3 in the skin is photolysed, upon exposure to UVB radiation, to produce pre-vitamin D_3, which in turn spontaneously isomerises to cholecalciferol (also able to be obtained from the diet). Cholecalciferol undergoes hydroxylation in the liver via vitamin-D-25-hydroxylase (*CYP27A1*) to yield the circulating vitamin D_3 pro-hormone, calcidiol. Activation of the pro-hormone (calcidiol) to the active hormone, calcitriol occurs in many tissues including the colon via hydroxylation by 1-α-hydroxy-vitamin-D-hydroxylase (*CYP27B1*). Calcitriol is inactivated via further hydroxylation by 1,25-dihydroxyvitamin-D-24-hydroxylase (*CYP24A1*). UVB = ultra-violet B radiation.

1.2.2 *Mechanism of action of calcitriol*

The vitamin D receptor (VDR) is a member of the superfamily of steroid/thyroid nuclear receptors and is an almost ubiquitously expressed ligand-dependent transcription factor (TF) (Deeb *et al.* 2007, Thorne and Campbell 2008). It is involved in the regulation of a very large number of genes (Norman *et al.* 1982, Reichel *et al.* 1989, Verstuyf *et al.* 2010). Correspondingly, gene expression studies have demonstrated that calcitriol regulates, directly and indirectly, a very substantial number of genes also (reviewed by (Bouillon *et al.* 2008).

The role of calcitriol is to convert the VDR into an active protein that can bind to specific DNA sequences and recruit co-regulatory complexes that are essential for the subsequent process of gene expression. The VDR has several conserved domains including a ligand-binding domain (LBD), a DNA-binding domain and a transcriptional activation domain (Pike *et al.* 2012).

Ligand binding causes conformational change in the VDR that increases, a) its association with another nuclear receptor, retinoid X receptor (RXR) and its ligand, 9-*cis* retinoic acid (Pereira *et al.* 2012), which in turn activates the VDR's DNA-binding ability, b) its interaction with the basal transcription apparatus, consisting of general TFs, RNA polymerase and the Mediator multiple protein complex, and c) its interaction with a number of co-activators precipitating the disengagement of co-repressors (Leyssens *et al.* 2013, Pereira *et al.* 2012, Pike *et al.* 2012).

In the absence of calcitriol, VDR interacts with the nuclear receptor co-repressor and the silencing mediator for retinoid co-repressors. These in turn bind histone de-acetylases (HDACs) that de-acetylate nucleosomes and prevent gene activation by controlling chromatin accessibility (Pereira *et al.* 2012).

In the presence of calcitriol binding, the conformational change in the LBD causes a loss of co-repressor binding and the association of co-activator complexes. The co-activators that are associated with the earliest stage of gene activation by VDR are the steroid receptor co-activators. They possess an intrinsic histone acetyl-transferase (HAT) activity and also recruit other histone acetyl-transferase enzymes to target promoters. The ensuing acetyl-transferase activity causes acetylation of nucleosomes at target genes and consequently facilitation of gene activation by allowing chromatin access (Pereira *et al.* 2012).

VDR-DNA binding (calcitriol bound or not) occurs within specific DNA sequences in the target genes termed vitamin D responsive elements (VDREs). The most frequent DNA sequence identified at these sites is comprised of two directly repeated half-sites of the consensus sequence A/G G G/T T C A (hexamer) separated by three nucleotides (Direct Repeat 3 (DR3)-type) (Kerner *et al.* 1989, Pike *et al.* 2012). Results from other studies examining DNA fragments near transcription start sites (TSS) suggest that the

separation in the consensus sequence can be three or four (Direct Repeat 4 (DR4)-type) nucleotides or that the consensus sequence can be two everted repeats, TGAACT…AGTTCA, separated by six to nine nucleotides (Everted Repeat 6-9 (ER6-ER9)-type) (Carlberg and Dunlop 2006, Carlberg and Seuter 2009). In these studies, the DR3-type was also found to be the most common consensus sequence, constituting <30% of all VDR binding sites. A substantial proportion of VDR binding sites were found to be neither of the described direct nor everted type sequences (Carlberg and Campbell 2012, Meyer *et al.* 2012).

To date, it has been thought that regulation of eukaryotic genes occurs largely through enhancers located a few kilo-bases from gene promoters (Pike 2011). However multiple regulatory elements for the VDR-RXR dimer have been identified on the genome and frequently removed from the promoter region by hundreds of kilo-bases. In a study performed in osteoblasts, VDREs were observed proximal to the promoters of known genes, but were more often observed distal to promoters in inter-genic regions surrounding transcription units or within introns. The data also indicates that genes are frequently modulated by multiple regulatory elements and that these elements are often located in clusters (Pike *et al.* 2012). These findings could be interpreted in several ways. The multiple VDREs in distant locations may, upon chromatin folding, result in enhancer clustering which may facilitate a multi-cluster gene activation mechanism. The diverse VDRE locations may aid in a modular activation mechanism or perhaps provide a synergistic activation mechanism. The association of additional nuclear receptors and TFs at ligand-induced VDR/RXR sites suggests a modular activation procedure (Pike *et al.* 2012). However, a modular activation mechanism does not necessarily exclude a multi-cluster activation method.

VDR mediates adjustments in gene expression indirectly via its ability to recruit sets of large and diverse co-regulatory multi-protein complexes. Co-regulatory multi-plexes have at least one VDR-interacting element with each of the multi-plexes undertaking a key step in the process of transcriptional regulation. For example, the Mediator multi-plex has a role in the recruitment of RNA polymerase II and its insertion at the pre-initiation complex. The details of how each of the multi-plexes functions to alter the expression of gene targets are beginning to emerge and may involve mechanisms that facilitate interpretation of epigenetic modifications. Epigenetic modifications facilitate

the expression of subsets of the genome thus determining a cell-type's unique phenotype. Although these mechanisms are unknown at present, it is specific DNA binding by the VDR that provides the initiation point for the directing of processes required to alter gene expression in a cell-type specific way (Pike *et al.* 2012).
In summary, the overarching principles of VDR-RXR action at target genomes include; a) the number of active VDREs on the genome is cell-type specific, b) VDREs are predominantly consensus DR3-type, c) enhancers are located promoter-proximal, promoter-distal or a combination, relative to the TSS; many enhancers are located in clusters hundreds of kilobases from their target genes (but may not be upon chromatin folding), d) enhancers are modular in nature, having binding sites for a number of different TFs.

1.2.3 *Vitamin D - role in calcium/phosphorous homeostasis and bone health*

The most widely accepted physiological function for the hormone, vitamin D, mediated mainly by calcitriol, is the modulation of calcium and phosphate homeostasis and bone mineralisation (Holick 2007, Verstuyf *et al.* 2010).

Calcitriol enhances calcium and phosphate absorption in the small intestine. In particular, it is known to modulate multiple gene products involved in the trans-epithelial transport of calcium in the intestine. Similarly, calcitriol regulates the expression of carriers responsible for the transportation of phosphate (Ajibade *et al.* 2010).

Adequate levels of calcium and phosphate in the blood promote the mineralisation of the skeleton. Calcitriol enhances the release of calcium and phosphate from bone to maintain their levels in the blood (Pike *et al.* 2012).

1.2.4 *Vitamin D – other health effects*

Research in the early 1980s revealed that calcitriol was able to modulate cellular proliferation and differentiation and function in a diverse array of cell types not directly involved in mineral/bone homeostasis (Abe *et al.* 1981). The modulation of these processes in the various cell populations result in a variety of actions including the control of bile acid metabolism (Makishima *et al.* 2002) and xenobiotic chemical degradation in the intestinal tract (Schmiedlin-Ren *et al.* 2001, Thummel *et al.* 2001),

control of keratinocyte differentiation in the epidermis, and of barrier function (Bikle *et al.* 2010), the regulation of sebaceous gland development and post-natal hair follicle cycling (Cianferotti *et al.* 2007), maintenance of normal blood pressure and cardiovascular integrity (Wu-Wong 2007, Wu-Wong 2009) and the regulation of both the innate and acquired immune systems (Hewison 2011).
Epithelial and immune cells of the gut mucosa express both vitamin D receptor (VDR) (section 1.3.4.2) and *CYP27B1*, signaling the potential for calcitriol to have a significant role in the regulation of the immune system of the colon (Verstuyf *et al.* 2010). A number of studies have demonstrated that calcitriol and VDR are involved in the innate immune response against bacterial infections in the colon (Lagishetty *et al.* 2010).

In the vitamin D deficient state (reviewed by (Hock 2014)), cathelicidin secretion is reduced, leading to impaired autophagy and consequently, dysregulation of phagocytosis, cytotoxicity and antigen processing and presentation. Vitamin D deficiency also impacts T- and B-lymphocyte induction and the number and performance of regulatory natural killer T-cells in the gastro-intestinal tract (Hock 2014).

Deficiency of calcitriol has also been linked with the incidence of a variety of chronic illnesses including cancers of the colon, breast, prostate, ovaries and oesophagus, and also including type I diabetes, hypertension, congestive heart failure and rheumatoid arthritis (Holick 2005b).

1.2.5 *Recommendations for vitamin D for optimal health*

Vitamin D levels are best quantified by measuring serum levels of the main circulating form of vitamin D, calcidiol, which captures systemic vitamin D acquired through both dietary sources and UVB irradiation. However, it is not clear to what extent calcidiol serum levels are representative of local vitamin D tissue status (Pereira *et al.* 2012).

Early studies regarding systemic vitamin D requirements focused on the prevention of rickets in children in industrialised communities in the 1800s (Prentice 2013). Vitamin D levels for bone health were established as serum calcidiol levels of <25 nmol/L being deficient and normal levels within the range of 40-50 nmol/L. Later studies investigating vitamin D requirements for risk reduction, for a variety of diseases, have

recommended calcidiol levels of 50 nmol/L for bone health and in excess of 80 nmol/L for optimal health benefits (Barger-Lux *et al.* 1998, Bouillon 2011, Gozdzik *et al.* 2008, Holick 2005a).

The US Institute of Medicine considers serum calcidiol levels of 50 nmol/L to be normal (Ross *et al.* 2011), whilst the US Endocrine Society defines serum calcidiol levels under 50 nmol/L as vitamin D deficient, levels between 50-75 nmol/L as vitamin D insufficient and levels above 75 nmol/L as vitamin D sufficient (Holick *et al.* 2011).

A study carried out in early 2012 looking at the lifestyles of two Tanzanian tribes found a mean serum calcidiol concentration of 115 nmol/L. The two Tanzanian tribes both experienced life-long, year-round exposure to tropical sunlight and wore moderate clothing and no sunscreen (Luxwolda *et al.* 2012). Holick suggests that 115 nmol/L should be the normal blood calcidiol level for optimal health (Holick 2013).

1.2.6 *Sunshine and vitamin D*

Skin exposure to UVB radiation provides upwards of 90% of the body's requirement for vitamin D (Holick 2011). During exposure to sunlight, UVB photons (wavelength 290-315 nm) penetrate the skin where they are absorbed by 7-DHC. 7-dehydrocholesterol is present in the plasma membrane of the cells of the dermis and epidermis. The absorption of UVB radiation causes 7-DHC to open its B ring, forming pre-vitamin D_3 (section 1.2.1) (Holick 2003).

Wavelengths of 290-315 nm are present in sunlight when the UV index is greater than 3. The UV index is 0 when dark and 10 at midday on a clear spring/summer day in the tropics. A UV index of 3 occurs on a sunny spring/summer day in temperate climates. A UV index of 11 is obtained in summer at low latitudes (approximately 0^{o}-25^{o}) or in areas of above average ozone layer depletion.

The quantity of UV photons that penetrate to the earth's surface is determined by a number of factors including the path length through the stratospheric ozone layer and the distance the UV radiation must travel through the atmosphere as a function of the solar zenith angle. The solar zenith angle is dependent on latitude, season and time of day (Holick 1995).

Holick suggested (in 2007) that at latitudes around 40°, a fair skinned person should attain maximum vitamin D requirements on a clear summer day by 5-10 minute exposures, 2-3 times weekly, on the face and forearms around midday. This recommendation was made on the basis that exposure of a young adult body to one minimum erythema dose (the lowest dose that produces pink erythema with distinct borders) of solar UVB radiation elevates the circulating level of calcidiol to a peak of approximately 52 nmol/L (Holick 2007). It was suggested that the time should be increased to 30 minutes for dark skin (see below) or on a cloudy summer day. Other reports have reached similar conclusions (Jones and Dwyer 1998, Reid *et al.* 1986). There is significant debate regarding optimal calcidiol serum levels at present. Bone health experts suggest aiming for calcidiol serum levels of 50 nmol/L whilst other health experts advocate for levels of 80 nmol/L or as high as 115 nmol/L (section 1.2.5).

Further UVB exposure results in conversion of pre-vitamin D_3 to the inactive isomers, tachysterol and lumisterol, serving as a reservoir for pre-vitamin D_3 (Webb and Holick 1988). Excess vitamin D is deposited in the adipose tissue. Vitamin D may be slowly released when calcidiol serum levels are low and adipose cell turnover occurs (Rosenstreich *et al.* 1971). Obese individuals do not have higher serum calcidiol levels, as might be expected, as vitamin D deposited in the adipose tissue may not be released due to relatively immobile fat stores (Arner *et al.* 2011, Liel *et al.* 1988, Wortsman *et al.* 2000). Weight reduction studies report that serum calcidiol concentrations increase when obese individuals lose body fat (Riedt *et al.* 2005, Tzotzas *et al.* 2010, Zittermann *et al.* 2009).

Excessive exposure to UVB radiation does not result in vitamin D toxicity due to a number of processes including conversion of pre-vitamin D_3 to inactive isomers, storage of vitamin D in adipose tissue and the degradation of vitamin D precursors once equilibrium is reached (Holick 1995).

Exposure to UVB radiation induces melanin production in the skin and higher concentrations of melanin in the skin have been linked with lower vitamin D synthesis. It has been proposed that melanin's broadband UV absorbent and photon reflective properties may equip melanin as a photo-protective agent (Brenner and Hearing 2008).

Consequently, as epidermal melanin content increases epidermal photo-protection may increase such that increased sun exposure would be required to produce the desired serum calcidiol levels. The theory that melanin interferes with vitamin D production is supported by a study undertaken by Matsuoka et al. Exposure of the entire body to UVB irradiation resulted in significantly higher serum calcidiol levels in white Americans when compared to both Black African Americans and South Asians (Matsuoka *et al.* 1991).

A number of studies have investigated the impact of latitude on vitamin D production and have found that at moderate to high latitudes (>40^{o}) calcidiol serum levels decrease. In a study by Engelsen it was shown that at latitudes of greater than 50^{o}, vitamin D synthesis was absent for between 3-4 months due to the low solar zenith angle (Engelsen 2010). Similarly, Matsuoka *et al* demonstrated that at latitudes above 40^{o}, the threshold level of UVB radiation required to induce vitamin D synthesis is not reached during the winter months (Matsuoka *et al.* 1989). However, storage of calcidiol in the adipose tissue during the spring, summer and autumn months may compensate for lack of vitamin D production in the winter (Creighton University 2015). Based on this, it could be argued that humans have evolved to store vitamin D in the fat tissue being laid down during the 'surplus seasons' of spring, summer and autumn, to be used in the 'leaner' winter season where vitamin D could be released with adipose cell turnover.

The cohort for the study, that is the subject of this thesis, was recruited from the Southern DHB region of NZ, situated approximately 44^{o} – 47.5^{o}S, well within the region demonstrated by other researchers to have inadequate UVB radiation during the winter months (Engelsen 2010, Matsuoka *et al.* 1989).

Commercial sunscreens absorb UV radiation and so prevent UV rays from penetrating the skin. A sunscreen with a UVB sun protection factor (SPF) of 8 can decrease the vitamin D synthetic capacity by 95%. A sunscreen with a UVB SPF of 15 can reduce the vitamin D synthetic capacity by 98% (Shrapnel and Truswell 2006). Glass almost completely blocks UVB radiation (Holick 1995).

Whilst it is generally accepted that UVB wavelengths of 290-315 nm produce optimal levels of pre-vitamin D_3, further research is needed to ascertain whether UVA radiation

is able to produce pre-vitamin D_3 in the skin. Research is also required to clarify the ideal wavelengths of UVB radiation, and UVA if applicable, for optimal pre-vitamin D_3 synthesis (Norval *et al.* 2010).

1.2.7 *Diet and vitamin D*

Ninety to ninety-five percent of an average person's daily vitamin D requirement is achieved through casual sun exposure, except at high latitudes during the winter months (section 1.2.5).

Vitamin D is naturally present in a very limited selection of food sources. Oily fish, including salmon, tuna, mackerel and herring, cod liver oil and sun-dried mushrooms typically provide 400-500 IU (10-12.5 mcg) per serving, translating to serum calcidiol levels of 10-12.5 nmol/L (Holick 2005a). Egg yolks and liver contain smaller amounts of vitamin D. Fortification of foods with vitamin D in New Zealand is limited, unlike the situation in the United States (US), Canada, the United Kingdom and Europe. Consequently, average daily vitamin D intakes in NZ are substantially below intakes in countries where fortification of food with vitamin D occurs (Shrapnel and Truswell 2006). A variety of supplements are available that contain varying quantities of vitamin D.

Dietary recommendations for vitamin D from the National Health and Medical Research Council of Australia assume minimal sun exposure and advocate a vitamin D intake of 5-15 mcg/day depending on age. This is equivalent to 200-600 IU/day and translates to calcidiol serum levels of 5-15 nmol/L (NHMRC and MOH 2005).

Vitamin D toxicity is typically not observed until serum levels exceed 375 nmol/L (Koutkia *et al.* 2001), and is more likely to occur from high intakes of dietary supplements than from high intakes of vitamin D rich or vitamin D fortified foods (Ross *et al.* 2011). Toxicity is not associated with UVB induced synthesis of vitamin D. The body has a number of mechanisms to deal with excess vitamin D as it is being produced (section 1.2.5). Excessive vitamin D leads to increased calcium levels, which can result in calcinosis (the deposit of calcium salts in soft tissues of the body, such as the kidneys, heart and lungs) and hypercalcaemia (high blood levels of calcium).

1.2.8 *Vitamin D levels in New Zealand*

In 1996/97 the New Zealand Health Survey, a population-based, nationwide survey of 6999 adolescents and adults aged 15 years and over, was undertaken. The 1997 National Nutrition Survey was conducted as an extension of the 1996/97 New Zealand Health Survey, where 4636 participants of the initial survey agreed to be involved in the nutritional extension survey. Serum samples were collected from 3369 participants. Surplus serum was available for 3008 of the 3369 serum samples collected. Vitamin D status was quantified by measuring calcidiol concentration in the surplus serum samples. Of the 3008 participants, 2946 had all of the required data for inclusion in a study published by Rockell et al. in 2006. Forty-eight and eighty-four percent of adults were reported to be vitamin D insufficient based on cut-offs of ≤50 and ≤80 nmol/L respectively. In participants aged ≥ 65 years the mean calcidiol levels were found to be insufficient, lower in women (43 nmol/L) than in men (55 nmol/L). Mean serum calcidiol levels were found to be lower in Māori (42 nmol/L) and Pacific Islanders (37 nmol/L) when compared to people of other ethnicities (51 nmol/L). Mean calcidiol levels changed markedly with season. The difference between spring and summer levels was 28 nmol/L (range 25-31) in men and 31 nmol/L (range 28-34) in women. This seasonal difference exceeds those found between New Zealanders of differing gender, age, body mass index (BMI) or ethnicity suggesting that UV exposure may be the prime determinant of calcidiol status in New Zealanders (Rockell *et al.* 2006).

In 2007, Livesey et al. conducted a study in Christchurch investigating circulating calcidiol levels alongside environmental UV exposure. Based on their study results, they postulated that most, if not all, of the healthy general population of Christchurch do not sustain adequate blood levels of calcidiol (>75 nmol/L) throughout the year. By July and August, 35% of the volunteer group was vitamin D deficient (<25 nmol/L) (Livesey *et al.* 2007). In a similar study undertaken in Auckland around the same time, circulating calcidiol levels were higher than in the Christchurch cohort, particularly in men. The mean calcidiol level for men was 102 nmol/L in March and 59 nmol/L in September. The higher circulating calcidiol levels in the Auckland cohort were likely due to the lower latitude (37^{o}S) but the researchers had no explanation for why the levels were markedly higher in men (Bolland *et al.* 2006).

Studies have been conducted in groups of women in Auckland. Interestingly, results from one study suggested that South Asian women were at high risk of vitamin D deficiency due to an indoor lifestyle and intentional sun avoidance (von Hurst *et al.* 2010).

Whilst differences between Māori and European New Zealanders' serum calcidiol levels are not statistically significant, it has been hypothesised that epidermal pigmentation is the predominant contributing factor to the differences seen (section 1.2.5).

Vitamin D deficiency in New Zealand may be due to an increasingly indoor lifestyle and use of sunscreens in a country where a high incidence of malignant melanoma is well publicised. Furthermore, populations living at latitudes greater than 40^{o} are probably unable to synthesise vitamin D for 3 to 4 months during the winter (section 1.2.6). In New Zealand this equates to south of Wanganui.

1.3 Vitamin D and Colorectal Cancer

1.3.1 *Population studies*

1.3.1.1 *Ecological studies*

In 1980, Garland and Garland demonstrated higher rates of colon cancer mortality in regions of the US with low solar radiation, suggesting that this observation may be due to lack of vitamin D synthesis in the skin (Garland and Garland 1980). In addition, deaths from colon cancer were higher in industrialised areas compared with rural areas. Later, the same authors reported an inverse association between vitamin D status and CRC in the US (Garland *et al.* 1989).

A more recent study in the US, examining UV radiation data acquired for July 1992, demonstrated an inverse association between UV radiation and colon and rectal cancer mortality for the years 1970 to 1994 (Grant 2002).

Ecological studies in Japan (Mizoue 2004) and China (Chen *et al.* 2010) also showed an inverse association between UV radiation and colon cancer mortality. In the Chinese study, and in line with the findings of Garland and Garland (Garland *et al.* 1989), an

inverse relationship was observed among rural residents, not urban residents, suggesting a lack of sun exposure linked to industrialisation (Chen *et al.* 2010).

In the Southern DHB region, lower solar radiation is experienced in the winter months (section 1.2.5). Furthermore, the incidence of CRC in the Southern DHB region is higher than the NZ rate of incidence for CRC (section 1.1.2).

The existence of confounding factors, dietary vitamin D, and misclassification of exposures, due to use of non-individual UV levels, are considerable limitations in these ecological studies.

1.3.1.2 *Epidemiological studies*

A systematic review of cohort studies revealed that dietary vitamin D from food was inversely correlated with CRC. The association for total vitamin D intake from food and supplements was not significant (Touvier *et al.* 2011). The risk ratios (RR) (95% CIs) for an increase of 100 IU/day were 0.95 (0.93-0.98) for dietary vitamin D from food and 0.98 (0.95-1.01) for total vitamin D from food and supplements. In the concurrent meta-analysis of total vitamin D and CRC, when one study was excluded, heterogeneity was reduced and the summary RR became statistically significant (RR=0.97, 95% CI=0.95-0.99). Findings have been variable for vitamin D supplement use, which could be due to behavioural confounding factors.

The integration and overall analysis of relevant epidemiological studies is limited. Many studies do not account for endogenous vitamin D synthesis from UV exposure and are restricted by measurement error of dietary vitamin D intake. Such shortcomings can be overcome by measuring circulating calcidiol concentration, which is a useful biomarker for providing an overall estimate of vitamin D status accounting for both UV exposure and dietary intake (Pereira *et al.* 2012).

The body of literature examining the relationship between serum calcidiol level and CRC risk supports an inverse association. Several nested case-control studies have examined whether high circulating calcidiol levels lowered the risk of developing CRC. Blood samples were collected prior to CRC diagnosis. The majority of the study populations were Caucasian (Braun *et al.* 1995, Feskanich *et al.* 2004, Garland *et al.* 1989, Jenab *et al.*

2010, Tangrea *et al.* 1997, Wactawski-Wende *et al.* 2006, Wu *et al.* 2007), one was a multi-ethnic population (Woolcott *et al.* 2010) and another was Japanese (Otani *et al.* 2007). Summary evidence indicated an inverse association for CRC. A stronger association was seen for rectal cancer. When prospective studies were included, analysis of 1822 colon and 868 rectal cancers suggested that elevated circulating calcidiol levels were associated with a significant reduction in CRC (OR=0.66, 95% CI=0.54-0.81 for upper versus lower categories). The inverse association was stronger for rectal cancer (OR=0.5, 95% CI=0.28-0.88 for upper versus lower categories) (Lee *et al.* 2011).

A meta-analysis of 35 independent studies supported an inverse association between serum calcidiol levels and the risk of CRC (Gandini *et al.* 2011). A systematic review of 18 prospective studies investigating the association of vitamin D intake or serum calcidiol levels and CRC risk was conducted with a total of 1,000,000 individuals. An inverse association between risk of CRC and both vitamin D intake and serum calcidiol level was found (Ma *et al.* 2011).

A prospective, observational study, demonstrated that in patients with CRC, elevated pre-diagnosis calcidiol levels were correlated with a significant improvement in overall survival (Ng *et al.* 2008). The study included 304 patients from the Nurses' Health Study (NHS) and the Health Professionals Follow-Up Study (HPFS) who were diagnosed with CRC between 1991 and 2002. Participants were observed until death, June 2005 for the NHS or January 2005 for the HPFS. Patients in the upper-most quartile had an adjusted hazard ratio (HR) for overall mortality of 0.52 (95% CI=0.29-0.94) compared with those in the lowest quartile. Patients in the upper-most quartile had an adjusted HR for CRC-specific death of 0.61 (95% CI=0.31-1.19) compared with those in the lowest quartile.

A prospective observational study found that CRC patients with post-surgery calcidiol levels in the upper tertile had improved survival (Zgaga *et al.* 2014). This Scottish study included 1598 stage I–III patients. To mitigate the potentially confounding effect of reverse causality (disease process causing vitamin D deficiency), the researchers tested for interaction between calcidiol level and genotype at the VDR locus. Interactions were detected between vitamin D concentration and the rs11568820 genotype for CRC-specific mortality (P=0.008) and the number of protective alleles for CRC-specific mortality (P=0.004). The gene-environment interaction between vitamin D

concentration and genetic variation at the VDR locus may indicate that vitamin D is able to actively protect against death.

Giovannucci proposed that the consistency of the relationship between vitamin D status and CRC risk signifies a causal association, despite the issues of population heterogeneity, the presence of confounding factors and using disparate approaches and techniques to measure vitamin D status (Giovannucci 2011). Recently, Giovannucci re-reviewed the epidemiology of CRC and vitamin D, commenting that despite employing numerous approaches to quantify vitamin D for diverse endpoints and in diverse populations the inverse association was highly consistent and therefore strongly indicative of a causal relationship (Giovannucci 2013). In accordance with this, the IARC stated that the epidemiological evidence for an association between the incidence of CRC and serum calcidiol levels was consistent and persuasive, but that large, prospective clinical trials were required to prove causality (IARC 2008).

1.3.2 *Human experimental studies*

To date, intervention trials examining vitamin D in human populations with CRC as a primary endpoint have not been conducted. Several studies have been undertaken examining vitamin D with CRC as a secondary endpoint. Two intervention studies with molecular endpoints have also been included in this section.

The first of the studies is a randomised, double-blind, placebo-controlled trial carried out in 36,282 post-menopausal women (50-79 years) in the US Women's Health Initiative. Participants received 400 IU/day vitamin D plus 1 g/day calcium or placebo, for an average of seven years. No effect on the incidence of CRC was seen. Unfortunately, the study suffered from several limitations including low dose of vitamin D, lack of patient protocol adherence and short follow-up period (seven years) (Wactawski-Wende *et al.* 2006). However, re-analysis of the data revealed that concurrent estrogen therapy increased CRC risk. Treatment with vitamin D plus calcium in the absence of estrogen was found to be beneficial (RR= 0.71; CI= 0.46-1.09) (Ding *et al.* 2008). A second published randomised controlled trial in 1180 post-menopausal women receiving 1100 IU/day plus 1.4-1.5 g/day calcium (or calcium alone) for four years failed to show any difference in CRC rates. However, the study was underpowered for this endpoint (Lappe *et al.* 2007).

Five large-scale trials are underway to examine the potential benefits of vitamin D in a number of chronic diseases. Two of these trials include cancer as a main outcome, the U.S VITAL study and the Finnish FIND study. In the VITAL study, 20,000 participants will receive either 2000 IU/day of vitamin D (expected to raise blood levels of calcidiol to ≥75nmol/L in most participants) or placebo for five years. Results from the VITAL trial are not expected until 2017. In the FIND study, 18,000 participants will receive either 1600 or 3200 IU/day of vitamin D or placebo for five years. Results from the FIND study are not expected until 2020 (Kupferschmidt 2012).

A New Zealand clinical trial, the ViDA study, is underway to determine whether vitamin D supplementation prevents cardiovascular and respiratory disease and also whether it prevents fractures. In addition, there are a number of secondary outcomes including CRC. The study has recruited 5100 participants aged 50-84 years. Participants will receive either 100,000 IU/month of vitamin D or placebo for four years. Results from the ViDA study are expected in 2016 (University of Auckland 2012).

In a number of human studies where vitamin D supplements have been given, some useful molecular data has been collected. In a study by Fedirko *et al*., 92 patients with at least one adenoma were randomised to one of four supplementation arms, 800 IU/day vitamin D and 2 g/day calcium, same dose vitamin D alone, same dose calcium alone or placebo for six months. The aim of the study was to develop biomarkers for CRC by firstly investigating their performance in normal rectal mucosa. The biomarkers selected included *BAX*, a regulator of apoptosis, and *p21,* a potent inducer of differentiation in colonocytes (among other roles). In response to vitamin D ($p=0.02$), *BAX* expression was enhanced in colorectal crypts. Similarly, in response to vitamin D ($p=0.005$) or calcium ($p=0.03$), *p21* expression was elevated in colorectal crypts (Fedirko *et al.* 2009a, Fedirko *et al.* 2009b). Furthermore, decrease in the expression of pro-inflammatory cytokines TNF-α, IL1β, IL6, and IL8, and the pro-inflammatory marker C-reactive protein was observed in the group supplemented with vitamin D alone (Hopkins *et al.* 2011). This result was not statistically significant but the overall inflammation z-score (sum of all measured inflammation biomarker scores for each participant) decreased significantly by 77% ($p=0.003$).

In 2006 Holt *et al*. published a study where consenting patients (19 of the >400 potential participants) underwent scheduled colonoscopies with partial polypectomy, after polyps were found during screening sigmoidoscopy. Biopsies of normal rectal mucosa were taken at colonoscopy. Patients were randomised to either 400 IU/day of vitamin D and 1.5 g/day of calcium (eleven patients) or placebo (eight patients) for six months. At study end, patients underwent proctosigmoidoscopy for completion of polyp removal and repeated normal rectal mucosa biopsy retrieval. Pre and post study-treatment normal rectal mucosa and colorectal polyps were tested for biomarkers of proliferation, apoptosis, colorectal adenoma and colon carcinoma progression. Proliferation of both normal rectal mucosa ($p<0.001$) and polyps ($p=0.025$ or $p<0.02$ depending on test) was significantly suppressed in the treatment group compared with the control group. Expression of *BAK1* (encodes pro-apoptotic protein) was enhanced ($p<0.005$) and expression of CRC-associated mucin (*MUC5AC*) was inhibited ($p<0.005$) in polyps when compared with the placebo group (Holt *et al.* 2006).

Autier *et al*. carried out a systematic review and meta-analyses examining prospective and intervention studies that assessed the impact of serum calcidiol levels on non-skeletal health outcomes in individuals aged 18 years and over. Prospective studies have shown associations between reduced serum calcidiol level and increased risk of many health outcomes including CRC. Randomised controlled trials (RCT) have not been able to substantiate the association between low calcidiol level and increased risk of disease. They concluded that the association between low serum calcidiol levels and increased risk of disease, including CRC, was not causal. It was hypothesised that low calcidiol concentrations may be precipitated by the disease process rather than increasing the risk of the disease. They also postulated that the common factor between low calcidiol levels and most non-skeletal health disorders was inflammation. With respect to CRC, they observed that vitamin D supplementation did not reduce the risk of CRC whereas non-steroidal anti-inflammatory medications did appear to reduce the risk (Flossmann *et al.* 2007). Higher levels of calcidiol at cancer diagnosis are associated with better outcomes, but Autier *et al*. did not find evidence of causality in their meta-analysis of clinical trials (Autier *et al.* 2014).

1.3.3 *Animal studies*

Studies conducted in experimental animals support both protective and therapeutic effects of vitamin D in CRC. Delivering a western style diet (high in fat, low in vitamin D and calcium) to wild-type mice induced colonic crypt hyperplasia and colon dysplasia. The number of pre-neoplastic lesions was elevated in intestinal carcinogenesis mouse models. Supplementation with vitamin D and calcium suppressed these effects in the mice (reviewed in (Lamprecht and Lipkin 2003) and (Ordonez-Moran *et al.* 2005).

Human CRC xenografts, established by implanting CRC cell lines subcutaneously into immuno-suppressed mice, are frequently utilised in pre-clinical anti-cancer drug development. Several studies have demonstrated the inhibitory nature of calcitriol towards the growth of colorectal xenografts (reviewed in (Ordonez-Moran *et al.* 2005), (Deeb *et al.* 2007) and (Kang *et al.* 2011). Numbers of colorectal tumours induced in mice and rats by various chemical carcinogens were also decreased by the administration of calcitriol (reviewed in (Ordonez-Moran *et al.* 2005).

The *APC* $^{min/+}$ mouse is a model of intestinal tumouri-genesis that carries a mutated allele of *APC* and spontaneously grows numerous neoplasias throughout the intestinal tract (Su *et al.* 1992). Administration of calcitriol decreases tumour load in *APC* $^{min/+}$ mice (Huerta *et al.* 2002). These findings have been confirmed more recently and the researchers also demonstrated that calcitriol administration reduces nuclear β-catenin levels, down-regulates the expression of *MYC* (encodes a transcription factor with roles in cell cycle progression and apoptosis) and elevates that of E-cadherin (epithelial cadherin, a cell-cell adhesion glycoprotein) in the colon of *APC* $^{min/+}$ mice (Xu *et al.* 2010).

The effect of disruption of VDR expression in colorectal carcinogenesis has also been investigated in genetically modified mice. VDR-deficient mice are susceptible to oncogene- and carcinogen-induced tumours (reviewed in (Bouillon *et al.* 2008)). Recently, two groups have sought to ascertain the impact of VDR deficiency on the initiation and progression of CRC in *VDR* $^{-/+}$ and *APC* $^{min/+}$ mice (Larriba *et al.* 2011, Zheng *et al.* 2012). These studies reported elevated tumour burden in *VDR* $^{-/-}$ *APC* $^{min/+}$ mice as compared to *VDR* $^{+/+}$ *APC* $^{min/+}$ mice. Enhanced activation of the WNT/β-catenin pathway (leading to increased cell proliferation) and an increase in the number of atypical colonic crypt foci were also apparent in the *VDR* $^{-/-}$ *APC* $^{min/+}$ mice (Larriba *et al.* 2011).

1.3.4 *Molecular studies*

1.3.4.1 *Tumouri-genesis in colorectal cancer*

Normal colonic epithelial cells undergo transformation and malignant progression as a result of the progressive accumulation of genetic and epigenetic changes. The somatic mutation in the tumour suppressor gene, *APC* is the earliest event in most colorectal adenomas. This mutation leads to the aberrant activation of the established WNT/ β-catenin signaling pathway (impacts cell proliferation) (Clevers 2006, Klaus and Birchmeier 2008). In conjunction, a substantial proportion of adenomas sustain activating mutations in *BRAF* (cell growth regulator) or *KRAS* (signal transduction pathways). Inactivating mutations in the transforming growth factor-β (TGF-β) pathway (*SMAD2, SMAD4* or *TGFBR2*) confer further malignant traits on adenoma cells (Markowitz and Bertagnolli 2009). In approximately 50% of CRC tumours, the adenoma-carcinoma transition is linked to the inactivation of the *TP53* tumour suppressor gene (Iacopetta 2003). The molecular mechanisms of tumour development and the acquisition of metastatic capacity are as yet not fully elucidated.

1.3.4.2 *Anti-tumoural activity of calcitriol in colorectal cancer cells*

The anti-tumoural activity of calcitriol in CRC depends on various processes at the cellular level. These processes include suppression of proliferation and angiogenesis and induction of differentiation, apoptosis and cell detoxification metabolism. The integrated impact of these processes, in a cell-type and cell-context dependent mode, may determine the anti-tumoural activity of calcitriol (reviewed in (Lamprecht and Lipkin 2003, Larriba *et al.* 2008) and (Krishnan and Feldman 2011)).

The earliest reported anti-neoplastic effects of calcitriol include the anti-proliferative and pro-differentiating effects on cancer cells *in vitro* and *in vivo* (in CRC: (Cross *et al.* 2003, Díaz *et al.* 2000, Huerta *et al.* 2002)).

1.3.4.2a *Proliferation*

In CRC, calcitriol regulates many genes involved in proliferation, including *GADD45A*, *MYC*, *FOS* and *JUN* (Meyer *et al.* 2012, Palmer *et al.* 2001, Pálmer *et al.* 2003). *GADD45A* is up-regulated by calcitriol and impacts cell cycle arrest following DNA damage. The c-*MYC* oncogene is over-expressed in most cancer types. Its suppression by calcitriol

occurs directly, via interaction with multiple VDREs and indirectly, via antagonism of the WNT/β-catenin signaling pathway (Palmer *et al.* 2001). Calcitriol directly inhibits β-catenin transcriptional activity in colon cancer cells by stimulating VDR binding to β-catenin. This prevents β-catenin translocation to the nucleus, precipitating its accumulation at the plasma membrane and eventual removal from the cell (Pendas-Franco *et al.* 2008a). In addition, inhibition of the WNT pathway, by calcitriol, occurs via its up-regulation of DICKKOPF-1, an extracellular WNT pathway inhibitor (Pendas-Franco *et al.* 2008b).

Calcitriol can modify cellular growth by affecting other important signaling pathways. For example, the TGF-β signaling pathway is activated by calcitriol and contributes to its anti-proliferative effects (Chen *et al.* 2002).

A study conducted in CRC cells indicated that calcitriol-mediated anti-proliferative activity is dependent on the dual role of the VDR, firstly as a transcription factor and secondly as a non-genomic initiator of the Rho-ROCK-p38MAPK-MSK (RRpMM) signaling pathway. This pathway has roles in antagonising the WNT/β-catenin pathway and in inducing an adhesive epithelial phenotype (Ordonez-Moran *et al.* 2008).

1.3.4.2b *Differentiation*

Calcitriol has numerous pro-differentiation effects in CRC cells. One of these effects is the up-regulation of expression of cell adhesion structures essential for the maintenance of the epithelial phenotype (Brehier and Thomasset 1988, Giuliano *et al.* 1991, Halline *et al.* 1994). These cell adhesion structures include E-cadherin in adherens junctions, actin cytoskeleton linked proteins and intermediate filaments (Palmer *et al.* 2001, Palmer *et al.* 2003). Calcitriol induces E-cadherin expression in conjunction with its disruption of β-catenin transcriptional activity (section 1.3.4.2f) (Pendas-Franco *et al.* 2008b). It also activates the RRpMM signaling pathway necessary for *CDH1* (encodes E-cadherin) induction and therefore the adhesive phenotype (Ordonez-Moran *et al.* 2008).

1.3.4.2c *Apoptosis*

Sensitisation to apoptosis by calcitriol in colorectal carcinoma cells involves enhancing the expression of the pro-apoptotic protein BAK1 and diminishing the expression of the anti-apoptotic protein BAG1 (Barnes *et al.* 2005, Díaz *et al.* 2000). Calcitriol also

stimulates apoptosis in CRC cells by promoting the expression of GOS2 (Palmer *et al.* 2003). GOS2 is a protein that blocks BCL-2 from forming anti-apoptotic heterodimers with BAX (a pro-apoptotic protein) (Welch *et al.* 2009).

Previous studies have indicated that sensitisation to apoptosis by calcitriol does not require an intact *TP53* tumour suppressor gene (Díaz *et al.* 2000, Hansen *et al.* 2001). However a recent study in mice has demonstrated that mutant p53 protein interacts with VDR and regulates the transcriptional activity of calcitriol. This regulation induces up-regulation of anti-apoptotic gene expression and down-regulation of pro-apoptotic gene expression, thereby converting calcitriol to an anti-apoptotic agent (Stambolsky *et al.* 2010).

1.3.4.2d *Angiogenesis*

Calcitriol may impact the angiogenic capacity of CRC cells through its interactions with hypoxia-inducible factor-1$^{\alpha}$ (HIF-1), a principal TF in hypoxia-induced angiogenesis. Calcitriol down-regulates the expression and transcriptional effects of HIF-1 (Ben-Shoshan *et al.* 2007). In addition, calcitriol strongly inhibits DICKKOPF-4, a weak WNT antagonist that stimulates angiogenesis and spread in cultured CRC cells. DICKKOPF-4 is over-expressed in human colon carcinomas (Pendas-Franco *et al.* 2008b).

1.3.4.2e *Detoxification*

Enterocytes express various enzymes involved in the detoxification of xenobiotics from both endobiotic and dietary sources that may be involved in the development of CRC. Phase I enzymes, predominantly members of the cytochrome P450 (CYP) super-family, initiate the detoxification of xenobiotics (Kaminsky and Zhang 2003). In CRC cells, calcitriol stimulates the expression of *CYP3A4*, the member of the CYP super-family primarily expressed in the intestinal tract (Thompson *et al.* 2002, Thummel *et al.* 2001). Phase II enzymes bind different charged groups to the products of phase I enzymes expediting their excretion (Kaminsky and Zhang 2003).

1.3.4.2f *Antagonism of the WNT/β-catenin pathway*

Calcitriol impedes several signaling pathways. Researchers have demonstrated that, in CRC cells, calcitriol suppresses the WNT/β-catenin pathway and consequently the activation of its target genes. These actions contribute to the antagonism of cell

proliferation and to the preservation of the differentiated phenotype (Larriba *et al.* 2007, Palmer *et al.* 2001). Antagonism of the WNT/β-catenin pathway by calcitriol occurs through various processes. Initially, it rapidly elevates the level of VDR bound to β-catenin, decreasing the interplay between β-catenin and the TFs of the T-cell factor/lymphoid enhancer factor (TCF/LEF) group and precipitating the inhibition of its target genes. Secondly, it causes β-catenin nuclear export coupled to E-cadherin collection at the plasma membrane adherens junctions, further disrupting the interplay between β-catenin and the TFs of the TCF/LEF family and thus causing additional inhibition of its target genes (Palmer *et al.* 2001). Lastly, it enhances the expression of DICKKOPF-1, an extracellular WNT inhibitor (Aguilera *et al.* 2007). Given that the WNT/β-catenin pathway is de-regulated in most adenomas and colorectal carcinomas and is regarded as the fundamental driving force in this neoplasia, its antagonism is probably critical for the anti-tumoural impact of calcitriol in CRC.

1.3.4.3 *Calcitriol target genes in colorectal cancer cells*

The ongoing investigation of calcitriol effects in the transcriptome of cancer cells has been enabled by the development of high-throughput technologies for gene expression profiling (reviewed in (Kriebitzsch *et al.* 2009)). In one study, oligonucleotide microarrays profiling CRC cells treated with calcitriol demonstrated enhanced regulation of two-thirds of the vitamin D transcriptome. Regulation of the remaining one-third was diminished. The genes regulated by calcitriol corresponded to a variety of functions, with a substantial proportion associated with transcription, metabolism and cell adhesion (Palmer *et al.* 2003).

1.3.4.4 *Mechanisms of resistance to calcitriol action in colorectal cancer*

CRC sensitivity to calcitriol is determined principally by the availability of calcitriol and the expression of VDR. Intracellular calcitriol levels are dependent on circulating levels of calcidiol and the activity of both *CYP27B1* (converts calcidiol to calcitriol) and *CYP24A1* (catabolises calcitriol) within the cell. *CYP27B1* expression and activity are enhanced in the early stages of carcinogenesis (adenomas, polyps) and radically suppressed in advanced colorectal tumours. *CYP24A1* expression and activity are enhanced in the later stages of CRC. Resistance to calcitriol occurs as a result of the alterations in calcitriol synthesis and accelerated calcitriol inactivation in advanced colorectal carcinomas (Bareis *et al.* 2001, Bises *et al.* 2004, Cross *et al.* 2001).

Similarly, VDR expression is elevated in the early stages of colorectal carcinogenesis whereas it is diminished in advanced stages causing unresponsiveness to calcitriol (Anderson *et al.* 2006, Cross *et al.* 2001, Larriba and Munoz 2005, Matusiak *et al.* 2005, Sheinin *et al.* 2000). SNAIL1, a repressor of E-cadherin also suppresses the expression of VDR and consequently prevents calcitriol anti-tumoural activity in CRC cells in culture and in xenografts (Larriba *et al.* 2007, Pálmer *et al.* 2004). SNAIL2 also down-regulates VDR expression. In cultured cells and colorectal tumours SNAIL1 and SNAIL2 exhibit an additive repressive impact on VDR expression (Larriba *et al.* 2010, Larriba *et al.* 2009).

MicroRNA species miR-27b and miR-298 decrease VDR levels in LS-180 colon cancer cells (Pan *et al.* 2009). This post-transcriptional reduction in expression of VDR caused by miRNAs may be an additional process responsible for the low VDR levels reported in advanced CRC and may therefore also contribute to calcitriol resistance in CRC cells.

1.4 Summary, primary aim and chapter outline

The burden of CRC worldwide is significant with an estimated 1.36 million new cases of CRC and an estimated 694,000 deaths due to CRC in 2012. In New Zealand, incidence rates of CRC are projected to decline, decreasing overall by approximately one-quarter in the 45-74 age group. The overall burden is projected to continue increasing as a result of increased population growth and an aging population.

Vitamin D is essential for bone health. Recently, it has been postulated that vitamin D may be important for a number of other health outcomes. To maintain bone health and obtain benefits from other health effects, many experts advocate serum calcidiol concentrations of >80 nmol/L. In NZ, vitamin D levels are frequently below this level, particularly during the winter months.

Epidemiological studies have reported an inverse association between vitamin D status and incidence of CRC. Higher serum vitamin D levels at the time of diagnosis or post-surgery have been associated with improved long-term outcome in CRC. Supplementation with vitamin D was reported to suppress colorectal tumour growth in

mouse xenografts. Vitamin D reportedly alters the gene expression of genes involved in cell proliferation, cell differentiation and apoptosis.

The primary aim of the trial described in this thesis, was to determine whether a single large dose of vitamin D given within this short window would be sufficient to elevate vitamin D levels to recommended optimal levels and have a measurable biological effect on vitamin D mediated genes in normal colonic and colorectal tumour tissue removed at surgery.

Chapter 2 describes the design and the methods for the clinical outcomes of the trial. Chapter 3 presents the clinical results and other secondary outcomes. Chapters 4 and 5 address the primary aim of determining whether a single large dose of vitamin D could have an impact on vitamin D pathways in normal and tumour tissue. Chapter 4 describes the methods used to measure this question, which is based on analysis of differential gene expression in colorectal tumours and normal bowel obtained from the study participants. Chapter 5 presents the results of the differential gene expression analysis. Chapter 6 summarises the findings of the thesis.

2. TRIAL DESIGN

2.1 Introduction

Incidence rates of CRC are projected to decline in New Zealand. The overall burden is projected to continue increasing as a result of increased size and structural aging of the NZ population (MOH 2010). As discussed in chapter 1 (section 1.3.1.2), epidemiological studies have reported an inverse association between serum calcidiol levels and the incidence of colorectal cancer. In accordance with this, Southland and Otago have the highest incidence of CRC in New Zealand (MOH 2011a) as well as the lowest exposure to effective UVB irradiation, responsible for upwards of 90% of the body's vitamin D requirements (Holick 2011).

Higher serum calcidiol levels at the time of diagnosis or post-surgery have been associated with improved long-term outcome in CRC (Ng *et al.* 2008, Zgaga *et al.* 2014), but there is no proven causal link. Nor has it been determined whether vitamin D supplementation at the time of diagnosis could alter outcome in CRC. Most New Zealanders have serum calcidiol levels below 80 nmol/L (Rockell *et al.* 2006), which is considered sub-optimal for a range of health outcomes, other than bone health (Hossein-nezhad and Holick 2013). In a placebo-controlled, randomised trial of six months of vitamin D supplementation in healthy volunteers, colonic crypt cells in subjects randomised to vitamin D expressed increased levels of pro-apoptotic markers and had reduced proliferation rates (Fedirko *et al.* 2009b).

There is only a short period between diagnosis and treatment of colorectal cancer, and therefore only a short window available to intervene prior to definitive treatment. The study described here was designed to utilise this window to increase vitamin D levels, and examine tissues removed at surgery to determine if there was any measurable biological effect. There have been no previous studies to determine whether vitamin D pathways are activated following short duration vitamin D supplementation, in either normal colonic epithelium or tumour epithelium.

The current chapter describes the overall design of the trial and methods used to measure the secondary outcomes. The details of the molecular and bioinformatic methods are provided later, in Chapter 4.

2.2 Study Design

The study was a randomised, double-blind, placebo-controlled trial of vitamin D supplementation in patients undergoing elective surgery for CRC at Dunedin Hospital.

2.3 Participants

Patients with a pre-operative diagnosis of CRC, confirmed by colonoscopy or computerised tomography colonography (CTC), and scheduled to undergo elective surgery at Dunedin Hospital were considered for participation. Following approval by the clinical team, potential participants were approached by a member of the trial team.

Patients with acute presentations, such as bowel obstruction or perforation, were not eligible due to the requirement for a minimum of seven days between randomisation and administration of study drug, and surgery.

2.3.1 *Consent and ethics*

After being informed of the diagnosis and planned treatment by their medical team, potential participants were approached by a member of the study team. The study was explained and an information sheet provided. After they had had an opportunity to ask questions and consider their participation, written informed consent was obtained from those wishing to participate.

Ethical approval for the study was given by the Lower South Regional Ethics Committee (LRS/10/11/054).

This trial is registered with the Australian and New Zealand Clinical Trials Registry; reference number ACTRN12610000936022.

2.3.2 *Inclusion criteria*

Patients were included in the study if all of the following were applicable:

- Patients with CRC, confirmed by colonoscopy or CTC, and scheduled for elective surgery at Dunedin Hospital.
- Male or female, aged 18 years or over.
- Participant willing and able to give informed consent for recruitment into the study and willing and able (in the Investigators opinion) to comply with all of the requirements of the study.
- Participant willing to have their general practitioner notified of their involvement in the study if required.
- Patient serum calcium level <2.6 mmol/L, and serum calcidiol level <125 nmol/L.

The reference range used for calcidiol serum levels by the Southern Community Laboratories is 50–150 nmol/L. The reference range used for serum calcium levels by the Southern Community Laboratories is 2.05-2.60 mmol/L. Calcium levels above 2.6 mmol/L indicate hypercalcaemia. In the literature calcidiol serum levels above 125 nmol/L indicate hypervitaminosis D. Vitamin D supplementation can increase serum calcium level to hypercalcaemic levels thus it was important to determine normal range calcium level before vitamin D supplementation (Jones 2008).

2.3.3 *Exclusion criteria*

Patients were excluded from the study if any of the following were applicable:

- Aged less than 18 years of age.
- Concomitant use of prescription strength cholecalciferol.
- Participation in research studies in the previous 12 weeks involving investigational product.
- Pregnant or lactating.
- Hypercalcaemia (>2.6 mmol/L).
- Hyperparathyroidism.
- History of kidney stones.
- Renal failure (requiring renal replacement therapy).
- Advanced liver disease (Childs-Pugh B or C).

- Any other significant disease or disorder which, in the opinion of the Investigator, may either put the patient at risk because of participation in the study, or may influence the result of the study or the participant's ability to participate in the study.

2.3.4 *Randomisation*

Consenting patients were randomised once the inclusion and exclusion criteria had been met. For assignment of the patients to the two arms of the trial, a computer generated randomisation list (Appendix 1) with 1:1 allocation in permuted blocks of 10 was used.

2.3.5 *Blinding and concealment*

Participants and researchers were blinded to treatment allocation until all data had been entered into the database for both primary and secondary outcomes. The data were checked and verified and the database "locked down" prior to unblinding.

A Dunedin Hospital pharmacist, who was independent of the study, dispensed the study medication (cholecalciferol) or placebo in capsules. These were identical in appearance and were provided in sequentially numbered containers, according to the computer generated randomisation list.

2.3.6 *Baseline measurements*

2.3.6.1 *Demographic*

Demographic information including age, gender, ethnicity and season of diagnosis was collected and entered into the database.

2.3.6.2 *Clinical*

Clinical data were collected from patient clinical notes and entered prospectively into the database. This included date, height and weight (used to calculate body mass index= BMI), co-morbidities, the American Society of Anaesthesiology score, and site of tumour.

The American Society of Anaesthesiology (ASA) score is the standard method used to classify peri-operative risk and is determined by patient co-morbidities (Dripps *et al.* 1961).

Tumour site was classified as right colon, left colon or rectum. The right colon included the ileo-caecal region, caecum, ascending colon, hepatic flexure and transverse colon. The left colon included the splenic flexure, descending colon and sigmoid. The rectum included recto-sigmoid junction, rectum and anal canal.

2.3.6.3 *Laboratory*

Blood samples were collected at baseline and again immediately prior to surgery for serum calcium and calcidiol levels (See inclusion and exclusion criteria, sections 2.3.2 and 2.3.3).

Calcium was measured using a photometric assay (with NM-BAPTA) and analysed on the Cobas 702 (Roche Diagnostics, Mannheim, USA) at Southern Community Laboratories, Dunedin. Vitamin D was measured using high-performance liquid chromatography tandem mass spectrometry on the API4000 Mass Spectrometer (Thermo Fisher Scientific, NY, USA) at Canterbury Health Laboratories, Christchurch.

2.3.6.4 *Clinico-pathological*

The staging of tumours was performed using the American Joint Committee on Cancer (AJCC) Method, which utilises the TNM classification. T stage (1-4) represents depth of tumour invasion through the bowel wall, N stage (0-2) represents extent of regional lymph node involvement, and M stage (0-1) represents the presence or absence of distant metastasis. The TNM classification is then transposed to AJCC stages I-IV (Edge and Compton 2010).

Tumour grade, also known as tumour differentiation, was classified according to the World Health Organisation Classification of Tumours (WHO 2000). Tumours were graded G1-G3 by the pathologist, representing well, moderately and poorly differentiated tumours respectively.

2.4 Intervention

Patients were randomised to a single dose of 200,000 international units (IU), equivalent to 5 mg of vitamin D in the form of cholecalciferol, or an identical placebo. They were instructed to take the dose between seven and 21 days prior to the scheduled

surgery date. The study treatment was self-administered and the date of study drug ingestion was recorded.

2.4.1 *Dose justification*

The pharmaco-kinetics of a single dose of 100,000 IU of cholecalciferol was characterised in healthy volunteers by Ilahi et al (Ilahi *et al.* 2008). Following administration, serum calcidiol levels rose from a baseline mean of 67.6 nmol/L (+/- 19.2 nmol/L) to a mean C_{max} of 104.8 nmol/L (+/- 22.7 nmol/L). T_{max} was attained at day seven and decreased linearly so that at day 84 mean calcidiol measures were no longer significant compared to baseline (mean 80.1 nmol/L). The dose was found to be safe and no subjects experienced hypercalcaemia at any of the measured time points.

In a pilot feasibility study, conducted by Judd et al, the effect of calcitriol treatment on hypertension was investigated (Judd *et al.* 2010). Two hundred thousand IU of calcitriol was administered weekly for three weeks and no adverse effects were observed. Three other studies also reported no adverse effects after a single loading dose of 5 mg of vitamin D (Mallet *et al.* 2010, Schurch *et al.* 1998, Witham *et al.* 2010).

The dose of 200,000 IU was therefore considered to be safe and likely to be well tolerated. The dose was also designed to reliably achieve adequate serum levels (>80 nmol/L) without the need to delay surgery.

2.5 Samples and Follow-up

2.5.1 *Specimen collection and storage*

Resection specimens were collected fresh from the operating theatre and taken to the pathology laboratory. The attendant pathologist dissected the specimen and removed normal and tumour tissue samples for the study. Tumour tissue samples were taken from within the invasive margin, avoiding the necrotic regions of the tumour, with the aim of collecting non-necrotic representative tumour tissue. Normal tissue was taken from at least 5 cm from the tumour to reduce the likelihood of contamination with tumour cells.

Tissue was either stored directly at -80°C or immersed in RNAlater® overnight before storing at -80°C. RNAlater® is an RNA stabilisation solution designed to stabilise and protect RNA for prolonged periods prior to processing of tissue samples. Pre-treatment with RNAlater® allows storage of tissue at -80°C for indefinite periods, without deterioration in the quantity or quality of RNA (Thermo Fisher Scientific, NY, USA).

Normal and tumour tissue, along with heparinised blood (pre-incision), were also collected for the immune function studies (section 2.6.2.3).

2.5.2 *Clinical follow-up*

Patients were followed from admission to discharge for length of hospitalisation and post-operative complications. Following discharge patient notes were reviewed for cancer recurrence and survival. All surviving patients remain on follow-up, which is continuing at the time of writing.

2.5.3 *Data storage and de-identification*

Demographic, clinical, laboratory and clinico-pathological data were entered prospectively into a FileMaker® database (FileMaker Inc., Santa Clara, California). Follow-up data were collected and entered iteratively. All key fields were compulsory so that all required data were collected.

Data were de-identifed by removing the name and National Health Number (NHI) from the database. The linking information was kept in a separate, password-protected, Excel spreadsheet (Microsoft Excel, Redmond, WA, USA) that was only accessible to clinical members of the trial team.

Completion of data entry and checking and cleaning of data were performed prior to unblinding.

2.6 Outcome Measures

2.6.1 *Primary outcome – differential gene expression analysis*

The primary endpoint of the study was a significant change (minimum 1.5-fold) in vitamin D transcriptome activity in CRC tumour tissue and the corresponding normal colon in patients receiving the active study medication. RNA was extracted from both normal and tumour tissue and profiled on Affymetrix GeneChip® PrimeView™ Human Gene Expression Arrays (Affymetrix, Cleveland, USA). Microarray gene expression data were then analysed for differential gene expression utilising a linear models approach.

A full description of the differential gene expression analysis methodology has been presented in the Differential Gene Expression Analysis Methods chapter (Chapter 4).

2.6.2 *Secondary outcomes*

2.6.2.1 *Vitamin D levels*

Vitamin D levels were measured at baseline (prior to intervention) and again immediately prior to surgery. In order to mitigate the risk of unblinding, the pre-incision vitamin D levels were not read or entered into the database until other endpoints had been measured and entered.

2.6.2.2 *Adverse events related to intervention*

Adverse events were recorded and classified according to the Common Terminology Criteria for Adverse Events (CTCAE) (US Department of Health 2009).

2.6.2.3 *Immune function measurements*

This work is the subject of another post-graduate thesis, carried out by Mr Edward Taylor under the supervision of Dr Roslyn Kemp, Department of Microbiology and Immunology, University of Otago. It is therefore only briefly described here.

The initial aim was to evaluate anti-tumour macrophage activation in the CRC tissue, and the ability of tumour infiltrating macrophages to prime T-cells *in vitro*. Unfortunately the number of macrophages that could be isolated from specimens was too few to undertake these investigations, thus the focus of the immune function measurement was adjusted to examining infiltrating T-cell subsets.

Differences in infiltrating T-cell subsets in the normal and tumour tissue were identified and quantified using a novel analytical flow cytometric approach, developed by the Kemp Group (Girardin *et al.* 2013).

2.6.2.4 *Post-operative recovery*

Post-operative complication and length of hospitalisation data were collected and entered prospectively into the database. Admission and discharge dates were used to determine length of hospitalisation. Post-operative complications were graded according to the Clavien-Dindo Classification for Surgical Complications (Dindo *et al.* 2004).

2.6.2.5 *Recurrence and survival*

Cancer recurrence and survival were obtained by review of the clinical notes as well as direct contact with clinical teams and General Practitioner for additional information when needed. Death certificates were obtained for all deceased patients. The date, method of diagnosis and site(s) of recurrence were recorded. The cause and date of death were recorded for patients who are deceased, and current disease status and date of last follow-up recorded for patients who remain alive.

2.7 Statistical Analysis

2.7.1 *Analysis for clinical outcomes*

The following statistical methods were used for both baseline and secondary endpoint measurements.

Study groups were assessed for comparability at baseline by Fisher's exact test for categorical variables (Fisher 1955) and by either Welch's two sample t-test for normally distributed continuous variables (Welch 1947) or the Wilcoxon rank-sum test for non-normally distributed continuous variables (Wilcoxon 1945).

Pairwise comparisons were performed using paired t-tests (Box 1987).

Survival was visualised using the Kaplan-Meier method (Kaplan and Meier 1958) and the differences in recurrence free survival and disease-specific survival were assessed using the log-rank test (Mantel 1966).

All of the analysis for the clinical outcomes was performed in R (version 3.1.2), a free online software programming language and software environment for statistical computing and graphics (R-Core-Team 2014). Code and session information are available on GitHub (GitHub 2015) at https://github.com/FrancescaMunro/Thesis_Code (Munro 2015).

2.7.2 *Differential gene expression analysis*

Detailed differential gene expression analysis methods can be found in the Results of Differential Gene Expression Analysis chapter (Chapter 5).

2.7.3 *Sample size*

The primary analysis was based on intent-to-treat and the sample size was calculated for the primary outcome.

Microarray power calculations were performed by the study team's statistician using the methodology of Liu and Hwang (Liu and Hwang 2007). The following is a brief summary of that process.

To assess the probable impact of vitamin D on gene expression in a CRC cell line the Affymetrix data set (GEO: GSE444) of Wood *et al* was examined (Wood *et al.* 2004). In this data set significantly differentially expressed genes had undergone a 1.5-fold (on average) alteration in expression in response to vitamin D. Based on these findings data generated using Affymetrix HG-U133+2.0 microarray in the University of Otago Cancer Genetics Laboratory were utilised to obtain an estimate of variability in CRC gene expression. The method of Liu and Hwang was used to estimate the sample size required to detect 90% of the vitamin D transcriptome displaying an average 1.5-fold alteration in expression in response to vitamin D.

It was thus ascertained that a sample size of fifty patients was required, 25 in each of the treatment and control arms of the study, to achieve the level of power required for the primary outcome.

2.8 Summary

This trial was designed to investigate the impact of a single large dose of vitamin D on the normal colonic and colorectal tumour tissue of patients diagnosed with CRC. The primary outcome was measured using differential gene expression and bioinformatic methods. These are described in chapter 4. Secondary outcomes were also measured including laboratory, pathological and post-operative clinical outcomes. The methods for these were described in this chapter.

3. RESULTS OF CLINICAL OUTCOMES

3.1 Introduction

The main aim of this trial was to investigate whether a single large dose of vitamin D given to CRC patients prior to surgery would have an effect on vitamin D activated pathways in the normal and tumour tissue. In the previous chapter the trial design was described. Chapter 3 reports the clinical results and other secondary outcomes of the trial.

3.2 Patient Recruitment

Recruitment occurred over 17 months between April 2011 and November 2012. The CONSORT diagram (Moher *et al.* 2010, Schulz *et al.* 2010) summarises the flow of patients through the study (Figure 3.1).

Of the 117 patients assessed for eligibility 58 patients were excluded from the study and 59 patients were randomised. Of the 58 patients excluded, 36 did not meet the inclusion criteria (section 2.3.2) and 22 were excluded for other reasons. These include tumour considered too small for study sample (eight), inadequate time available between notification and surgery (six), study staff unavailable (six) and other logistical reasons (two).

The recruitment target was reached based on power calculation for sample size (section 2.7.3).

In the control arm both tumour and normal tissue samples were collected from 25 of the 30 patients. Of the other five patients, one did not proceed to surgery, the resected tissue of one patient was erroneously placed in formalin in theatre, and the pathologist was not able to provide a tumour sample in the remaining three.

Of the 29 patients recruited to the treatment arm, both tumour and normal tissue samples were collected from 26. The pathologist was unable to provide a tumour sample

for three patients. Normal tissue was collected for all but one patient, as staff did not request the normal sample (Figure 3.1).

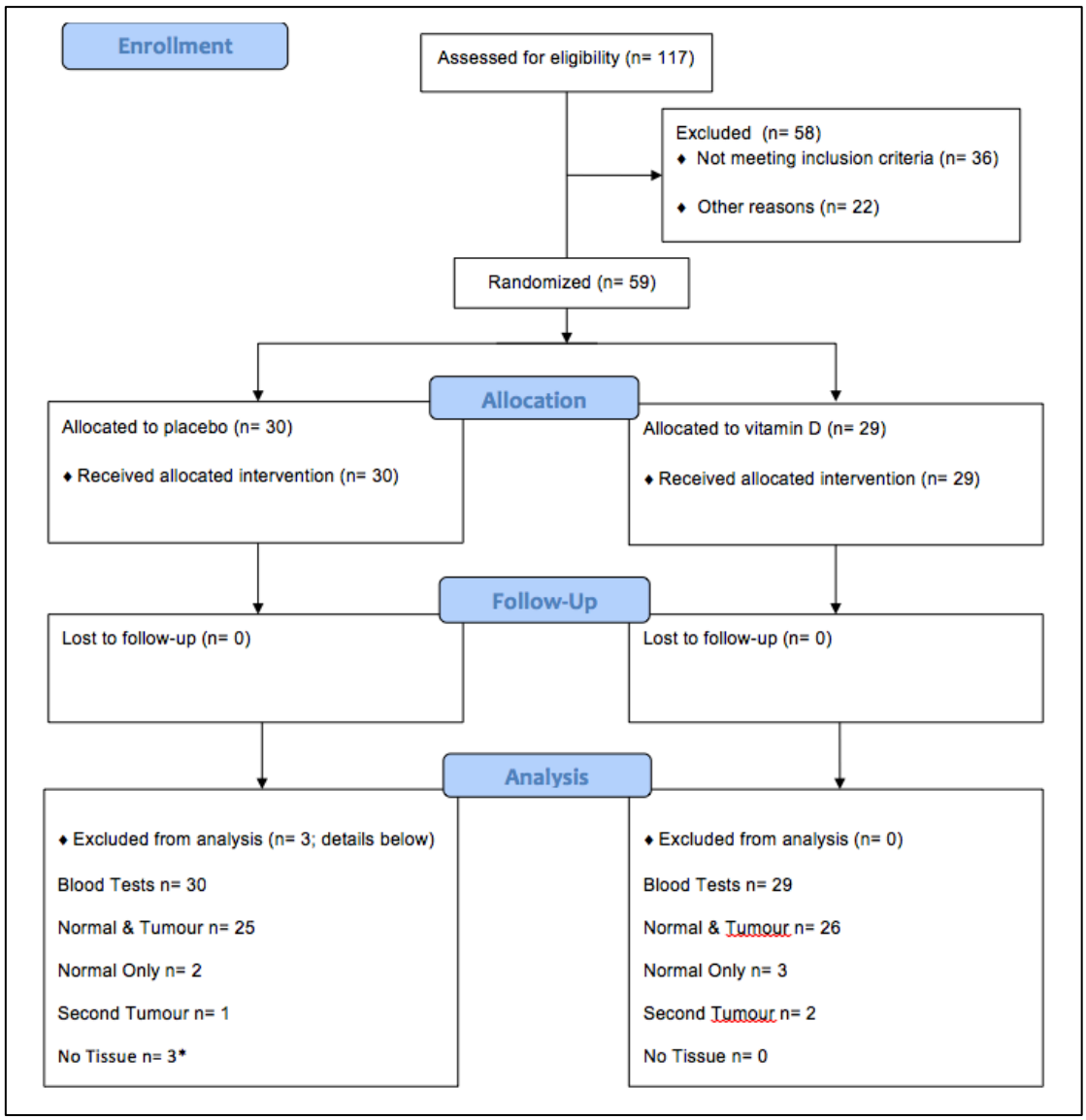


Figure 3.1. CONSORT diagram: displaying the flow of patients through the study. *1- no surgery, 1- tissue placed in formalin, 1- no tissue collected.

3.3 Baseline Patient Characteristics

3.3.1 *Demographic*

Baseline patient demographic characteristics of age, gender, ethnicity and season of diagnosis are summarised in Table 3.1. There were no significant disparities between the two groups.

The average age across the study cohort was 70 years, with a range of 38-86 years. Two participants were Māori. This is not unexpected considering the small sample size of the study and the relatively low proportion of Māori living within the Southern District Health Board region.

More patients were enrolled during autumn than other seasons, but there were no major differences in season of diagnosis between the treatment and control arms.

Characteristic	Placebo Group (n=30)	Vit D Group (n=29)	Total (n=59)	P Value
Age (years), median (range)	70(46-84)	71(38-86)	70(38-86)	0.98
Sex, n(%)				0.44
Men	14(47)	17(59)	31(53)	
Women	16(53)	12(41)	28(47)	
Ethnicity, n(%) (n=58)*				0.9
NZ European	24(83)	22(76)	46(79)	
NZ European/Māori	1(3)	0(0)	1(2)	
Māori	0(0)	1(3)	1(2)	
Dutch	1(3)	1(3)	2(3)	
Other European	1(3)	2(7)	3(5)	
Declined	2(7)	3(10)	5(9)	
Season of diagnosis, n(%)				0.7
Winter	5(17)	8(28)	13(22)	
Spring	8(27)	5(17)	13(22)	
Summer	6(20)	7(24)	13(22)	
Autumn	11(37)	9(31)	20(34)	

Table 3.1. Baseline patient demographic characteristics. Vit D= vitamin D. *Information not collected for the patient that did not go forward for surgery.

3.3.2 *Clinical*

Baseline patient clinical characteristics of BMI, ASA score and tumour site were well matched between the two arms of the study (Table 3.2).

The most common site for a tumour was the ascending colon (right colon), accounting for half of the cohort.

Characteristic	Placebo Group (n=30)	Vit D Group (n=29)	Total (n=59)	P Value
BMI (kg/m^2), mean (sd)	29.2(4.1)	28.0(4.7)	28.6(4.4)	0.31
ASA group, n(%)				0.38
I	6(20)	5(17)	11(18)	
II	21(70)	18(62)	39(66)	
III	2(7)	6(21)	9(15)	
IV	1(3)	0(0)	1(2)	
Tumour site, n(%)				0.32
Right colon	16(53)	12(41)	28(47)	
Left colon	7(23)	5(17)	12(20)	
Rectum	7(23)	12(41)	19(32)	

Table 3.2. Baseline clinical characteristics. Vit D= vitamin D. BMI= body mass index. ASA= American Society of Anaesthesiology.

3.3.3 *Laboratory*

Baseline serum calcidiol levels were measured for all of the 59 randomised participants, and were similar between the two study groups (Table 3.3). The baseline levels were higher in patients recruited to the study during summer and autumn.

One patient had a baseline calcidiol level of 104 nmol/L. This patient was taking cod liver oil, which is known to contain high levels of vitamin D. The patient is included in the intent-to-treat analysis.

All participants had serum calcium levels within the normal range.

Baseline calcidiol levels were significantly higher in patients who were diagnosed in summer and autumn (Table 3.3).

Characteristic	Placebo Group (n=30)	Vit D Group (n=29)	Total (n=59)	P Value
Baseline calcidiol (nmol/L), mean (sd)	58(27)	58(23)	58(25)	0.96
Baseline calcium (mmol/L), mean (sd)	2.26(0.13)	2.29(0.13)	2.27(0.13)	0.39
Baseline calcidiol (nmol/L)/Season of diagnosis*				<0.001[#]
Winter			45(23)	
Spring			45(26)	
Summer			73(14)	
Autumn			64(24)	

Table 3.3. Baseline serum levels. *mean(sd) #Welch's two sample t-test (Welch 1947).

3.3.4 *Clinico-pathological*

Baseline clinico-pathological characteristics of tumour stage and grade are shown in Table 3.4. These are similar in the treatment and control arms.

Characteristic	Placebo Group (n=30)	Vitamin D Group (n=29)	Total (n=59)	P Value
Tumour stage TNM, n(%) (n=59)				0.86
I	5(17)	7(24)	12(21)	
II	11(37)	11(38)	22(38)	
III	11(37)	9(31)	20(34)	
IV	3(10)	2(7)	5(8)	
Tumour grade, n(%)				1
G1	0(0)	0(0)	0(0)	
G2	24(83)	24(83)	48(81)	
G3	5(17)	5(17)	10(17)	
Missing*	1(3)	0(0)	1(2)	

Table 3.4. Baseline patient characteristics. *One patient did not go forward for surgery.

3.4 Outcome Measures

3.4.1 *Primary outcome – differential gene expression analysis*

A full description of the results of the differential gene expresson analysis is presented in the Results of Differential Gene Expression Analysis chapter (Chapter 5).

3.4.2 *Secondary outcomes*

3.4.2.1 *Vitamin D and calcium levels*

Post treatment calcidiol and calcium levels (levels determined from pre-incision bloods) were measured again immediately prior to surgery in all 58 patients who went forward for surgery (Table 3.5).

Pairwise comparisons were performed to examine serum calcidiol levels before and after study drug. Patients randomised to vitamin D had a significant increase in calcidiol levels above their baseline levels.

Post treatment calcium levels remained below 2.6 mmol/L in both groups.

	Placebo Group (n=29)	**Vitamin D Group** (n=29)	**Total** (n=58)*	**P Value**
Postdose calcidiol (nmol/L), mean (sd)	49(19)	87(22)	69(28)	<0.001[#]
Pre & post calcidiol comparison	0.2	<0.001[$]	0.036	
Postdose calcium (mmol/L), mean (sd)	2.26(0.09)	2.25(0.12)	2.25(0.10)	0.61
Pre & post calcium comparison	0.97	0.19	0.35	

Table 3.5. Secondary outcomes: calcidiol levels. *One patient did not go forward for surgery. #Wilcoxon rank-sum test (Wilcoxon 1945). $Paired t-test.

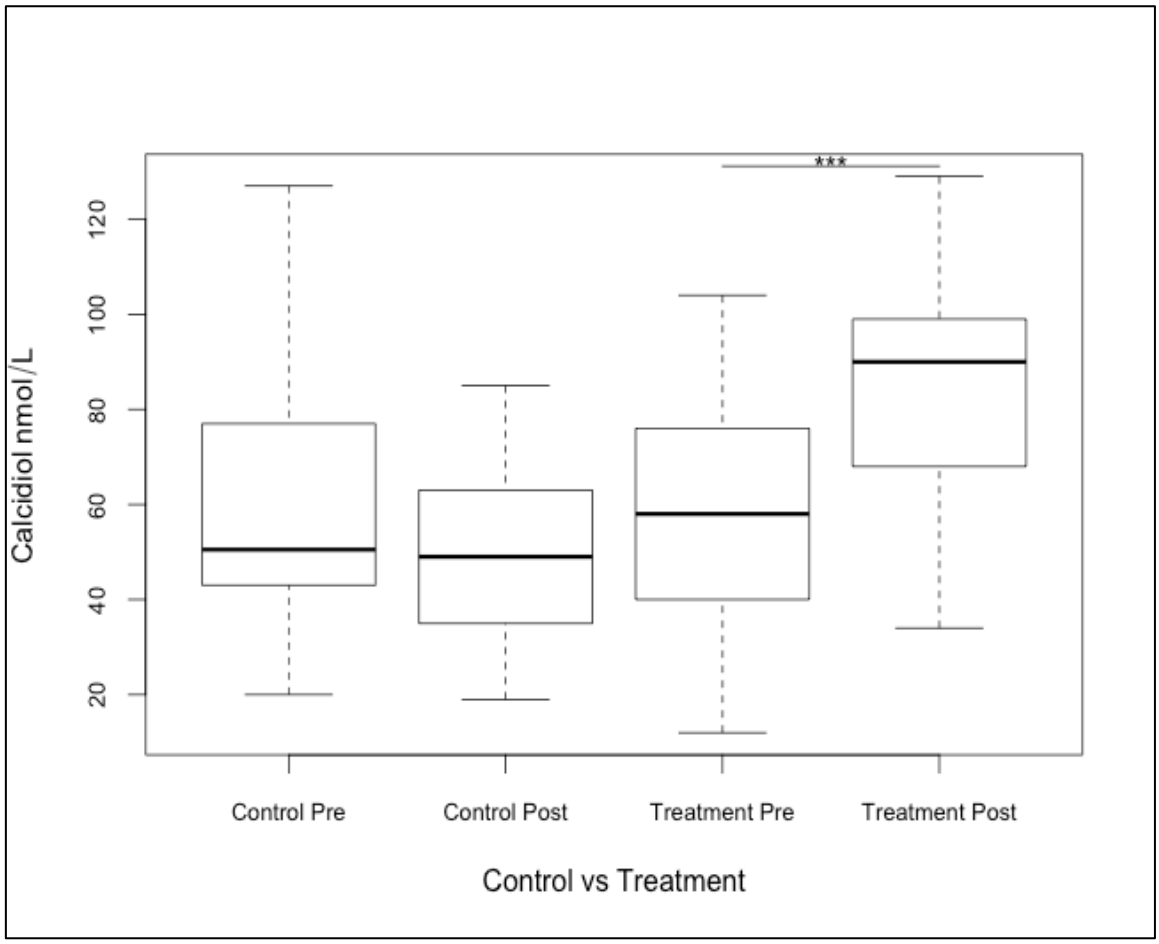


Figure 3.2. Boxplot of calcidiol levels at baseline and post treatment in the study groups. Pre= pre-study treatment (baseline). Post= post-study treatment. ***p<0.001 (paired t-test).

3.4.2.2 *Adverse events related to intervention*

No adverse events were reported related to the study drug in either of the control or treatment arms.

3.4.2.3 *Post-operative recovery*

The median length of hospitalisation was 7.5 days (Table 3.6).

The Clavien-Dindo classification is a well validated method of describing post-operative complications. The numbers of complications were similar between the two groups with a total of 23 and 21 for the control and treatment groups respectively, and no difference in Clavien-Dindo grade.

	Placebo Group (n=29)	Vitamin D Group (n=29)	Total (n=58)*	P Value
Day stay (days), median (range)	7(4-45)	8(4-37)	7.5(4-45)	0.83*
Post-op complications, n(%)	23(79)	21(72)	43(74)	0.78
Number of post-op complications, n(%)				0.94
0	8(28)	10(35)	18(31)	
1	9(31)	9(31)	18(31)	
2	4(14)	3(10)	7(12)	
3	5(17)	3(10)	8(14)	
4	1(3)	2(7)	3(5)	
5	0(0)	1(3)	1(2)	
6	2(7)	1(3)	3(5)	
Clavien-Dindo classified complications, n(%)				0.24
I	22(46)	11(26)	33(56)	
II	23(48)	27(64)	50(85)	
IIIa	2(4)	2(5)	4(7)	
IIIb	1(2)	2(5)	3(5)	
IVa	0(0)	0(0)	0(0)	
IVb	0(0)	0(0)	0(0)	
V	0(0)	0(0)	0(0)	

Table 3.6. Secondary outcomes: post-operative recovery. *Wilcoxon rank-sum test.

3.4.2.4 *Recurrence and survival*

At the time of reporting, the median follow-up time is 29 months with a range of 1-44 months. To date 11 of the 59 study patients have had tumour recurrence. Recurrence and survival data are shown in Figures 3.3 and 3.4. There are currently no significant differences between the groups.

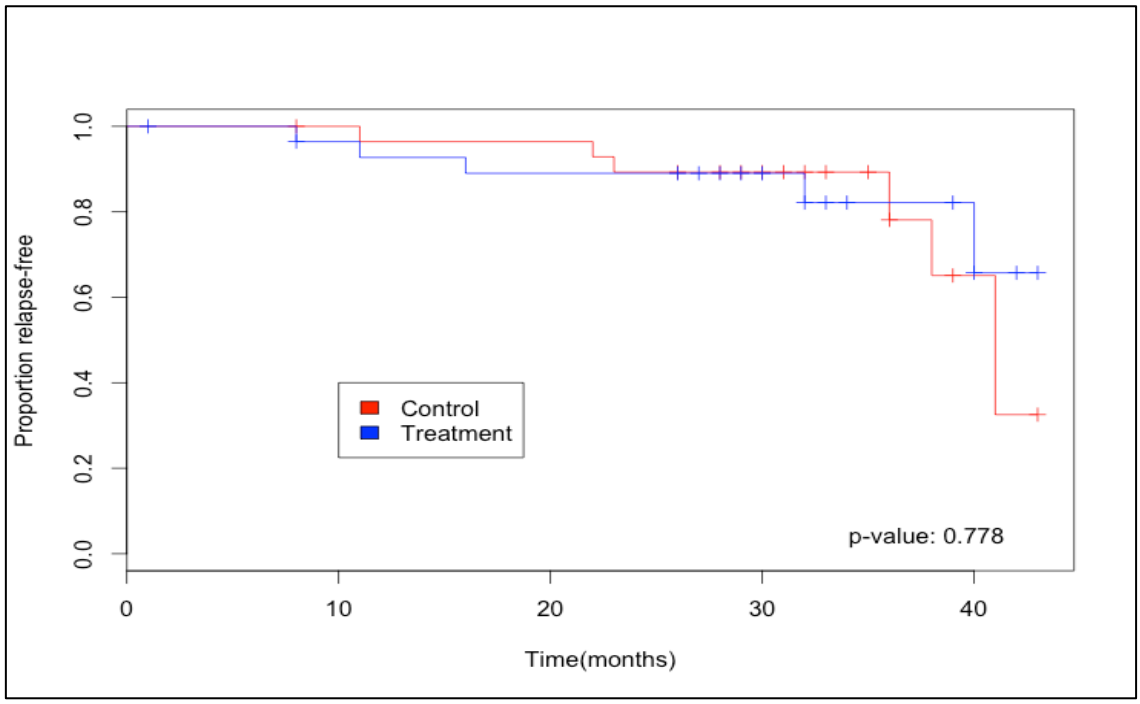


Figure 3.3. Kaplan-Meier curve for recurrence-free survival in the treatment and placebo groups. There are currently no significant differences in recurrence-free survival between the groups.

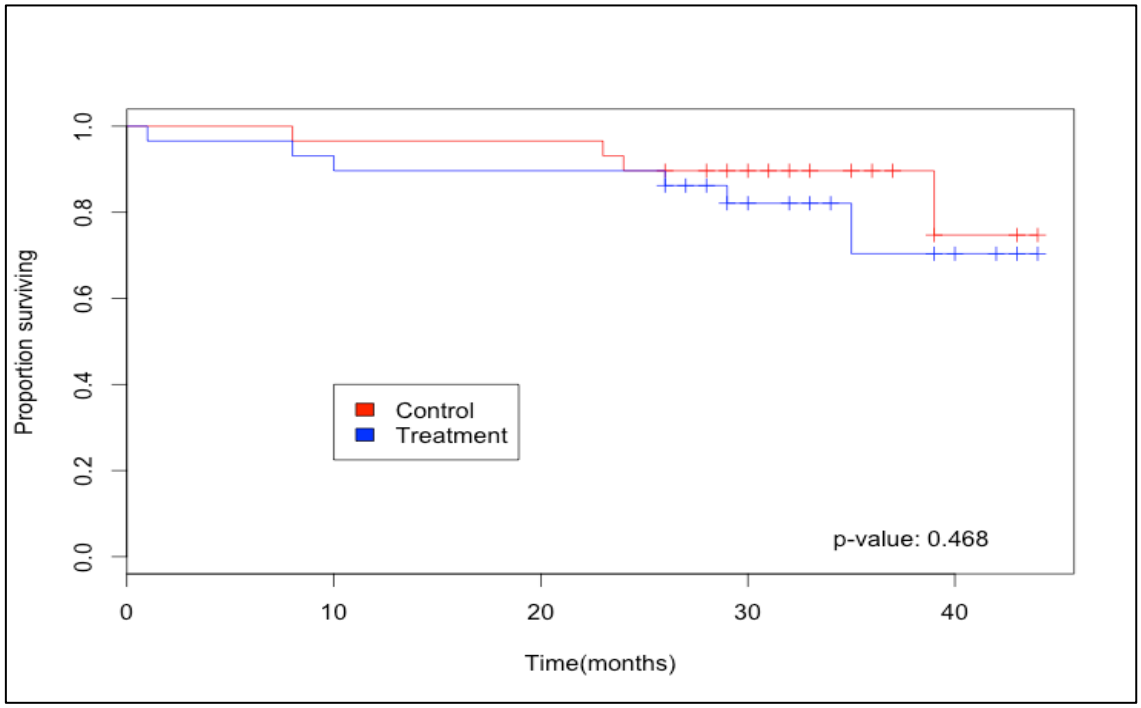


Figure 3.4. Kaplan-Meier curve for overall survival in the treatment and placebo groups. There are currently no significant differences in overall survival between the groups.

3.5 Discussion

The aim of the study was to determine whether a single large dose of vitamin D, administered at least one week prior to surgery in patients with CRC, would have a measurable effect on vitamin D activated pathways in the resected tumour tissue and non-tumour bowel. The recruitment target was 25 patients in each of the treatment and control arms of the study. This was exceeded to ensure that at least 25 patients in each group had tumour and non-tumour bowel specimens available in accordance with the *a priori* power calculation. No adverse events were reported related to administration of the study drug. Patients in the treatment arm achieved significantly higher levels of vitamin D by the time of surgery, which is a necessary pre-condition for the primary aim of the study.

The complication rate and length of day stay were within the expected range for the cohort of patients (DeBarros and Steele 2013, University of Auckland). No differences in post-operative recovery, cancer recurrence or survival were expected in this study, which was not powered for clinical endpoints, and none were observed. Nevertheless patients will be followed up for a minimum of five years.

Holick (2005a) and other researchers have proposed that the optimal serum calcidiol concentration should be at least 80 nmol/L (section 1.2.5). The mean post-treatment calcidiol level in the treatment group was 87 nmol/L compared to the placebo group of 49 nmol/L. Although the post-treatment mean calcidiol levels for both groups may be within the 'normal' range (section 1.2.5), the treatment group underwent a 60% increase in mean calcidiol serum level. This 60% increase may lead to observable changes in the expression of vitamin D responsive genes in the tumour and corresponding normal tissue (the primary outcome of the study).

Serum calcidiol concentrations were found to be significantly higher in patients recruited during summer and autumn. Vitamin D synthesis in the skin varies depending on the season. This phenomenon has been described (Holick 1995) and is summarised in section 1.2.6. Research carried out in New Zealanders by Rockell et al. (2006) found that mean calcidiol concentrations varied markedly with season with the greatest difference observed between summer and spring levels. Similar findings were reported

by other New Zealand researchers, Livesey et al. (2007) and Bolland et al. (2006) (section 1.2.8).

In summary, this chapter has described the study participants and outcomes for the laboratory, pathological and post-operative clinical outcomes of the study. The recruitment target was met and the pre-operative vitamin D levels in the treatment arm were significantly increased. No differences in clinical outcome were observed. The remainder of the thesis will concentrate on the differential gene expression analysis of tumour and non-tumour bowel specimens obtained at surgery.

4. METHODS FOR DIFFERENTIAL GENE EXPRESSION ANALYSIS

4.1 Introduction

The primary objective of the study was to investigate whether a single large dose of vitamin D given to CRC patients prior to surgery could alter the activity of the vitamin D transcriptome (vitamin D responsive genes) in tumour tissue and corresponding normal bowel. Vitamin D has been shown to impact gene expression in tumour tissue both *in vitro* and *in vivo* (sections 1.3.3, 1.3.4.2 and 1.3.4.2a-e). Cellular activities impacted by vitamin D in tumour tissue have comprised a range of effects including cell proliferation, cell differentiation, apoptosis, detoxification and angiogenesis. Differential gene expression and bioinformatic methods were employed to measure alterations in gene expression in the normal and tumour tissue collected from participants in the study.

Chapter 4 describes the methods used to measure the primary outcome. The results of the differential gene expression analysis are reported in the following chapter (Chapter 5).

4.2 Methods for Differential Gene Expression Analysis

4.2.1 *Microarray methods*

RNA extraction was performed using the Norgen Total RNA Purification Kits (Norgen Biotek, Thorold, Canada). RNA purity was assessed using the Implen NanoPhotometer® (Implen GmbH, Munchen, Germany). RNA quality was assessed utilising the Agilent 2100 Bioanalyser (Agilent Technologies, Palo Alto, California, USA). All work was carried out in accordance with the manufacturers' instructions.

RNA labeling was performed using the NuGEN Encore® Biotin Module labeling kit (NuGEN Technologies, San Carlos, USA), according to the manufacturers' directions.

The RNA was hybridised to and profiled on Affymetrix GeneChip® PrimeView™ Human Gene Expression Arrays (Affymetrix, Cleveland, USA), following the manufacturers' protocols.

The Affymetrix GeneChip® PrimeView™ Human Gene Expression Arrays were scanned using the Affymetrix GeneChip® Scanner 3000 7G Plus.

The Affymetrix GeneChip® PrimeView™ Human Gene Expression Arrays provide coverage of the whole genome with a total of 49,395 probe sets. They contain eleven probes per probe-set for well-annotated sequences and nine probes per set for non-well-annotated sequences (Affymetrix 2015).

4.2.2 *Microarray quality control*

Quality assessment of the microarray data was performed using the arrayQualityMetrics package (Kauffmann *et al.* 2009), available from the Bioconductor project (Gentleman *et al.* 2004). The package generates a microarray quality metrics report comprising a series of visualisations explained by text. Report generation is expedient and allows for prompt assessment of data quality. The package can be run using either raw or normalised data (section 4.2.3).

The quality metrics report deals with a number of array parameters separated into sections. The first section is committed to between array comparisons where distances between arrays are presented in a false colour heatmap and principal component analysis (PCA) (Pearson 1901) is visualised in a scatterplot. The distance between two arrays is calculated as the mean absolute difference between the data of the arrays. A bar chart of the sum of distances compared to other arrays is also provided. A threshold is established based on the distribution of the values across all arrays and represented as a vertical line; outliers are found outside (to the right) of this line.

The PCA scatterplot represents the distribution of the arrays along the first two principal components. PCA is a dimension reduction and visualisation method. In this context, PCA is employed to translate the multivariate data vector of each array onto a two-dimensional plot. Overall array similarity (or dissimilarity) is displayed in the spatial arrangement of the points in the scatterplot. The PCA plot is an interactive plot where placing the cursor on a point on the plot will render its sample name. Outliers are portrayed by a larger dot.

The second section reports array intensity distributions utilising boxplots and density plots. Summaries of the signal intensity distributions for each array are illustrated as boxplots. Generally, it is expected that the boxes will be in similar positions on the plot and have similar widths. Detection of outliers is achieved by calculating the Kolmogorov-Smirnov (KS) statistic between each array's distribution and the overall distribution of the combined data (Chakravarti and Laha 1967). A bar chart of the KS statistics is provided where a threshold is determined based on the distribution of the values across all arrays and represented as a vertical line. Outliers are found outside of the threshold line.

The density histograms depict density estimates (smoothed histograms) of the array intensity data. The plots are superimposed on a single graph enabling the observation of unusual distributions more readily. Generally, the histograms are expected to have similar shapes and ranges. The graph is interactive; placing the cursor on a curve on the plot will render its sample name. Background noise here refers to technical noise and includes differences in sample preparation, differences in labelling reaction efficiency, production differences between the microarrays and differences in experimental factors, for example, non-specific cross hybridisation. High levels of background noise shift an array's histogram to the right. The tail of a histogram can be shortened if the array signal intensity is insufficient.

Section three concerns variance-mean dependence and is visualised in a density plot. The density plot represents the standard deviation of the signal intensities for the arrays versus the rank of their mean. The running median of the standard deviation is shown using red dots connected by lines. After the intensity data has been normalised and transformed to a logarithm-like scale, it is generally expected that the red line will be approximately horizontal, which would signify no significant trend.

There is an Affymetrix-array specific section generated if these arrays are used and assesses array quality. This section was only produced in the quality control report generated before normalisation of the expression data. Boxplots of Relative Log Expression (RLE) and Normalised Unscaled Standard Error (NUSE) with associated outlier barcharts are rendered. An RNA digestion plot indicating RNA degradation is also produced in this section. Both the RLE and NUSE are derived from a probe-level model

that calculates an expression measure utilising M-estimator robust regression (Hogg 1979). RLE plots are generated using log-scale estimates for the expression of each probe-set on each array. For each probe-set on each array, ratios are computed between the expression of a probe-set and the median expression of the probe-set across all of the arrays. Relative log expression boxplots are expected to centre around zero. If they do not or are more spread than the majority of the arrays, they are considered to potentially problematic in terms of quality.

Detection of outliers is achieved by calculating the KS statistic between each array's RLE values and the pooled, overall distribution of the RLE values (Chakravarti and Laha 1967). A bar chart of the KS statistics is provided where a threshold is determined based on the distribution of the values across all arrays and represented as a vertical line. Outliers are found outside of the threshold line. Standard error (SE) estimates are normalised so that for each probe-set the median SE across all the arrays is equal to one. Normalised Unscaled Standard Error boxplots are expected to be centred around one. An array with elevated values relative to the other arrays is typically of lower quality. Detection of outliers is achieved by calculating the 75% quantile, N_a, of each array's NUSE values. Outliers have a large N_a. A bar chart of the N_a values is provided where a threshold is determined based on the distribution of the values across all arrays and represented as a vertical line. Outliers are found outside of the threshold line.

The RNA digestion plot provides a general assessment of RNA degradation. RNA degradation occurs from the 5' end of a transcript. Heavy degradation would theoretically result in a systematic shift towards lower signal values for the probes closer to the 5' end. For every array the probes are ordered from the 5' end of the target transcripts. Each line in the plot illustrates the average degradation effects calculated from all the probe-sets for an array. The higher the slope, the stronger the degradation effect. Arrays with slopes deviating from the majority probably contain problematic data.

The final section attends to individual array quality illustrated in MA plots. The MA plot is a visualisation of the distribution of the difference between log intensities (M) for an array versus the average log intensities (A) for the array. The array of interest is compared to a 'median array' consisting of the median log intensities across the set of

experiment arrays. Generally, it is expected that the majority of the distribution would be focussed along the M=0 axis, signifying no trend. Two intensities are measured from the microarray 'spot' (where RNA hybridises to probes), the spot and background intensity. The background intensity is taken from the periphery of the spot and is used to remove 'noise' from the signal intensity. If a trend exists in the lower range of A, differing background intensities between the array of interest and the 'median array' are frequently the issue and may be resolved using background correction. Background correction can be addressed when normalising the expression data and is described in the following section (section 4.2.3). If a trend exists in the upper range of A, saturation of signal intensities may be indicated and in subtle cases may be amended using quantile (non-linear) normalisation.

Detection of outliers is achieved by calculating Hoeffding's D statistic on the paired distribution of M and A for each array (Hoeffding 1948). Hoeffding's D statistic is a non-parametric test of independence, which depends on the rank order of the observations, in this case the rank order of M for an array versus the rank order of A for the same array. (The non-parametric model is not completely without parameters but the number and nature of the parameters are not determined in advance and are thus accommodating of the data). The range for D in this context is -0.5 to 1.0. The higher the D statistic, the more dependent the variables being tested. MA plots for arrays with the highest and lowest Hoeffding's D statistics are presented, along with a bar chart of the Hoeffding's D statistics. A threshold of 0.15 is used in this context. Outliers are found outside of the threshold line.

In the quality control report generated before normalisation of the data, a false colour representation plot of the arrays' spatial distributions of feature intensities was produced along with a bar chart identifying the outliers. The false colour plots assist in identifying patterns that may be generated by various issues including air bubbles and spatial gradients in the hybridisation chamber. The plot appears uniform if the features are randomly distributed. The colour scale is proportional to the ranks of the probe intensities. The detection of outliers is achieved by computing, F_a, the sum of the absolute value of low frequency Fourier coefficients. Fourier coefficients are estimated intensity values for presenting features. A bar chart of the F_a values is provided where a

threshold is determined based on the distribution of the values across all arrays and represented as a vertical line. Outliers are found outside of the threshold line.

Quality assessment reports were generated both before and after data normalisation and are discussed in chapter 5.

4.2.3 *Data normalisation*

The objective of data normalisation (pre-processing) is to remove experimental artefacts without removing the biological signal. There are many algorithms that can be used to normalise microarray data. Benchmarking of the various methods is an ongoing exercise. At present, the Robust Multi-chip Average (RMA)-based normalisation procedures (Bolstad *et al.* 2003, Irizarry *et al.* 2003a) are considered to be amongst the leading methods (Irizarry *et al.* 2003b, Welsh *et al.* 2013). Welsh et al. (2013) searched the Gene Expression Omnibus, ArrayExpress and microarray repositories to find the three most commonly used software packages for processing Affymetrix microarrays. The RMA method was the most used approach of the three (the others being MAS5 and dChip) that featured in this search, due to accuracy in processing signal intensities and the low level of noise generated by the procedure. Another significant strength of the procedure is its use of the complete set of arrays for both estimation of signal and normalisation (refer below for further information).

In this study, data normalisation was performed using RMA. This approach consists of three components including background correction/adjustment, quantile normalisation and per-gene summaries.

Probe intensities are assumed to be a combination of signal and background noise. The RMA background adjustment estimates the signal (and attempts to remove the noise), assuming that the signal is exponentially distributed and the background noise is normally distributed. In earlier Affymetrix arrays (compared with the arrays used in the present study), mismatch probes (MM) were incorporated in the array design to estimate background noise. Each perfect match probe (PM) had an MM, which had the same sequence as the PM bar a single nucleotide. Subtracting the MM signal from the PM signal provided an estimate of the expression signal. Affymetrix GeneChip® PrimeView™ Human Gene Expression Arrays (used in this study) incorporate background probe

collections (BGP) for estimation of background noise. They provide an estimate of the background noise based on coarsely modelled sequence that is considered to be close enough to the background signal provided by MM. The advantage of using BGP is that they require approximately 50% less space on the array, allowing an approximately 2-fold increase in PM content (Affymetrix 2005, Affymetrix 2007). Benchmarking suggests that RMA performed without the background adjustment component is advantageous. This is because while background adjustment minimises bias in point estimates, it also greatly elevates the variance of expression measures (Bolstad 2002, Freudenberg 2005). In the present study, normalisation was performed both with and without background correction.

The aim of quantile normalisation is to ensure that the distribution of the probe intensities is the same across the arrays (removal of array effects) thus enabling comparisons to be made between arrays. The only assumption made by the quantile normalisation method is that the signal intensities of each array originate from the same underlying distribution. After quantile normalisation, all gene-chips have the same density function (Bolstad *et al.* 2003, Freudenberg 2005).

A per-gene summary is the log intensity for a probe-set (gene), representing the gene expression level for that gene. This summarisation process reduces the effect of outliers within each probe-set and provides a similar result to taking the median of each normalised probe-set. The median polish algorithm is used to accomplish this by fitting a model to the data that takes into account the differences in probe affinities. The method provides robust estimates in two ways. Firstly, using medians in place of means renders the algorithm less sensitive to outliers and secondly, estimations are based on all of the gene-chips for the investigation allowing information to be acquired from other arrays (Freudenberg 2005).

4.2.4 *Genes with vitamin D responsive elements*

Genes responsive to vitamin D have enhancers termed vitamin D responsive elements (VDRE) (section 1.2.2). A VDRE gene list was created to investigate their expression as a separate group. The list was constructed upon consultation of the literature. Genes were added to the list when they featured as calcitriol targets or as explicitly containing at

least one VDRE. Extensive lists were obtained from work carried out by Wang et al and Meyer et al (Meyer *et al.* 2012, Wang *et al.* 2005).

The Wang et al. (2005) study involved a genome wide investigation to identify direct calcitriol targets in a squamous cell carcinoma cell line (SCC25). Large scale *in silico* and microarray based techniques were utilised. These techniques were complementary, with microarrays used to identify regulated genes, and *in silico* screens used to identify potential target genes independent of the tissue of expression. Microarray analyses identified more than 900 regulated genes, but as these were expressed in SCC25 they were not included in the VDRE gene list (Vitamin D activity is cell-type and cell-context specific; section 1.2.2). Screening for consensus DR3 and ER6 VDRE elements (section 1.2.2) in the human genome, lying within -10 kb to +5 kb regions of genes, presented over 1300 genes. These genes were included in the VDRE gene list (Wang *et al.* 2005).

Meyer et al. (2012) carried out investigations examining the VDR/RXR (section 1.2.2) cistrome in CRC tumour cells. Chromatin immunoprecipitation-seq and gene expression analyses were performed to identify the VDR/RXR cistrome and the genes it moderates. Genes found to be regulated in this study were added to the VDRE gene list. VDRE de novo motif discovery analyses were also undertaken. An extensive list of genes from this work was also added to the gene list (Meyer *et al.* 2012).

GeneCards is an online human gene compendium, which provides gene name aliases and related gene data as well as links to various databases for further information (Rebhan *et al.* 1997). GeneCards was used to obtain Entrez Gene official gene symbols for the VDRE list genes where they were not already provided. The VDRE gene-set comprises over 3130 genes. The VDRE list genes were matched to the appropriate probes from the primeview symbols list using the *AnnotationDbi::as.list* function in R (R-Core-Team 2014). Three separate lists contributing to the VDRE gene list, with their respective citations can be viewed in Appendix 2. The lists include a 'miscellaneous' list with VDRE gene contributions gathered from many papers, a Meyer list and a Wang list.

4.2.5 *Differential expression analysis (genes)*

4.2.5.1 *Limma*

Differential expression (DE) analysis of the microarray data was performed using the

limma package (Ritchie *et al.* 2015) available from the Bioconductor project (Gentleman *et al.* 2004). The limma package is widely used (Yaari *et al.* 2013) and as the candidate had some previous experience with it it was selected for use in this analysis. The limma package utilises a linear models approach to identify individual genes that have been significantly differentially expressed. Once the expression data has been normalised, a linear model is fitted to the gene expression signals to detect the genes that have undergone significant changes in expression. For statistical analysis and evaluating DE, limma employs an empirical Bayes method to generate a modified test statistic, based on shrunken estimates of the per-gene variances. Creating a modified test statistic has a two-fold effect. Firstly, the impact of genes with very small standard deviations is controlled, limiting any major influence on the overall gene expression results. Secondly, significant gene expression changes are highlighted, by making it more probable that they will have a large-fold change (Smyth 2004).

The expression data was presented to the linear models and Bayesian shrinkage estimation functions in limma as a matrix. The data matrix is submitted in conjunction with a design matrix, which specifies experimental grouping, allowing the comparison of gene expression between experimental conditions.

A summary of the results of the linear model can be generated as a table. The table presents the most highly differentially expressed genes (DEG), ordered by p-value. The p-value is a function of the observed sample results and is used in testing statistical hypotheses. The null hypothesis is rejected if the p-value is equal to or less than the significance threshold. The table also incorporates other summary statistical information for each gene including the log-fold change (log base 2 in this case), average expression (log2), t-statistic, an adjusted p-value and B-statistic. The log-fold change represents, for a given gene, the magnitude of the difference in the log-expression between experimental conditions. The average expression denotes the average log-expression level, for a given gene, across all arrays. A negative t-statistic signifies a down-regulated gene and conversely a positive t-statistic represents an up-regulated gene. The adjusted p-value is the p-value adjusted for multiple testing (section 4.2.5.2). A significance threshold is selected before the testing is performed, usually 5% or 1%. The B-statistic is the log-odds that the gene is differentially expressed. However the B-statistic does not take multiple hypothesis testing into account (section 4.2.5.2). A

threshold proportion of DEG is required by limma for the B-statistic to be calculated. A B-statistic of zero corresponds to a 50% chance that the gene is differentially expressed, whereas a B-statistic of greater than zero signifies that the gene is more likely to be differentially expressed.

4.2.5.2 *Multiple testing correction*

During microarray data analysis multiple hypothesis testing is performed to assess DE for each gene. Performing multiple tests has the potential to substantially inflate the Type 1 error rate (i.e. false positive results, where the null hypothesis is wrongly rejected) to levels that are unacceptable depending on the significance threshold. A significance threshold of $\alpha=0.05$ would render (on average) 1000 Type 1 errors from 20,000 tests in which the null hypothesis was always true (the order of testing in the current study). Type 1 errors can be controlled using various approaches, the most common of which fall into two categories, family-wise error rate control (Hochberg and Tamhane 1987) and False Discovery Rate (FDR) control (Benjamini and Hochberg 1995). Family-wise error rate controlling methods are focused on ensuring that the likelihood of a single testing error across all tests performed is small, and thus imparts a high degree of certainty regarding the results. However, this level of control is very conservative, making it likely that some genes that undergo DE will not be identified. FDR controlling procedures allow a small proportion of Type 1 errors in the overall number of rejected hypotheses. It is a less conservative approach and is consequently endowed with an increased probability of Type 1 errors. However, FDR methods are less likely to generate Type 2 errors (false negatives) and therefore achieve greater power.

The Benjamini and Hochberg (BH) method, an approach to control the FDR was selected to control the Type 1 error rate in the present study (Benjamini and Hochberg 1995). The BH procedure controls the FDR at the pre-selected significance threshold. In this method the p-values are sorted and ranked. The smallest p-value is assigned a ranking of one, the second a ranking of two et cetera. The largest value is assigned a ranking of N. Each p-value is multiplied by N and divided by its allocated rank to produce adjusted p-values. Monotonicity of the p-value sequence (non-decreasing values from smallest to largest) is preserved by setting any adjusted p-values to the value of the next lowest ranked p-value in cases where the adjustment results in a decrease). To restrict the FDR to the significance threshold, for example $\alpha=0.05$, genes with an adjusted p-value less

than 0.05 are selected.

4.2.5.3 *Normal, tumour and paired analyses*

The primary endpoint for this study was to detect a 1.5-fold change in the expression of vitamin D responsive genes in the tumour tissue. This level of change is considered to represent a biologically meaningful change.

Analysis was performed comparing gene expression in the normal tissue between the study groups, for all genes. Expression data for the normal tissue, from both the control and treatment groups, was collated in a matrix in R (R-Core-Team 2014). A design matrix was generated specifying which of the normal tissue samples were in either of the control or treatment groups. These matrices were presented to the linear model and Bayesian shrinkage estimation functions in limma (section 4.2.5.1).

The raw p-values for the VDRE genes were extracted from analysis conducted in the normal tissue and adjusted to correct for the number of VDRE genes.

Analysis was performed comparing gene expression in the tumour tissue between the study groups, for all genes, following the process described for analysis of gene expression in the normal tissue between the trial groups. The raw p-values for the VDRE genes were extracted from analysis conducted in the tumour tissue and adjusted to correct for the number of VDRE genes.

Gene expression analyses were performed in paired normal and tumour tissue comparing gene expression between the study groups, for all genes. Normal and tumour tissue pairing has the advantage of removing differences that arise due to variation between patients. Once the normal and tumour tissue gene expression data were paired, the tumour gene expression values were subtracted from the normal gene expression data and a matrix, termed a *difference matrix*, was produced. The difference matrix along with a design matrix specifying the treatment group allocation for each set of paired normal and tumour tissue, were presented to the linear model and Bayesian shrinkage estimation functions in limma. The raw p-values for the VDRE genes were extracted from analysis conducted in the paired analysis and adjusted to correct for the number of VDRE genes.

4.2.6 *Differential expression analysis (gene-sets)*

4.2.6.1 *Gene-sets and pathways*

Relevant alterations in gene expression can be subtle at the individual gene level. Also examining genes in isolation disregards the complex interactions within cells and tissue. For these reasons it can be more helpful to investigate gene pathways rather than individual genes (Mootha *et al.* 2003, Nam and Kim 2008, Subramanian *et al.* 2005).

A gene-set can be defined as the genes of a biological pathway but can equally constitute any collection of genes that a researcher may want to investigate the expression of. Biological pathways include innumerable cell functions ranging in activity from transcriptional regulation through signal transduction and metabolic activity to structural and cellular activity. Having significant pathways identified by gene set analysis (GSA) provides information that can potentially elucidate biological activity in the tissue of interest.

There are a large number of online databases available for assessing lists of genes. In the present study GeneSetDB (Araki *et al.* 2012), Gene Ontology (GO) (Ashburner *et al.* 2000), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto 2000), Reactome (Croft *et al.* 2014, Milacic *et al.* 2012) and Wikipathways (Kelder *et al.* 2012) were employed. Investigators can submit a list of gene-sets to a database whereupon pathways are returned. Depending on the database's capabilities, users can interact with the database to access related biological information.

GeneSetDB integrates gene-set information from 26 public databases, producing a comprehensive meta-database, with a specific emphasis on human disease and pharmacology. The database was generated to allow the investigator to access many databases, or a subset thereof, and their information from a single site.
Reactome is a database of reactions, pathways and biological processes. It has a focus on visualisation of biological pathways and the user has the ability to interact with the source data, which is downloadable in various formats. Each pathway component is supported by citations providing verification of the activity represented. Processes are inferred from non-human data when no human data exists if an expert biologist and reviewer deem it appropriate.

Wikipathways is designed for community curation of biological pathways information. Its focus is on gene, protein and metabolite systems. In addition to a pathway diagram, a bibliography and list of constituent genes and proteins is provided with links to public resources. Users can access lists of pathways that share the same pathway node. The database has over 1950 pathways available to interact with.

The KEGG database is an amalgamation of a number of databases which are categorised into systems, genomic, chemical and health information. The core of the KEGG database is the KEGG pathway database integrating genes, RNAs, proteins, chemical compounds and chemical reactions and also disease genes and drug targets. The database has approximately 220 pathways available.

The GO project is a bioinformatics initiative with the three main objectives of continuing to develop gene ontology, assigning ontology to gene/gene products and developing software and databases for the first two objects. GO terms are divided into three domains, molecular function, cellular component and biological process. The molecular function domain deals with gene product activities at the molecular level. The cellular component section is involved with the cell componentry and its extracellular environment. The biological process area concerns sets of molecular events engaged in the functioning of integrated systems, cells, tissues, organs and organisms. Each domain contains functional categories, each consisting of a list of GO terms representing increasingly specific biological activity. The GO terms have defined relationships to other terms both within and between domains. Individually, a GO term has an identifier, name, assigned domain, definition with citations and other related information. The GO project has approximately 40,000 terms available.

4.2.6.2 *Over-representation analysis and gene-set analysis*

A list of the most highly differentially expressed genes from a differential expression analysis and a collection of gene-sets are used in over-representation analysis (also known as enrichment analysis). Fisher's Exact Test is used to test whether there are more genes in the presented list, that are present in a given gene-set than would be expected by chance. The method is limited by only taking into account the size of a gene-set and not the ranks (e.g. based on DE test statistics) of the genes within a pathway.

A new concept in gene-set analysis was proposed by Mootha et al. (2003) with the gene-set enrichment analysis (GSEA) technique. It is a computational procedure that determines whether a presented gene list exhibits statistically significant, consistent differences between two biological conditions. The method uses previously defined gene-sets (e.g. pathways), which the presented gene list is compared against in the following manner. The genes in the gene list are ranked according to the association of their expression profiles with a phenotype of interest (e.g. DE between cases and controls). The extent of the association is then measured by a running sum statistic, designated the enrichment score (ES), which increases when a gene from a gene-set is encountered in the gene list and decreases when the opposite is true. An ES for each gene-set is determined and the maximum ES (MES) is found. Assessment of the statistical significance of the MES is performed using permutation testing of the samples. One thousand permutations are performed and the MES calculated for each permutation. An empirical distribution of the MES under the null hypothesis is generated. A p-value for the observed MES can then be calculated (Mootha *et al.* 2003). This method was limited in calculating only the significance of the most highly regulated gene-set in the dataset.

A variety of methods for gene-set analysis (GSA) have been inspired by the work of Mootha et al. (2003). A comprehensive list of algorithms and tools for GSA can be found in a study conducted by Nam and Kim (Nam and Kim 2008) and includes the significance analysis of function and expression (SAFE) method (Barry *et al.* 2005). Another approach is the *geneSetTest* function in limma (section 4.2.6.3) (Ritchie *et al.* 2015). These methods also account for the rank of the genes within the pathways based on their levels of differential expression. The pathways are then themselves ranked according to the ranks of their constituent genes. The different GSA methods use various statistical tests to rank the genes. For example, the SAFE method uses the two-sample t-test, which orders the genes according to the absolute value of their test statistics, based on their level of DE relative to their variability. GeneSetTest utilises t-statistics generated in single gene analysis for ranking the genes (section 4.2.6.3). Ranking the pathways is also performed using a variety of statistical tests, for example both the SAFE and geneSetTest methods utilise the Wilcoxon Rank-Sum statistic to rank the pathways (this is true for geneSetTest if “ranks.only=T”).

Two categories of hypotheses can be tested in GSA, self-contained tests and competitive tests (Goeman and Bühlmann 2007). The self-contained null hypothesis assumes that none of the genes in the gene-set are associated with the phenotype. This is the strictest null hypothesis, is therefore relatively easy to reject, yielding many significant gene-sets and exhibiting high statistical power (Wang 2010). Self-contained tests compute p-values by permuting the samples (phenotype of interest). The competitive null hypothesis assumes that genes in the gene-set are no more associated with the phenotype than genes outside the gene-set. In competitive tests, the p-values are generated by permuting the genes. Self-contained tests can be used for single gene, gene-set or genome-wide studies. In contrast, competitive tests can only be applied where a comparison can be made between results within a gene-set to those outside the gene-set (Fridley *et al.* 2010). Most methods can utilise both types of testing.

4.2.6.3 *Tools used for over-representation and gene-set analyses*

The online databases, GeneSetDB, Reactome and Wikipathways have functionality for performing enrichment analysis (section 4.2.6.1) (Araki *et al.* 2012, Croft *et al.* 2014, Kelder *et al.* 2012, Milacic *et al.* 2012). A list of the most highly DEG from individual gene expression analysis conducted in limma (section 4.2.5.3) was uploaded into the databases. The databases returned gene-sets (e.g. pathways) where there were more genes in the presented gene list that were members of the gene-set, than would be expected by chance. GeneSetDB results are returned with a raw p-value and an FDR adjusted p-value (section 4.2.5.2). Reactome returns pathways with FDR adjusted p-values whereas Wikipathways does not return p-values at all, preventing comment on the significance of the pathways returned.

GATHER is an online tool for enrichment analysis with several capabilities including gene ontology (Chang and Nevins 2006). Upon uploading a list of differentially expressed genes into GATHER (e.g. a list of genes identified as significantly differentially expressed by limma), a table containing a range of information is returned. The information presented includes a Bayes factor (derived from a modified Fisher's Exact Test) to quantify the evidence supporting an association between a gene-set and a given annotation (function). The magnitude of the Bayes factor corresponds to the strength of the evidence for the association. A natural log of the Bayes factor (ln(Bayes factor))

value over 6.0 is the cut-off value for showing conclusive evidence supporting an association between the genes presented to GATHER and the given annotation (Chang and Nevins 2006).

Gene-set analysis enrichment was performed using the *geneSetTest* function in limma (Ritchie *et al.* 2015). The function tests the hypothesis that the presented gene-set tends to contain genes with significantly higher ranks (in terms of DE) in comparison to randomly selected gene-sets of the same size and calculates a p-value. As a competitive test (section 4.2.6.2), the *geneSetTest* function tests the gene-sets for DE relative to the other gene-sets. In permuting the genes, the function assumes that the genes are independent, that is that the genes in the gene-set are no more correlated than randomly determined genes. An indicator matrix of the genes from individual gene expression analysis performed in limma, matched with pathways that they participate in from the KEGG and GO databases (section 4.2.6.1) is created first. This matrix is used to indicate which genes are members of the gene-sets to be analysed. The matrix along with the matched t-statistics for the genes being presented for GSA is then presented to the *geneSetTest* function, which uses a Wilcoxon Rank-Sum test to order the pathways based on the ranking of the t-statistics of the genes within the pathways (section 4.2.6.2). Multiple hypothesis testing (section 4.2.5.2) was corrected for using the Benjamini and Hochberg method, based on the total number of pathways being examined (Benjamini and Hochberg 1995). Pathways are returned with adjusted p-values.

4.2.7 *Visualisation methods*

4.2.7.1 *Heatmaps*

Heatmaps allow the visualisation of substantial quantities of data that are stored in a matrix- or spreadsheet-like format. In molecular biology, heatmaps are typically utilised to illustrate gene expression across experimental conditions. Clustering algorithms utilising distance measures are employed to group gene expression variables according to similarity. The clustering algorithms can use two approaches, the agglomerative and devisive approaches. The agglomerative method makes distance comparisons beginning with individual objects. Conversely, the devisive method makes distance comparisons beginning with the entire group.

Distance between expression values can be measured in various ways. Selecting a distance measure can depend on variable type, measurement scale and knowledge of the discipline. The Euclidean distance is often selected when comparing profiles of respondents across variables. It is only appropriate for data measured on the same scale. It is defined as the square root of the sum of the squared differences between the corresponding components of two vectors.

Distance measures of pairs of variables can be represented by a distance matrix. The distance matrix is used by the clustering algorithm to generate a dendrogram (cluster tree). The agglomerative approach clusters utilising linkage methods, single, complete and average. The single linkage procedure measures distance between nearest neighbours. Complete linkage measures distance between farthest neighbours and has the advantage of being sensitive to outliers. Average linkage takes the average distance between the variables in question.

In the present study a modification of the *heatmap.2* function from the gplots package (Warnes *et al.* 2015) (available from the Bioconductor project (Gentleman *et al.* 2004)) was used to generate heatmaps (code provided by Associate Professor Mik Black). The modification provides the ability to have several clustering colour bars in the margin above a heatmap. This function utilises the hierarchical clustering algorithm which defaults to the agglomerative method of conducting distance comparisons and complete linkage. The function uses Euclidean distance as the default measure of distance between expression profiles. The hierarchical cluster analysis utilises a group of dissimilarities for the expression variables. Each variable is allocated to a group (cluster). The algorithm proceeds iteratively to make comparisons between the groups, joining the two most similar groups at each step. This process continues until there is a single cluster. At each step the distances between the groups are recalculated by a dissimilarity update formula according to the specified linkage method chosen (complete, in this case) and the specified distance measure method chosen (Euclidean, in this instance).

A red/green colour map is often used in heatmaps representing expression data. The red colouring represents elevation in gene expression and the green colouring reduction in

gene expression. Magnitude of elevation or reduction in gene expression is directly related to the brightness of colour.

Patterns in the heatmap may indicate an association between gene expression and the experimental conditions being tested. Clustering for a range of characteristics can be performed. For example, if clustering for a clinical characteristic, the clinical information can be presented to the *heatmap* function (in this case) in the form of a matrix. The *heatmap* function will perform the usual clustering of the expression data for the experimental conditions being investigated. It will also show clustering of the clinical variable of interest via a coloured bar across the top of the heatmap specifying, using colour dictated by the user, where the samples both with or without the clinical characteristic can be located in the heatmap. This allows for possible associations between gene expression and clinical variables to be more readily seen.

4.2.7.2 *Barcode plots*

The barcode plot enables the visualisation of pathway gene statistics computed during gene expression analysis (Ritchie *et al.* 2015).

The enrichment curve above the barcode plot is calculated by tri-cube weighted moving average, the traditional weight function used for LOESS (a local regression method) (Cleveland 1979). In this method the data points closest to the local point of estimation are given more weight than data points further afield. The point of estimation is based on the local data subset and thus 'moves' as each data subset is addressed for local regression fitting. The curve is a visualisation of the local regression fitting.

In the present study, the t-statistics generated from analyses performed in limma were presented to the *barcodeplot* function (section 4.2.5.1). As negative t-statistics represent down-regulated gene expression and positive t-statistics represent up-regulated gene expression, visualising these in a graph can enable the viewer to evaluate whether the pathway is being essentially up- or down-regulated. A pathway may appear primarily up-regulated whilst containing negative t-statistics (or vice versa). Genes in the pathway exhibiting a negative t-statistic may have been down-regulated allowing the overall impact of the pathway to be one of up-regulation. That is, the inhibited gene may need to be down-regulated in order for the pathway to be up-regulated overall. In the context of

the present study, where comparisons were being made between treated and untreated patients, an up-regulated pathway would be interpreted as being up-regulated in the treated group compared with the untreated group.

4.2.8 *Metagene creation for summarising gene-sets*

Summarising the activity of a gene-set provides the opportunity to investigate associations between the pathway and clinical variables. Summarisation of pathway activity is performed by reducing the dimension of the gene-set expression data matrix and extracting the strongest signal. This is carried out via singular value decomposition (Alter *et al.* 2000), using the *svd* function of the base package in R (R-Core-Team 2014). The data matrix is reduced to a series of principal components and singular values typifying the extent of variation in each principal component. The first principal component relates to the first singular value representing a dimension of the data in which the majority of the variation is detected. In the context of this study, the first principal component provides a summary (metagene (Spang *et al.* 2002, West *et al.* 2001)) of the gene activity for the pathway in question. However, it is important to note that the variation may not be due to differences in pathway activity but due to differences in experimental or clinical factors. Investigating the level of metagene expression between experimental conditions or clinical variables may provide insight into the underlying biology of the conditions being investigated.

Comparisons of the level of metagene between the two arms of the study were tested using the Welch's two sample t-test (Welch 1947).

4.3 Summary

This chapter described the methods used to measure the primary outcome. Ribonucleic acid extracted from normal and tumour tissue of study participants was profiled on microarray gene-chips. Data analysis was performed comparing the treatment and placebo groups in normal, tumour and paired analyses.

5. RESULTS OF DIFFERENTIAL GENE EXPRESSION ANALYSIS

5.1 Introduction

This chapter presents the results for the primary outcome of this study. The study was devised to measure alterations in the gene expression of the vitamin D transcriptome (vitamin D responsive genes) in the tumour tissue and corresponding normal tissue of CRC patients. The patients were supplemented with a single large dose of vitamin D prior to surgery. To measure the molecular impact of vitamin D on its responsive genes, RNA was extracted from the normal and tumour tissue and profiled on microarray gene-chips. Bioinformatic methods were employed to analyse gene expression data generated by the arrays. Gene expression data were analysed for significant alterations in expression in single genes, and in gene-sets, between the treatment and control groups of the study. Specifically, the primary endpoint was to measure a 1.5-fold change in the activity of vitamin D activated genes. This chapter presents the findings for the primary endpoint and outcome.

Analysis of gene expression was conducted for (1) normal, (2) tumour and (3) paired tissue. Each of these analyses was performed making comparisons between the treatment and placebo arms of the trial. This chapter is divided into sections based on each of the three analyses listed above. Within each section single gene, over-representation and gene-set analysis is presented. The gene-set analysis section is further sub-divided for analysis involving data from the KEGG database and GO project.

5.2 Microarray quality control and data normalisation

The arrayQualityMetrics package (Kauffmann *et al.* 2009), for R was used to ascertain the quality of the array expression data (section 4.2.2). The expression data were assessed using the package both pre- and post-normalisation. Summaries of the reports follow. The full reports can be found in Appendices 3 and 4 respectively.

5.2.1 *Microarray quality control - pre-normalisation*

Comparisons made between arrays depicted five outliers in the false colour heatmap and associated bar chart. Each of these outlier arrays exhibited a sum of distances outside the threshold of 91.2. Outliers were not highlighted in the PCA scatterplot (Appendix 3).

The boxplots for the array intensity distribution were not similar in position or width. One outlier was identified which also appeared in the between array comparisons. The density histograms also were not of similar shapes and sizes, indicating differences in the intensity distributions of the arrays.

The variance-mean density plot illustrates a positive association between the mean and standard deviation. This indicates that the standard deviation of the array signal intensities were not independent of the rank of their mean, signifying that the arrays do not have similar signal intensity.

The Affymetrix specific plots presenting the Relative Log Expression (RLE) and the Normalised Unscaled Standard Error (NUSE) with their associated bar charts identified seven and nine outliers respectively. These outliers were also identified in the post-normalisation report. All of the arrays had a similar slope in the RNA degradation plot signifying that none of the arrays contained any problematic data. The slope was not particularly high indicating that RNA degradation did not appear to be a problem. However, RNA integrity numbers (RINs) were measured for three patient normal and tumour samples in duplicate (12 samples). One set of patient normal and tumour tissue had been immersed in RNAlater® overnight before freezing at -80^{o}C and the other was put directly into the -80^{o}C freezer. The RINs had a post-adjusted range from 2.3-3.9 and were not significantly different between the RNAlater® and the -80^{o}C group. RINs upwards of seven (in a range from 0-10) are considered to indicate good quality RNA (ie little degradation) (Marx 2004). RINs were not measured for all of the study tissue samples as they were irreplaceable specimens and thus needed to be used regardless. The indication from the 12 RINs measured is that the RNA quality was low. Placement of tissue samples either in RNAlater® or directly into the -80^{o}C freezer varied with a range of 30-90 minutes depending on the availability of a pathologist. The delay in stabilising the RNA may have resulted in the lower quality RNA observed in this study. The fact that

the RNA degradation plot did not indicate an issue may be due to all of the RNA samples being of low quality. (For more information on the RNA purity and quality see Appendices 5a and 5b respectively). Despite these low scores (which are not uncommon for tumour specimens), it is still possible to obtain acceptable expression data from low quality RNA. While more subtle differences in gene expression may be lost, substantial changes are usually still measurable (L. McNoe, Otago Genomics and Bioinformatics Facility; personal communication, 2015).

The individual array quality MA plots and corresponding bar chart presented 25 outliers. Twenty five of the Hoeffding's statistics calculated for the arrays were outside the threshold of 0.15.

The false colour plots for the arrays' spatial distributions of feature intensities and associated bar chart identified one outlier (Appendix 3).

5.2.2 *Microarray quality control - post-normalisation*

Between array comparisons highlighted seven outliers in the false colour heatmap and associated bar chart. These arrays each exhibited a sum of distances outside the threshold of 49.8. The post-normalisation threshold was greatly reduced compared to the pre-normalisation threshold of 91.2. Eight outliers were highlighted in the PCA scatterplot, including the seven observed in the heatmap. Normalisation of the data had the effect of diminishing the distances between the arrays, which can be observed when comparing these plots between pre- and post-normalisation (Appendix 4).

The boxplots for array intensity distribution were similar in position and width. The plots and associated bar chart featured the same eight outliers as in the between array comparisons. The eight outliers exceeded the threshold KS statistic of 0.081, indicating that their array intensity distributions were outside the overall intensity distribution for the combined data. The post-normalisation threshold was substantially reduced compared to the pre-normalisation threshold of 0.335. The density histograms were of similar shapes and sizes. Normalisation ensures that the distribution of the probe intensities is the same across the arrays by removing any array effects, and thus places the arrays on a common scale.

The variance-mean density plot for the array data was still not approximately horizontal after normalisation as expected. There was a difference between the plots pre- and post-normalisation, however, with the normalisation process having markedly decreased the strength of the mean-variance relationship.

The individual array quality MA plots and corresponding bar chart presented no outliers. All of the Hoeffding's statistics calculated for the arrays were well within the threshold of 0.15. The quantile normalisation step of the normalisation process ensured that the probe intensity distribution was the same across the arrays.

Outliers featuring in the quality control (QC) exercise were still included in the analysis. They were not considered to be extreme and thus their overall impact was likely to be relatively minor. Alongside this, maintaining the power of the study by including 25 arrays in each group being compared in the analyses was important for robust results to be achieved (Appendix 4).

Data were normalised utilising the RMA method (Irizarry *et al.* 2003a) both with and without background adjustment. In accordance with current benchmarking, it was found that normalising the data without background adjustment was preferable (section 4.2.3).

5.3 Normal tissue - control versus treatment

Normal tissue was collected from 56 patients. Twenty-seven of the patients were in the control arm and 29 were in the treatment arm of the trial. RNA from each normal tissue sample was profiled on a separate microarray gene-chip.

5.3.1 *Single gene analysis*

Individual gene analysis was performed where a linear model was fitted to the gene expression data in order to detect the genes that had undergone significant changes in expression. Differential expression was evaluated using an empirical Bayes method, which generates a modified test statistic (t-statistic). The modified t-statistic has the dual effect of controlling the overall impact of genes with very small standard deviations and highlighting significantly expressed genes (section 4.2.5.1).

Following is a linear model results summary table of the statistical information derived from the analysis of gene expression in the normal tissue between the control and treatment groups (Table 5.1). The table includes summary statistical information for the top 10 differentially expressed genes. Multiple testing was adjusted for using the FDR controlling method of Benjamini and Hochberg (Benjamini and Hochberg 1995). No genes were found to have adjusted p-values below the significance threshold of 0.05.

Gene	Log fold change	Average expr	t statistic	p-value	Adjusted p-value	B statistic
LCN2	1.21	8.92	3.72	<0.001	1	-0.9
SULT1C2	0.32	6.1	3.65	<0.001	1	-1.03
ZNF137P	-0.15	6.32	-3.34	<0.01	1	-1.65
DUOX2	0.59	7	3.28	<0.01	1	-1.75
LYPD6B	0.2	6.45	3.12	<0.01	1	-2.05
C10orf116	-0.52	9.88	-3	<0.01	1	-2.28
SLC38A4	-0.42	6.46	-2.82	<0.01	1	-2.57
SLC7A11	0.36	6.86	2.82	<0.01	1	-2.58
DUOXA2	0.34	6.84	2.62	<0.05	1	-2.92
RFT1	-0.12	7.28	-2.61	<0.05	1	-2.94

Table 5.1. Linear model results summary table for normal tissue analysis. Summary statistical information derived from analysis of gene expression in the normal tissue between the two study groups. The table shows summary information for the top 10 genes (based on raw p-values). After adjustment for multiple testing, no genes demonstrated an adjusted p-value below the significance threshold of 0.05.

A heatmap of the 100 most highly differentially expressed genes (DEG) from this analysis can be viewed in Appendix 6 (Figure 4A).

The p-values for the VDRE gene list were extracted from the linear model results summary table and adjusted for multiple testing using the number of genes in the VDRE gene list. The VDRE gene list comprised 3136 genes (section 4.2.4). None of the adjusted p-values were significant. All of the adjusted p-values were 1, implying that the observed differences in expression could have occurred by chance.

5.3.2 *Over-representation analysis*

Presenting a list of the top 100 DEG to the Wikipathways database for enrichment analysis (section 4.2.6) identified 110 pathways exhibiting gene enrichment in the treatment group compared with the placebo group. These pathways covered a wide range of activity from regulation of DNA replication through fatty acid beta-oxidation, vitamin D receptor and metabolism pathways to oxidation by cytochrome P450, cell differentiation, regulation of apoptosis and focal adhesion. The selected pathways above are interesting in that the majority of them have featured in the literature with evidence supporting the role of vitamin D in modulating these activities. However, this is a hand-picked list of eight pathways from the 110 returned from the over-representation analysis in Wikipathways. Results are not returned with p-values or FDRs thus preventing comment on the significance of the pathways. The results from Wikipathways were not helpful due to too many pathways being returned and the fact that they were produced without significance information.

Submitting a gene list of the 50 most DEG from single gene analysis to the GeneSetDB database, returned a number of pathways showing gene enrichment in the treatment arm compared to the control arm. These pathways were significant at the 0.05 significance threshold based on FDR control. When all of the available databases were selected for enrichment analysis, the Vitamin Metabolic Process term (GO:0006766) featured from the GO project with an FDR adjusted p-value of 0.013. This GO term refers to a biological process including chemical reactions and pathways involving vitamins, both water- and fat-soluble. When the 'pathway' databases were selected for enrichment analysis two vitamin metabolism Reactome pathways featured, Metabolism of Vitamins and Cofactors, and Metabolism of Water-soluble Vitamins and Cofactors. Both pathways had an FDR adjusted p-value of 0.015. Submitting a list of the top 100 DEG and selecting all the databases for enrichment analysis returned the same GO term as when 50 genes were presented to GeneSetDB. The FDR adjusted p-value was 0.029. Selecting only the 'pathway' databases did not return any significant pathways when 100 genes were submitted. Vitamin processing pathways featured in the Wikipathways enrichment analysis and also in the limma analysis described below.

Uploading lists of 50 or 100 of the top DEG into the Reactome database and GATHER tool did not return any significant pathways.

5.3.3 *Gene-set analysis - KEGG*

Results from GSA performed in the normal tissue compared between the two groups using the *geneSetTest* function in limma and the pathway data from the KEGG database are presented in Table 5.2. An FDR adjusted p-value of 0.05 was chosen as the cut-off for determining pathway significance. Eleven pathways significantly enriched for highly ranked genes were returned. They were all metabolism pathways. The Fatty Acid Metabolism pathway (40 genes) shares genes with the PPAR Signaling (69 genes, 17 shared) and Propanoate Metabolism (32 genes, 11 shared) pathways. It has genes in common with other metabolism pathways also, but not to the same degree. For example, it shares seven genes with the tyrosine pathway (41 genes). The remaining metabolism pathways also share genes although in lower numbers. Many of the significant metabolism pathways may arise simply because they have genes in common with an 'authentically' significant pathway. The cytochrome P450 pathways share a considerable number of their genes. The pathways have 60 and 61 genes, 50 of which are in common.

The Fatty Acid Metabolism and Metabolism of Xenobiotics by Cytochrome P450 pathways were selected as representatives for the group of significant pathways as they were two of the more highly significant pathways. Barcode plots for the Fatty Acid Metabolism and Metabolism of Xenobiotics by Cytochrome P450 KEGG pathways can be seen in Figures 5.1 and 5.2. Alongside these are boxplots presenting pathway metagene expression between the treatment and control arms of the trial.

KEGG id	KEGG Pathway	p-value	Adjusted p-value (fdr)	Pathway size
"01100"	Metabolic pathways	<0.001	<0.001	1074
"00071"	Fatty acid metabolism	<0.001	<0.001	40
"00982"	Drug metabolism - cytochrome P450	<0.001	0.002	61
"00980"	Metab of xenobiotics by cytochrome P450	<0.001	0.003	60
"04977"	Vitamin digestion and absorption	<0.001	0.003	23
"00350"	Tyrosine metabolism	<0.001	0.012	41
"00590"	Arachidonic acid metabolism	<0.001	0.018	56
"03320"	PPAR signaling pathway	<0.001	0.018	69
"04974"	Protein digestion and absorption	<0.001	0.022	75
"00640"	Propanoate metabolism	0.0011	0.024	32
"00030"	Pentose phosphate pathway	0.003	0.05	26

Table 5.2. Significant KEGG pathways featured in GSA of normal tissue compared between study groups. KEGG id = KEGG pathway identification number. Pathway size= the number of genes in the pathway. p-value adjusted using the Benjamini and Hochberg FDR correction method. The more highly significant genes can be seen represented in the top half of the 'adjusted p-value' column of the table, with values ranging from <0.001 to 0.003.

The first of the significant KEGG pathways, 'Metabolic Pathways' is an amalgamation of eleven different metabolic pathways and includes energy, carbohydrate, lipid and nucleotide metabolism, metabolism of vitamins and co-factors and xenobiotics biodegradation and metabolism. This 'pathway' is defined as a metabolism overview map by KEGG. It may feature as a significant 'pathway' here because it incorporates some of the other featured pathways, for example, the cytochrome P450 pathways. A barcode plot was not created for the 'Metabolic Pathways' pathway as it contains many pathways and too many genes to be a useful visualisation of the pathway gene expression information.

Vitamin D involvement has been reported in the degradation of xenobiotics in the intestinal tract (Schmiedlin-Ren, 2001, (Thummel *et al.* 2001). In CRC cells, calcitriol enhances the expression of *CYP3A4* (Thummel *et al.* 2001), (Thompson *et al.* 2002), an oxidising enzyme that detoxifies small xenobiotic molecules expediting their excretion from the body (section 1.3.4.3e). *CYP3A4* can be found in both of the Metabolism of Xenobiotics by Cytochrome P450 and Drug Metabolism – Cytochrome P450 pathways alongside 19 and 15 genes from the VDRE gene list, respectively. Of the 60 and 61 genes

in each of these pathways, 50 genes are shared. The pathways appeared to be down-regulated in the normal tissue of the treated arm compared with the untreated group.

A metagene was created for the Metabolism of Xenobiotics by Cytochrome P450 pathway and represents the first principal component (summary) of the pathway expression data (section 4.2.8). There was no difference in the level of pathway metagene expression in the normal tissue between the two groups. This indicates that the genes in the pathway do not have a consistent pattern of expression in the normal tissue across the treatment arm, although the barcode plot indicates a tendency for these genes to be down-regulated in the normal tissue of the treated patients (Figure 5.1).

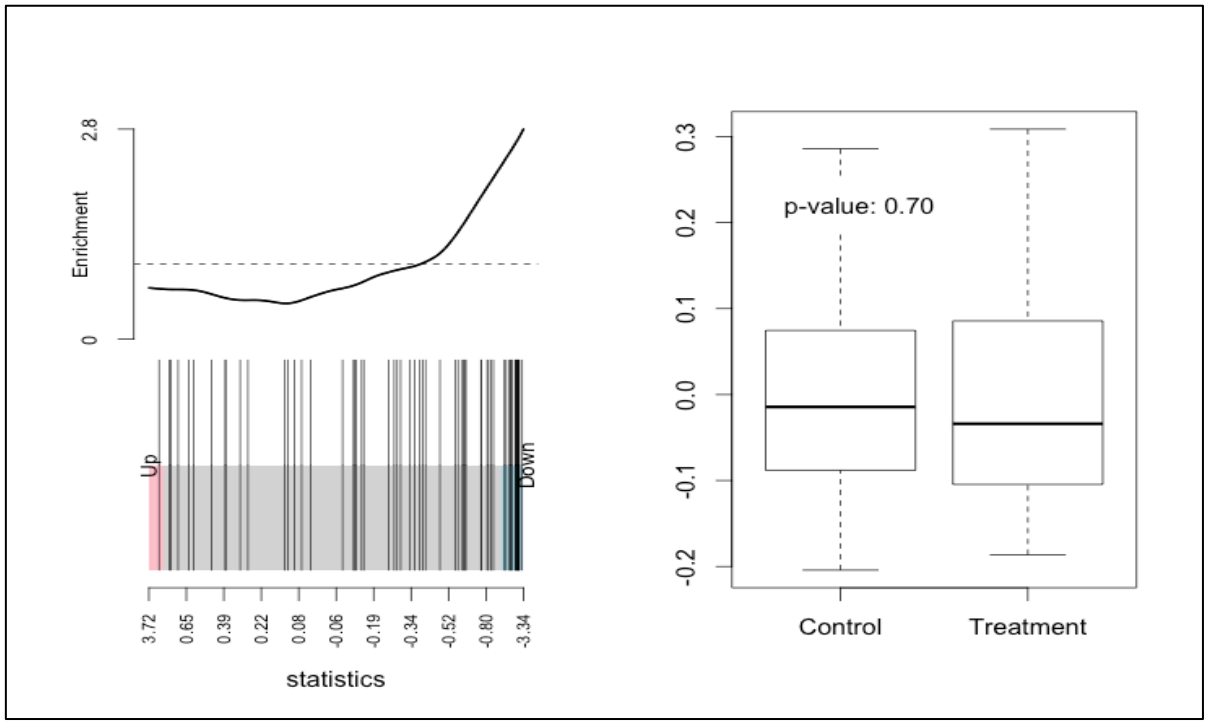


Figure 5.1. Barcode plot for the Metabolism of Xenobiotics by Cytochrome P450 pathway and boxplot of 'Metabolism of Xenobiotics by Cytochrome P450 pathway' metagene expression. The barcode plot presents the t-statistics for the genes of the Metabolism of Xenobiotics by Cytochrome P450 pathway. The enrichment curve is the local regression of the pathway t-statistics. The pathway has 60 genes. Many of the genes for the pathway are down-regulated. The pathway overall appears down-regulated in the normal tissue of the treated patients. The boxplot presents the level of metagene expression in the normal tissue compared between the treatment and control arms of the trial. There is no difference in the level of metagene expression between the two groups.

Fatty acids are preferentially utilised by heart and skeletal tissue as a high yield fuel source; large quantities of adenosine triphosphate (ATP) are released. Short chain fatty acids are produced by bacteria residing in the colon. Colonic epithelia use these as their main energy source. In KEGG, the Fatty Acid Metabolism pathway is defined as a Fatty Acid Degradation pathway where long chain fatty acids are catabolised, releasing acetyl coenzyme A (acetyl coA). Acetyl coA enters the tricarboxylic acid cycle (TCA). Nicotinamide adenine dinucleotide (NADH) produced in the TCA is fed into the oxidative phosphorylation pathway (electron transport). The net outcome of the closely linked TCA and oxidative phosphorylation pathway is the oxidation of nutrients yielding ATP. Vitamin D up-regulated protein 1 (VDUP1) has been proposed as a regulator of fatty acid utilisation (Kim et al. 2007). The Fatty Acid Metabolism pathway was down-regulated in the normal tissue of treated patients compared with the untreated patients (Figure 5.2).

A Fatty Acid Metabolism pathway metagene was generated (section 4.2.8). The level of metagene expression was not different in the normal tissue between the treated and untreated participants, demonstrating that the pattern of pathway gene expression is inconsistent in the normal tissue across the treated group (Figure 5.2).

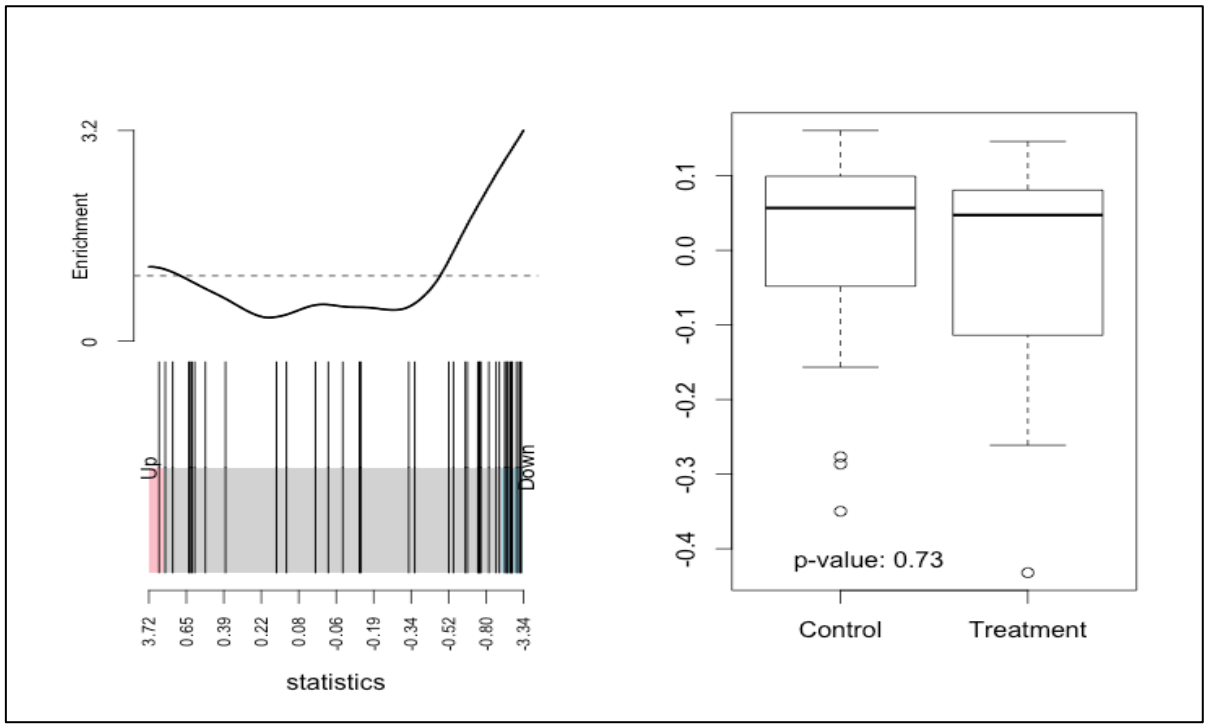


Figure 5.2. Barcode plot for the Fatty Acid Metabolism pathway and boxplot of 'Fatty Acid Metabolism pathway' metagene expression. The barcode plot presents the t-statistics for the genes of the Fatty Acid Metabolism pathway. The enrichment curve is the local regression of the pathway t-statistics. The pathway has 40 genes. Many of the genes for the pathway are down-regulated. The pathway overall appears down-regulated in the normal tissue of the treated arm. The boxplot presents the level of metagene expression in the normal tissue compared between the two groups. There is no difference in metagene expression between the groups.

The vitamin digestion and absorption pathway involves the digestion and absorption of both water-soluble and fat-soluble vitamins including fat-soluble vitamin D in the form of cholecalciferol from the diet. The vitamin digestion and absorption pathway was down-regulated in the analysis in normal tissue in the treatment arm compared to the placebo arm. Visualisation plots of this result are not included.

The fatty acid metabolism, cytochrome P450 and vitamin digestion and absorption pathways are all down-regulated in the normal tissue of the treatment arm compared to the control arm. These findings are discussed further with the findings of the gene-set analysis utilising pathway information from the GO project in section 5.3.3.

5.3.4 *Gene-set analysis - GO*

Table 5.3 presents the results from GSA performed in the normal tissue comparing the two groups using the *geneSetTest* function in limma and the pathway/process data from

the GO project. An adjusted p-value of 0.05 was selected as the cut-off for establishing pathway significance. Fourteen pathways/processes were identified that were significantly enriched for highly ranked genes in the analysis. The first two 'processes' are the most significant 'pathways' returned. They contain too many genes to visualise well in a barcode plot. The significant processes represented in this analysis are distinct in their activity until approaching some of the metabolic pathways lower in the table (5.3). Very few genes are shared by the latter pathways. For example, the Retinoid Metabolic Process pathway shares three genes with the Triglyceride Catabolic Process pathway. A barcode plot for the Negative Regulation of Growth pathway can be seen in Figure 5.3. A boxplot presenting pathway metagene expression is plotted alongside.

GO id	GO Term	p-value	Adjusted p-value (fdr)	Pathway size
GO:0005615	Extracellular space	<0.001	<0.001	965
GO:0044281	Small molecule metabolic process	<0.001	<0.001	1356
GO:0045926	Negative regulation of growth	<0.001	0.013	17
GO:0000278	Mitotic cell cycle	<0.001	0.025	380
GO:0005788	Endoplasmic reticulum lumen	<0.001	0.026	165
GO:0071294	Cellular response to zinc ion	<0.001	0.026	12
GO:0007586	Digestion	<0.001	0.027	44
GO:0030212	Hyaluronan metabolic process	<0.001	0.027	22
GO:0042953	Lipoprotein transport	<0.001	0.036	14
GO:0048015	Phosphatidylinositol-mediated signaling	<0.001	0.036	128
GO:0071682	Endocytic vesicle lumen	<0.001	0.036	14
GO:0001523	Retinoid metabolic process	<0.001	0.038	54
GO:0005975	Carbohydrate metabolic process	<0.001	0.039	320
GO:0019433	Triglyceride catabolic process	<0.001	0.042	19

Table 5.3. Significant GO terms featured in GSA of normal tissue compared between study groups. GO id = GO term identification number. Pathway size = the number of genes in the pathway. p-value adjusted using the Benjamini and Hochberg FDR correction method. The top two 'processes' are the most significant 'pathways' featured in this analysis based on the adjusted p-value information.

In the GO project the Negative Regulation of Growth is a biological process term. It is described as being "any process that stops, prevents or reduces the rate or extent of growth, the increase in size or mass of all or part of an organism." Vitamin D moderates many genes involved in proliferation (section 1.3.4.2a). Research investigating vitamin D's impact on proliferation has frequently featured the WNT/β-catenin signaling

pathway. The barcode plot of the Negative Regulation of Growth pathway presents a down-regulated pathway in the normal tissue of the treated group compared with the untreated group (Figure 5.3).

The pathway metagene boxplot depicts a lack of difference in the expression of the metagene between the treated and untreated trial participants (Figure 5.3). This signifies that the pathway genes do not have a consistent pattern of expression in the normal tissue across the treatment group.

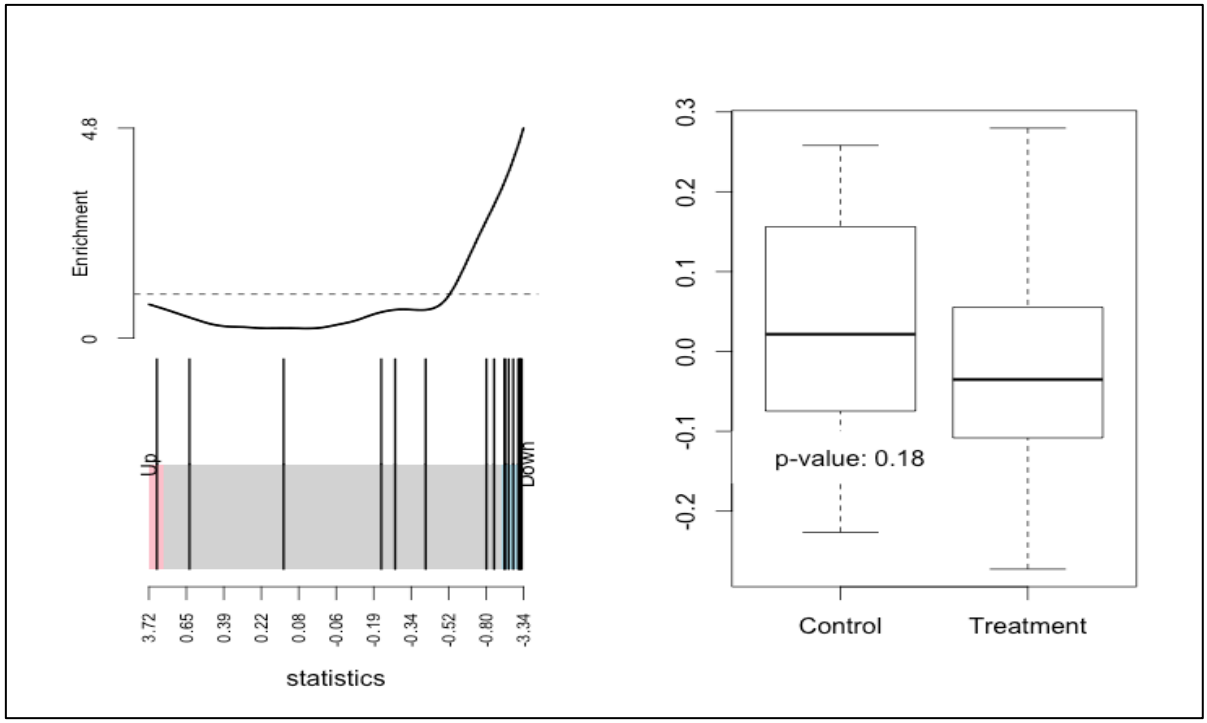


Figure 5.3. Barcode plot for the Negative Regulation of Growth pathway and boxplot of 'Negative Regulation of Growth pathway' metagene expression. The barcode plot presents the t-statistics for the genes of the Negative Regulation of Growth pathway. The enrichment curve is the local regression of the pathway t-statistics. The pathway has 17 genes. Most of the genes for the pathway are down-regulated. The pathway overall appears down-regulated in the normal tissue of the treatment arm. The boxplot presents the level of metagene expression in the normal tissue compared between the two groups. There is no significant difference in metagene expression between the groups.

Differences in gene expression in the normal tissue of the treatment arm compared with the placebo arm, seem apparent in the down-regulation of fatty acid metabolism, xenobiotic metabolism, vitamin digestion and absorption and the negative regulation of growth. Vitamin processing pathways with gene enrichment were also identified in the over-representation analysis. When comparing the level of the pathway metagene

expression between the groups, for the identified pathways there was no significant difference. This indicated that the gene expression pattern for the identified pathways was not strongly consistent in the normal tissue across the treated patients compared with the untreated group. It is not immediately obvious why these pathways might be down-regulated in response to vitamin D in the normal tissue. A comparison of these findings with those from analysis performed in the tumour tissue may provide some insight.

5.4 Tumour tissue - control versus treatment

Tumour tissue was collected from 51 patients. Three of the patients had two primary tumours. Twenty-five of the patients were in the control arm and 26 were in the treatment arm equating to 26 tumour tissue samples for the control arm and 28 for the treatment arm. Thus the total number of tumour tissue samples was 54. RNA from each tumour tissue sample was profiled on a separate microarray gene-chip.

5.4.1 *Single gene analysis*

Analysis of individual gene expression was performed in tumour tissue comparing the treatment and control groups. This was achieved utilising the limma package, which uses a linear models method and an empirical Bayes modified t-statistic. The linear model results summary table is presented for the 10 most highly differentially expressed genes (Table 5.4). After adjustment for multiple testing using the FDR controlling procedure of Benjamini and Hochberg (1995), no genes displayed adjusted p-values below the significance threshold of 0.05.

Gene	Log fold change	Average expr	t statistic	p-value	Adjusted p-value	B statistic
MSRB2	-0.41	9.03	-3.47	<0.01	0.97	-1.32
C5orf35	-0.31	6.31	-3.32	<0.01	0.97	-1.62
CBX2	0.32	6.73	3.31	<0.01	0.97	-1.64
SLC43A2	0.18	7.76	3.29	<0.01	0.97	-1.68
MPP7	-0.47	7.62	-3.29	<0.01	0.97	-1.69
CMAS	-0.37	8.5	-3.26	<0.01	0.97	-1.74
CDR2L	0.2	7.37	3.25	<0.01	0.97	-1.75
GRAMD1C	-0.31	6.36	-3.25	<0.01	0.97	-1.76
DERA	-0.39	8.71	-3.21	<0.01	0.97	-1.82
TMLHE	-0.16	6.69	-3.19	<0.01	0.97	-1.87

Table 5.4. Linear model results summary table for tumour tissue analysis. Summary statistical information derived from analysis of gene expression in the tumour tissue comparing the two study groups. The table shows summary information for the top 10 genes (based on raw p-values). After adjustment for multiple testing no genes demonstrated an adjusted p-value below the significance threshold of 0.05.

A heatmap of the 100 most highly DEG from this analysis in tumour tissue can be viewed in Appendix 6 (Figure 4B).

The VDRE gene list was examined for differential gene expression. The p-values were extracted from the statistical summary table and adjusted for multiple testing using the number of genes in the VDRE gene list, 3136. None of the adjusted p-values were significant with a range of 0.98-1. This infers that the observed alterations in expression could have occurred by chance. This result raises the question of whether the vitamin D infiltrated the tumour tissue.

5.4.2 *Over-representation analysis*

The top 100 DEG from the limma analysis were uploaded into the Wikipathways database for enrichment analysis (section 4.2.6). One hundred and twenty two pathways were identified with gene enrichment in the treated patients compared with the untreated patients. These pathways covered a wide range of activity, as for the normal tissue. A similar selection of pathways, as was selected from enrichment analysis in the normal tissue, can be extracted from the list of 122 pathways, with a number of unique additions. These unique additions include the MAPK Signaling, Fatty Acid Biosynthesis, Glycogen Metabolism and TCA (tricarboxylic acid cycle) and Respiratory Electron Transport pathways. Some of the selected pathways arise in analyses elsewhere in this

study and some of them are associated with the hallmarks of cancer. The MAPK Signaling Pathway was identified in the GSA, using KEGG data, below and also in the over-representation analysis of the paired analysis. The Fatty Acid Biosynthesis pathway was selected because fatty acid pathways feature in the normal and tumour analyses in this study. Energy pathways were chosen because of their connection with the emerging cancer hallmark, deregulation of energy metabolism. Deregulation of energy metabolism by tumour cells can ensure an increased supply of energy for the up-regulated cell proliferation of the tumour. These are a hand-picked list of pathways from the 122 returned pathways of the enrichment analysis in Wikipathways. The over-representation analysis performed by Wikipathways gave too many pathways, and without significance data, to be useful.

Uploading a gene list of the most highly differentially expressed genes, from single gene analysis, into both the GeneSetDB and Reactome databases featured pathways with gene over-representation in the treated participants compared with the untreated patients. The Pentose Phosphate and Propanoate Metabolism pathways each with an FDR adjusted p-value of 0.047 featured from enrichment analysis in GeneSetDB when the top 50 differentially expressed genes were submitted and the 'pathway' database selected. Metabolism pathways feature in the following GSA utilising KEGG and GO data.

Three pathways were identified with gene enrichment in the treated versus the untreated group when the top 100 DEG were submitted to Reactome. The Attenuation Phase (FDR <0.01), HSF1-dependent Transactivation (FDR <0.01) and HSF1 Activation (FDR <0.05) pathways are all involved in the cellular response to stress pathways. When the top 50 DEG were uploaded no significant pathways were produced.

Presenting a list of the 50 or 100 most DEG to the GATHER tool did not return any significant pathways.

5.4.3 *Gene-set analysis - KEGG*

Significant pathways from GSA performed in the tumour tissue where comparisons were made between the two study groups are displayed in Table 5.5. The analysis was carried out using the *geneSetTest* function in limma and data from the KEGG database. Significant enrichment was observed in 12 pathways. An adjusted p-value of 0.05 was

chosen as the cut-off for ascertaining pathway significance. The Fatty Acid Metabolism pathway (40 genes) shares genes with the beta-Alanine Metabolism pathway (22 genes, 9 shared), the Propanoate Metabolism pathway (32 genes, 11 shared), the PPAR Signaling Pathway (69 genes, 17 shared) and the Peroxisome pathway (77 genes, 10 shared). The beta-Alanine Metabolism (22 genes) and Propanoate Metabolism (32 genes) pathways also share genes (12 genes). The PPAR Signaling Pathway (69 genes) and Peroxisome pathway (77 genes) have genes in common too (13 genes). All of these pathways appear to be down-regulated. Some of them may feature in this analysis because they share genes with a 'significant' pathway, such as the Fatty Acid Metabolism pathway.

The PPAR Signaling Pathway does modulate genes involved in fatty acid transport and oxidation and therefore may be authentically identified as a significant pathway. Its barcode plot is not particularly convincing however, and thus has not been included. No further work was done with this pathway. Oxidative Phosphorylation and the neurodegenerative pathways listed in the table (5.5) share over half of their genes. This observation may explain why the neurodegenerative diseases appear here. Barcode plots for the Fatty Acid Metabolism and Oxidative Phosphorylation KEGG pathways can be seen in Figures 5.4 and 5.6. Boxplots showing pathway metagene expression are presented in these figures also.

KEGG id	KEGG Pathway	p-value	Adjusted p-value (fdr)	Pathway size
"00071"	Fatty acid metabolism	<0.001	<0.001	40
"05012"	Parkinson's disease	<0.001	<0.001	118
"00280"	Valine, leucine & isoleucine degradation	<0.001	0.002	44
"01100"	Metabolic pathways	<0.001	0.003	1074
"00410"	beta-Alanine metabolism	<0.001	0.004	22
"00190"	Oxidative phosphorylation	<0.001	0.005	120
"00640"	Propanoate metabolism	<0.001	0.005	32
"03320"	PPAR signaling pathway	<0.001	0.006	69
"05010"	Alzheimer's disease	<0.001	0.006	161
"05016"	Huntington's disease	<0.001	0.01	176
"04146"	Peroxisome	0.0011	0.02	77
"04010"	MAPK signaling pathway	0.002	0.026	264

Table 5.5. Significant KEGG pathways featured in GSA of tumour tissue compared between study groups. KEGG_id = KEGG pathway identification number. Pathway size = the number of genes in the pathway. p-value adjusted using the Benjamini and Hochberg FDR correction method. Many of the pathways appear significant based on their adjusted p-values but a number of the featured pathways probably arise due to having genes in common with 'genuinely' significant pathways. For example, the neurodegenerative diseases have many genes in common with the Oxidative Phosphorylation pathway.

The Fatty Acid Metabolism pathway featured in the present analysis as for analysis in the normal tissues (section 5.3). The pathway appears more strongly down-regulated here based on the enrichment curve above the plotted t-statistics of the pathway barcode plot (Figure 5.4). This may indicate that the Fatty Acid Metabolism pathway may be more down-regulated in the tumour tissue of the treated participants compared to the untreated group. The 'Metabolic Pathways' pathway also featured in the previous analysis. The majority of the remaining pathways comprise metabolic and neurological disease pathways.

A difference in the level of 'Fatty Acid Metabolism pathway' metagene (section 4.2.8) expression was seen in the tumour tissue between the groups with a p-value of 0.03 (Figure 5.4). Correction for multiple testing was not performed in the metagene analysis thus the significance threshold of 0.05 is fairly liberal. The down-regulation of the Fatty Acid Metabolism pathway metagene expression in the treatment arm compared with the untreated group signifies that the gene expression pattern of the pathway genes was overall consistent in the tumour tissue across the patients of the treatment group.

A heatmap of the Fatty Acid Metabolism pathway genes across the treatment and control groups was generated to visualise the expression of the pathway genes (Figure 5.5). The colour green in the heatmap depicts inhibited expression of genes. The black colouration in the 'Treatment/Control' colour bar indicates the treatment group. The overarching impression in the treated group is that the pathway genes were predominantly down-regulated. For some tumour tissue however, this did not appear to be the case. In five of the tumour tissue samples, more of the pathway genes were up-regulated. Examining the clinical data for the corresponding five patients did not provide any information on why these tumour samples might be responding differently to vitamin D. Also the metagene is highly expressed in the group on the right of the heatmap, indicated by the red colouring in the metagene colour bar. In this grouping the group of genes in the bottom section are up-regulated (also indicated by red) and the balance of the genes are down-regulated. For the grouping on the left side of the heatmap the metagene is down-regulated, indicated by the green colouring in the metagene colour bar. Gene expression in that grouping is clear-cut and appears to be distinctly different from the neighbouring group.

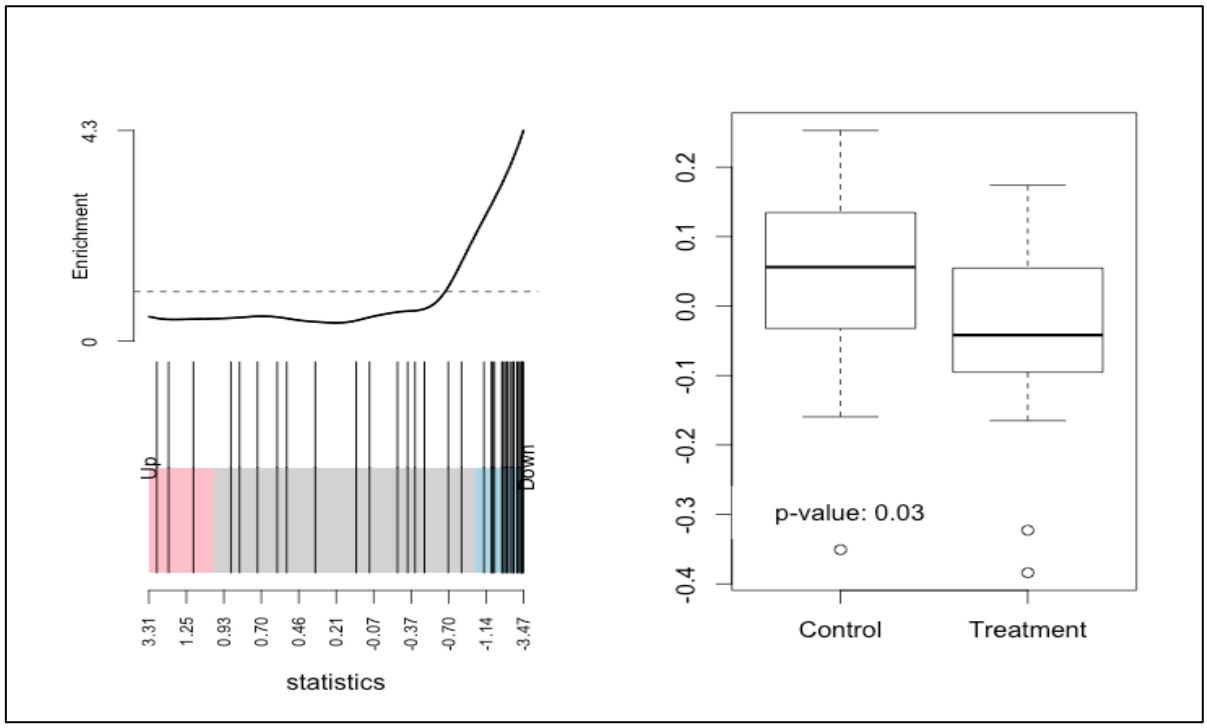


Figure 5.4. Barcode plot for the Fatty Acid Metabolism pathway and boxplot of 'Fatty Acid Metabolism pathway' metagene expression. The barcode plot presents the t-statistics for the genes of the Fatty Acid Metabolism pathway. The enrichment curve is the local regression of the pathway t-statistics. The pathway has 40 genes. Many of the genes for the pathway are down-regulated. The pathway overall appears down-regulated in the tumour tissue of the treated patients. The boxplot presents the level of the pathway metagene in the tumour tissue compared between the treatment and control groups. The expression of the metagene is lower in the treated group. No correction for multiple testing was performed in the metagene analysis, thus the significance threshold of 0.05 is fairly liberal.

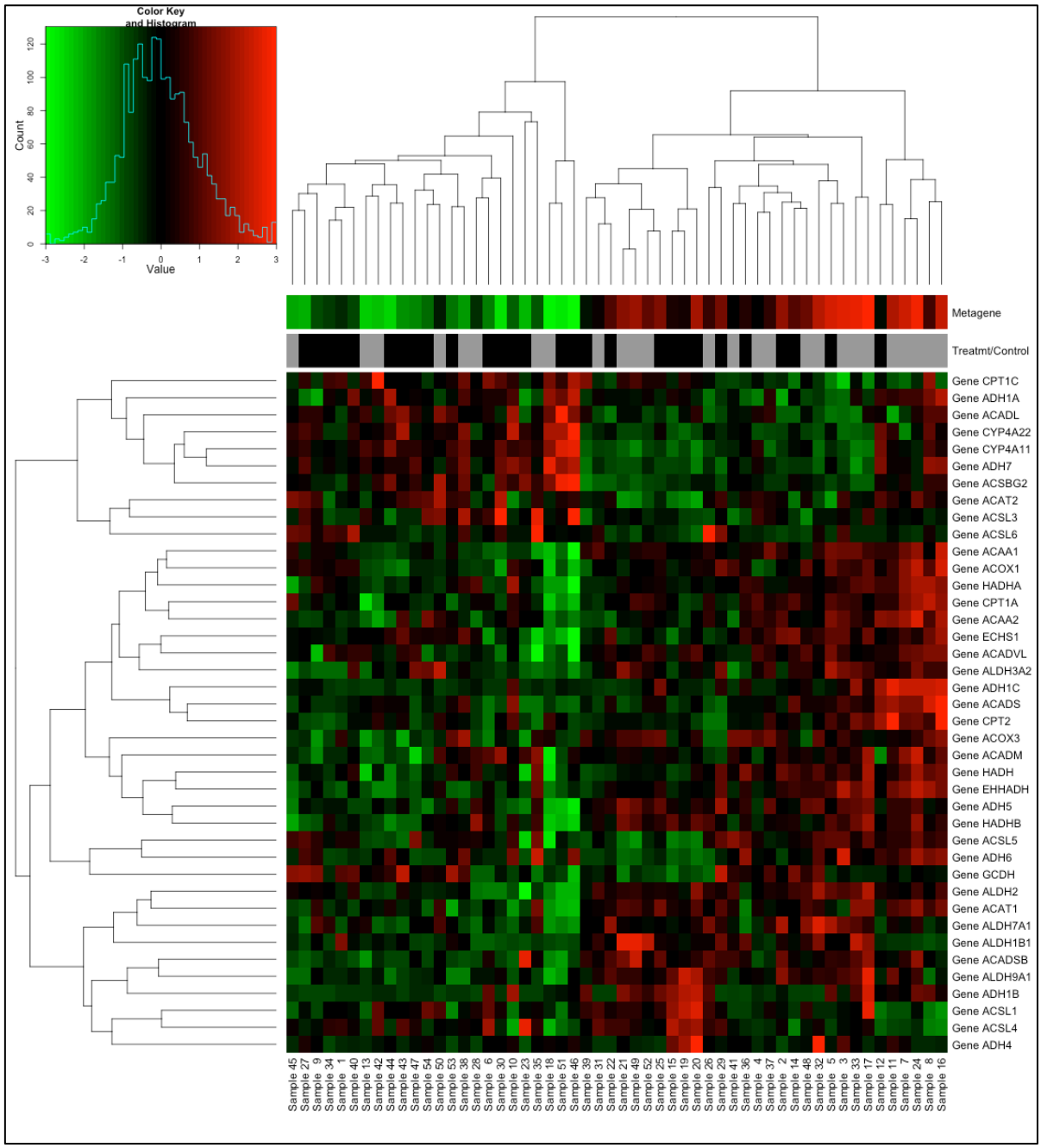


Figure 5.5. Heatmap of the Fatty Acid Metabolism pathway genes across the normal and tumour samples. The black colouration in the 'Treatment/Control' colour bar represents the treatment group. The overall indication is of down-regulation of the pathway genes in the tumour tissue of the treatment arm compared to the placebo arm.

In KEGG the Oxidative Phosphorylation pathway is a member of the energy metabolism pathways subset. The energy metabolism pathways subset belongs to the conglomerate of metabolic pathways designated as a metabolism map. Oxidative phosphorylation is the process by which the energy released by oxidation of nutrients in the mitochondria is used to reform ATP. Although oxidative phosphorylation is an essential part of metabolism, it produces reactive oxygen species causing cell damage and contributes to

disease. Energy metabolism is frequently de-regulated in cancer (Hanahan and Weinberg 2011). The Oxidative Phosphorylation pathway appeared to be down-regulated in the tumour tissue of the treatment group compared with the control group (Figure 5.6).

There was no difference in the level of the Oxidative Phosphorylation pathway metagene expression in the tumour tissue between study groups signifying that the pathway genes did not undergo consistent gene expression in the tumour tissue of the treated patients (Figure 5.6).

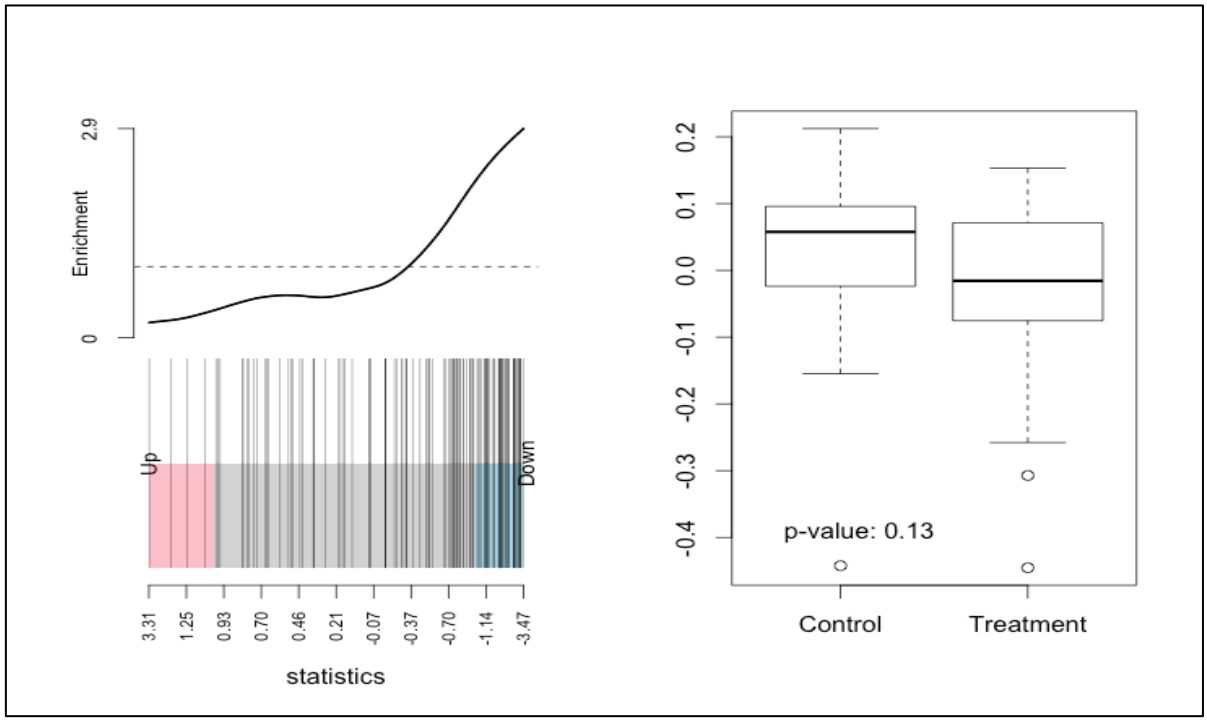


Figure 5.6. Barcode plot for the Oxidative Phosphorylation pathway and boxplot of 'Oxidative Phosphorylation pathway' metagene expression. The barcode plot presents the t-statistics for the genes of the Oxidative Phosphorylation pathway. The enrichment curve is the local regression of the pathway t-statistics. The pathway has 120 genes. Many of the genes for the pathway are down-regulated. The pathway overall appears down-regulated in the tumour tissue of the treated arm. The boxplot presents the level of the metagene in the tumour tissue compared between the two groups. There is no difference in metagene expression between the treatment and control arms.

5.4.4 *Gene-set analysis - GO*

GSA was performed for the tumour tissue using the geneSetTest in limma and data from the GO project. Significant pathways from the analysis are presented in Table 5.6. An adjusted p-value of 0.05 was chosen as the cut-off for establishing pathway/process significance. Nine pathways were identified with significant enrichment of highly ranked genes. The Fatty Acid beta-Oxidation pathway (33 genes) shares genes with the Peroxisome pathway (88 genes, 11 shared). The peroxisome contains peroxidases involved in a range of metabolic processes including lipid catabolism (fatty acids are a subgroup of the lipid category). This Peroxisome pathway is a cellular component pathway and thus is engaged in the formation of the peroxisome not its function. The pathway may feature here because of its shared genes with the Fatty Acid beta-Oxidation pathway. There are 78 shared genes between the Mitochondrial Inner Membrane (295 genes) and Respiratory Electron Transport Chain (87 genes) pathways. Barcode plots for the Fatty Acid beta-Oxidation and Respiratory Electron Transport Chain pathways can be seen in Figures 5.7 and 5.9. Boxplots depicting metagene expression level in the groups can be seen alongside. A barcode plot was not included for the Mitochondrial Inner Membrane pathway as it contains too many genes (295 genes) for good visualisation of the expression data. However, it was possible to discern that the pathway was down-regulated overall.

GO id	GO Term	p-value	Adjusted p-value (fdr)	Pathway size
GO:0006635	Fatty acid beta-oxidation	<0.001	<0.001	33
GO:0005743	Mitochondrial inner membrane	<0.001	<0.001	290
GO:0022904	Respiratory electron transport chain	<0.001	0.007	86
GO:0005777	Peroxisome	<0.001	0.008	88
GO:0044237	Cellular metabolic process	<0.001	0.017	119
GO:0006120	Mitochondrial electron transport, NADH to ubiquinone	<0.001	0.02	38
GO:0044281	Small molecule metabolic process	<0.001	0.028	1356
GO:0008150	Biological process	<0.001	0.032	547
GO:0007269	Neurotransmitter secretion	<0.001	0.048	53

Table 5.6. Significant GO terms featured in GSA of tumour tissue compared between study groups. GO_id = GO term identification number. Pathway size = the number of genes in the pathway. p-value adjusted using the Benjamini and Hochberg FDR correction method. The top four pathways are considered to be the most significant based on their adjusted p-values.

Beta-oxidation begins once fatty acids have been imported into the mitochondrial matrix. It is the complete oxidation of a fatty acid molecule and yields ATP molecules. VDUP1 has been proposed as a regulator of fatty acid utilisation (Kim et al. 2007). The Fatty Acid beta-Oxidation pathway was down-regulated in the tumour tissue in the treated group compared with the untreated group (Figure 5.7).

A difference in the level of the 'Fatty Acid beta-Oxidation pathway' metagene expression was seen in the tumour tissue between the treated and untreated participants with a p-value of 0.028 (Figure 5.7). The metagene analysis was not corrected for multiple testing and thus a significance threshold of 0.05 is liberal. Down-regulation of the Fatty Acid beta-Oxidation pathway metagene in the treated group compared with the control group indicates that the pattern of gene expression of the pathway genes was relatively consistent in the tumour tissue of the treated patients.

A heatmap of the Fatty Acid beta-Oxidation pathway genes across the treated and untreated patients was rendered to examine the level of expression of the pathway genes (Figure 5.8). The pathway gene expression appears to be reduced for the majority of the genes in the tumour tissue of the treated patients compared with the untreated group. Seven of the tumour tissue samples exhibit a different pattern of gene expression where the expression of the majority of the pathway genes is enhanced. Five of these samples showed a different pattern of gene expression for the Fatty Acid Metabolism pathway (featured in the tumour tissue when using data from the KEGG database, Section 5.4.3), where the majority of their genes were also up-regulated. Inspection of clinical data was not elucidating regarding this finding.

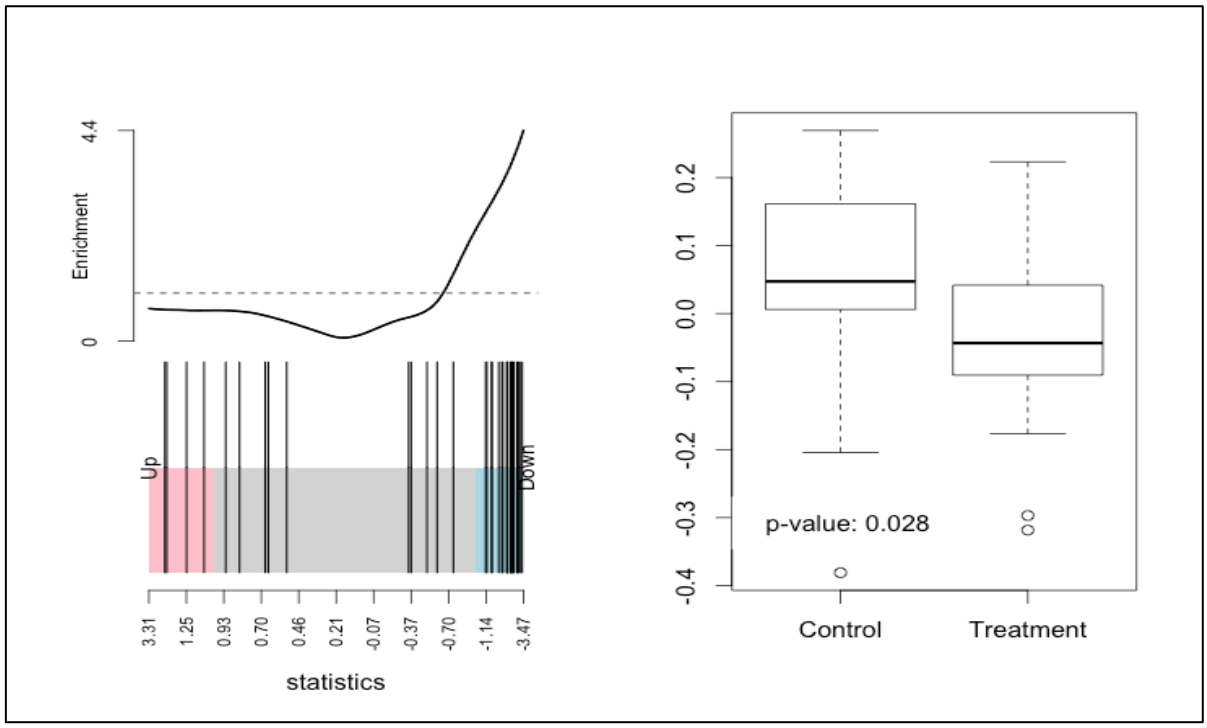


Figure 5.7. Barcode plot for the Fatty Acid beta-Oxidation pathway and boxplot of 'Fatty Acid beta-Oxidation pathway' metagene expression. The barcode plot presents the t-statistics for the genes of the Fatty Acid beta-Oxidation pathway. The enrichment curve is the local regression of the pathway t-statistics. The pathway has 33 genes. Many of the genes for the pathway are down-regulated. The pathway overall appears down-regulated in the tumour tissue of the treatment arm. The boxplot presents the level of the pathway metagene in the tumour tissue compared between the study groups. There is a difference in metagene expression between the trial groups. The metagene analysis was not corrected for multiple testing, thus the significance threshold of 0.05 is fairly liberal.

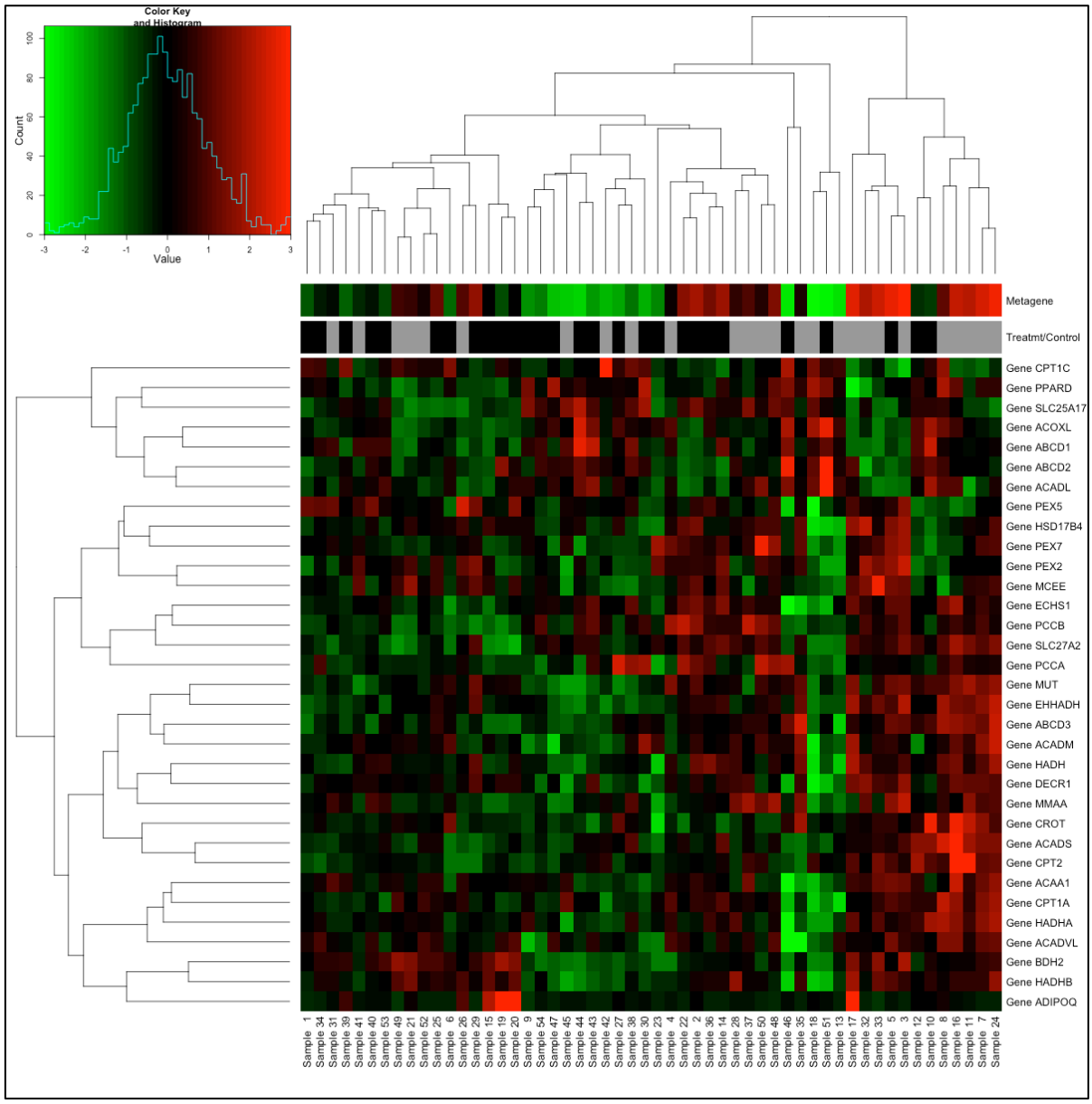


Figure 5.8. Heatmap of the Fatty Acid beta-Oxidation pathway genes across the normal and tumour samples. The treatment group is represented by black in the colour bar. Overall the pathway genes appear to be down-regulated in the tumour tissue of the treated group.

The respiratory electron transport chain comprises a series of protein complexes, embedded within the inner mitochondrial membrane. The chain is charged with setting up an electro-chemical proton gradient across the membrane. This is achieved when electrons are transferred between donors and acceptors, coupled with proton transfer across the membrane. The electro-chemical proton gradient is used to power the production of ATP from adenosine diphosphate (ADP) and inorganic phosphate (Pi). De-regulation of energy metabolism often occurs in cancer (Hanahan and Weinberg 2011). The Respiratory Electron Transport Chain pathway was down-regulated in the tumour

tissue of the treatment arm compared with the placebo arm (Figure 5.9).

There was no difference in the level of 'Respiratory Electron Transport Chain pathway' metagene expression in the tumour tissue when comparing the groups signifying that the pattern of pathway gene expression was not consistent across the tumour tissue of the treatment group (Figure 5.9).

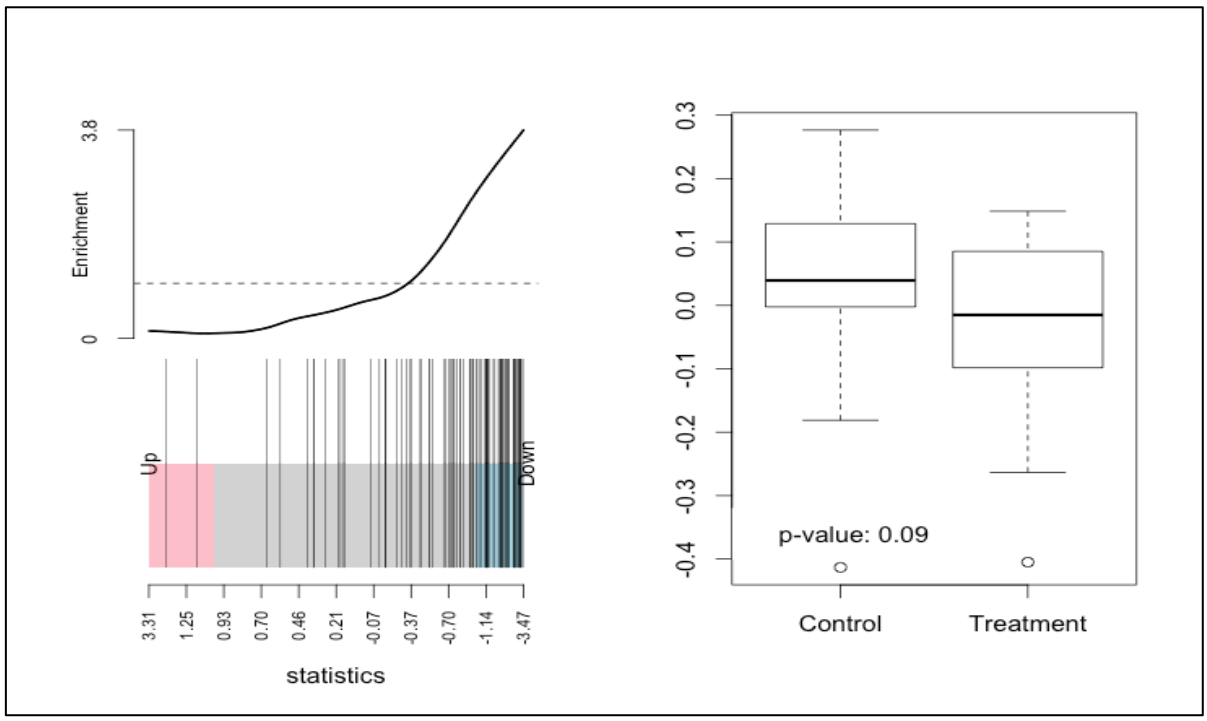


Figure 5.9. Barcode plot for the Respiratory Electron Transport Chain pathway and boxplot of 'Respiratory Electron Transport Chain pathway' metagene expression. The barcode plot presents the t-statistics for the genes of the Respiratory Electron Transport Chain pathway. The enrichment curve is the local regression of the pathway t-statistics. The pathway has 87 genes. Most of the genes for the pathway are down-regulated. The pathway overall appears down-regulated in the tumour tissue of the treatment group. The boxplot presents the level of the metagene in the tumour tissue compared between arms of the study. There is no difference in metagene expression between the groups.

Despite the barcode plot not being a helpful visual aid for the Mitochondrial Inner Membrane pathway (too many genes at 295), it was possible to discern that the pathway was down-regulated overall. The mitochondrial inner membrane is involved in the oxidative phosphorylation process. This process is described previously in this section, where the respiratory electron transport chain is discussed. In GO, this pathway is defined as a cellular component pathway and is thus concerned with the construction of the inner mitochondrial membrane.

In the tumour tissue differences in gene expression are evident between the control and treatment arms of the study. Metabolism pathways featured in the various database enrichment analyses and in GSA performed using limma. The activity of several energy metabolism pathways were demonstrated to be down-regulated in the tumour tissue of the treated arm compared with the untreated arm. Pathway metagene expression for two of the pathways, Fatty Acid Metabolism and Fatty Acid beta-Oxidation, showed a relatively consistent difference when compared between the study groups.

The difference in expression of the metagene may be due to either experimental conditions or clinical factors. However, as this study is a randomised controlled trial, the clinical factors are expected to be balanced across the treatment and control arms of the trial. The baseline demographics show this to be the case (section 3.3.1, 3.3.3 and 3.3.4). Establishing the quality of the expression data using the arrayQualityMetrics package (section 4.2.2) and normalising the data using the RMA method (section 4.2.3), ensured that the data used in these analyses were of sound quality and that experimental artefacts were removed. Thus the significant down-regulation of both the Fatty Acid Metabolism and Fatty Acid beta-Oxidation pathways observed in the tumour tissue of the treated group can be viewed as an authentic result. The fact that multiple testing was not corrected for in the metagene analysis needs to be borne in mind. The result indicates that the pattern of pathway gene expression was overall consistent in the tumour tissue of the treated patients compared with the untreated group.

The Fatty Acid Metabolism pathway was also down-regulated in the normal tissue of the treated group compared with the untreated group. This result might suggest that normal vitamin D activity in the normal colonic tissue, when vitamin D levels are at proposed optimal levels, includes modulation of energy metabolism. The tissue was collected while the patient was undergoing surgery and thus the bowel would not have been functioning as usual. There would have been lessened need for energy expenditure in the colon and perhaps time for the down-regulation of energy metabolism under these circumstances.

The results from the tumour tissue analysis did not shed any light on the other pathways featured in the normal tissue analysis. Perhaps these other pathways were also down-

regulated during surgery. Otherwise, the down-regulation of the xenobiotic metabolism, vitamin digestion and absorption and negative regulation of growth pathways observed in the normal tissue analysis may be baseline indicators only. This may be a useful picture of normal vitamin D activity in the colonic cells when vitamin D is at recommended optimal levels. Vitamin D may regulate xenobiotic metabolism, vitamin digestion and absorption and negative regulation of growth pathways on a 'business as usual' basis. Presumably the xenobiotic and vitamin pathway would be up-regulated by vitamin D if the need arose. For example, an increase in foreign material in the bowel might signal, through vitamin D, an enhanced expression of the xenobiotic pathway. Similarly, as a matter of course, vitamin D may inhibit the negative modulation of growth pathway and thus enable cell growth under normal conditions and depending on other conditions in the bowel.

The fact that the Fatty Acid Metabolism and Fatty Acid beta-Oxidation pathway metagene expression was different between the groups is a potentially useful finding as it corroborates the apparent down-regulation of these pathways in the tumour tissue and will be discussed further in the chapter discussion.

5.5 Paired analysis - control versus treatment

Paired normal and tumour tissue was collected from 51 patients. Twenty-five of the patients were in the control arm and 26 were in the treatment arm of the study. RNA from each normal and tumour tissue sample was profiled on a separate microarray gene-chip.

Investigating differences in gene expression between the paired normal and tumour tissue of participants allows the opportunity to determine whether any alterations in expression are consistently observed across the cohort. If consistent changes are observed across the entire group, it could be that the differences are due to the impact of vitamin D in the tissue.

As a basis for comparison with the paired normal and tumour tissue analysis, differences in gene expression were examined across the paired tissue without incorporating the effect of vitamin D (sections 5.5.1 and 5.5.2).

5.5.1 *Paired analysis - without group comparison - KEGG*

The paired normal and tumour tissue was analysed for pathways exhibiting enrichment of highly ranked genes utilising the limma package and KEGG data. Results from this analysis are presented in Table 5.7. Pathway significance was determined with an FDR adjusted p-value of 0.05. Thirteen significant pathways were identified. The majority of the most significant pathways are translation pathways. As these pathways were only required as a comparison for the paired analysis across the treatment and control groups, visualisation plots were not employed here.

KEGG id	KEGG Pathway	p-value	Adjusted p-value (fdr)	Pathway size
"03010"	Ribosome	<0.001	<0.001	86
"03050"	Proteasome	<0.001	<0.001	44
"03008"	Ribosome biogenesis in eukaryotes	<0.001	<0.001	71
"00970"	Aminoacyl-tRNA biosynthesis	<0.001	<0.001	41
"03013"	RNA transport	<0.001	<0.001	142
"03018"	RNA degradation	<0.001	0.005	70
"00670"	One carbon pool by folate	<0.001	0.007	18
"00230"	Purine metabolism	<0.001	0.01	152
"04512"	ECM-receptor interaction	<0.001	0.018	84
"04110"	Cell cycle	0.002	0.037	123
"03030"	DNA replication	0.002	0.039	36
"03020"	RNA polymerase	0.003	0.046	27
"04976"	Bile secretion	0.003	0.047	69

Table 5.7. Significant KEGG pathways featured in GSA of the paired analysis WITHOUT comparison between study groups. KEGG_id = KEGG pathway identification number. Pathway size = the number of genes in the pathway. p-value adjusted using the Benjamini and Hochberg FDR correction method. The most significant pathways in the table are primarily translation pathways.

5.5.2 *Paired analysis - without group comparison - GO*

The paired normal and tumour tissue was analysed for pathways significantly enriched for highly ranked genes utilising the limma package and GO data. Pathway significance was determined with an FDR adjusted p-value of 0.05. Ninety-four significant pathways were produced, 52 of them with an adjusted p-value of <0.001. The range of activity of the 52 pathways included mainly translation and cell cycle pathways with a small number of immune response pathways featured also. The top fifteen significant

pathways from this analysis are presented in Table 5.8. As these pathways were only required as a comparison for the paired analysis, visualisation plots were not generated here.

GO id	GO Term	p-value	Adjusted p-value (fdr)	Pathway size
GO:0016070	RNA metabolic process	<0.001	<0.001	240
GO:0016071	mRNA metabolic process	<0.001	<0.001	218
GO:0010467	Gene expression	<0.001	<0.001	649
GO:0006412	Translation	<0.001	<0.001	235
GO:0000278	Mitotic cell cycle	<0.001	<0.001	382
GO:0003735	Structural constituent of ribosome	<0.001	<0.001	143
GO:0019083	Viral transcription	<0.001	<0.001	80
GO:0006614	SRP-dependent cotranslational protein	<0.001	<0.001	105
GO:0006413	Translational initiation	<0.001	<0.001	129
GO:0006415	Translational termination	<0.001	<0.001	86
GO:0044822	Poly(A) RNA binding	<0.001	<0.001	1031
GO:0006414	Translational elongation	<0.001	<0.001	92
GO:0031145	Anaphase-promoting complex-dependent*	<0.001	<0.001	79
GO:0000184	Nuclear-transcribed mRNA catabolic process	<0.001	<0.001	110
GO:0051439	Regulation of ubiquitin-protein ligase activity#	<0.001	<0.001	75

Table 5.8. Significant GO terms featured in GSA of the paired analysis WITHOUT comparison between study groups. GO_id = GO term identification number. Pathway size = the number of genes in the pathway. *Full pathway name= Anaphase-promoting Complex-dependent Proteasomal Ubiquitin-dependent Protein Catabolic Process. #Full pathway name = Regulation of Ubiquitin-protein Ligase Activity involved in Mitotic Cell Cycle. p-value adjusted using the Benjamini and Hochberg FDR correction method. This is a subset (top 15) of the 94 significant pathways that were produced in this analysis, 52 of which had an adjusted p-value of <0.001.

In the paired analysis performed independent of vitamin D, many pathways were identified with significant differences in expression between the normal and tumour tissue of the pairings. Multiple significant differences were anticipated between the normal and tumour tissue thus this finding was not unexpected. The differences in pathway expression were consistent across the patients.

5.5.3 *Single gene analysis - paired analysis*

The paired normal and tumour tissue was analysed for individual gene expression using the limma package. Comparisons were performed between the treatment and control

groups. Ten of the top differentially expressed genes are presented in a statistical summary table (Table 5.9). No genes were found to have significant expression changes after adjustment for multiple testing (Benjamini and Hochberg 1995). All genes displayed adjusted p-values above the significance threshold of 0.05.

Gene	Log fold change	Average expr	t statistic	p-value	Adjusted p-value	B statistic
SLC43A2	-0.3	-0.05	-4.04	<0.001	0.95	-0.94
LMF1	-0.2	0.04	-3.7	<0.001	0.95	-1.51
EXOC6B	-0.19	0.19	-3.38	<0.01	0.95	-2.01
N4BP3	-0.16	-0.03	-3.22	<0.01	0.95	-2.26
LRRC42	0.21	-0.04	3.13	<0.01	0.95	-2.39
KCND3	-0.17	-0.01	-3.05	<0.01	0.95	-2.51
CNTROB	-0.17	-0.04	-3.05	<0.01	0.95	-2.52
SPTA1	-0.14	-0.04	-2.98	<0.01	0.95	-2.61
C21orf56	-0.25	-0.08	-2.96	<0.01	0.95	-2.65
TMCO4	0.18	0.02	2.87	<0.01	0.95	-2.78

Table 5.9. Linear model results summary table for the paired analysis. Summary statistical information derived from analysis of gene expression differences in the paired analysis between the treatment and control arms. The table shows summary information for the top 10 genes (based on raw p-values). After adjustment for multiple testing, no genes demonstrated an adjusted p-value below the significance threshold of 0.05.

A heatmap of the top 100 DEG from the paired analysis comparing the study groups can be viewed in Appendix 6 (Figure 4C).

The VDRE gene list was examined for differential expression. The VDRE gene list comprises over 3130 genes. P-values for these genes were taken from the linear model results summary table. They were adjusted for multiple testing using the number of genes in the VDRE gene list. None of the adjusted p-values were significant with a range of 0.96-1, suggesting that the observed expression changes could have happened by chance.

5.5.4 *Over-representation analysis - paired analysis*

The Wikipathways database identified 98 pathways exhibiting over-representation of genes in treated patients compared with untreated patients, when the top 100 DEG were uploaded to the database. Pathway activity ranged from MAPK, WNT signaling and regulation of DNA replication through fatty acid biosynthesis, the metabolism of steroid hormones and vitamin D and the vitamin D receptor pathway to cell differentiation, apoptosis modulation and signaling and miRNA regulation of DNA damage response. This is a hand-picked list of nine pathways from the overall 98 returned by the Wikipathways database. Being presented with so many pathways without significance information is not very useful.

A gene list of the 100 most highly DEG, from the single gene analysis in paired normal and tumour tissue, was presented to the GeneSetDB database. This process returned the MAPK Signaling Pathway (FDR <0.01) with enrichment of genes in the treated group compared with the untreated group when the 'pathway' database was selected along with a cut-off FDR adjusted p-value of 0.05. MAPK signaling pathways moderate diverse cellular functions including cell proliferation, differentiation, apoptosis and cell migration. Vitamin D has been reported to have roles in cell proliferation, differentiation and apoptosis (sections 1.3.4.2 and 1.3.4.2a-e). Though there is a difference in the expression of these pathways between the trial groups, the nature of that difference cannot be ascertained. The MAPK Signaling Pathway also featured in the over-representation analysis and GSA (using KEGG data) in the tumour tissue.

Six pathways were identified with over-representation of genes in the treated versus the untreated group when the 100 most highly DEG were submitted to the Reactome database. Most of the pathways had an FDR adjusted p-value of <0.001. All of the pathways were involved in the cellular response to stress processes. When the top 50 DEG were uploaded, seven significant pathways were identified, again the majority involved in the cellular response to stress processes and with FDRs ranging from <0.001 to <0.05.

Submitting the top 50 or 100 most DEG to the GATHER tool did not produce significant pathways.

5.5.5 *Gene-set analysis - paired analysis - KEGG*

The paired tissue was analysed for pathways significantly enriched for highly ranked genes using the limma package and KEGG data. Comparisons were made between the treatment and control groups. The significant results from this analysis are presented in Table 5.10. Two pathways with significant gene enrichment were identified. Pathway significance was determined with an FDR adjusted p-value of 0.05. A barcode plot for the Ribosome KEGG pathway can be seen in Figure 5.10. A boxplot showing Ribosome pathway metagene expression is presented in the figure also.

KEGG id	KEGG Pathway	p-value	Adjusted p-value (fdr)	Pathway size
"03010"	Ribosome	<0.001	0.0012	86
"04270	Vascular smooth muscle contraction	<0.001	0.035	113

Table 5.10. Significant KEGG pathways featured in GSA of the paired analysis compared between the treatment and control groups. KEGG_id = KEGG pathway identification number. Pathway size = the number of genes in the pathway. p-value adjusted using the Benjamini and Hochberg FDR correction method.

KEGG categorises the Ribosome pathway as a translation pathway, a subset of the overarching genetic information processing group of pathways. The ribosome comprises two main components, a small subunit which translates the messenger RNA (mRNA) and a large subunit, which connects the amino acids to produce a polypeptide chain/protein. The data used to plot the barcode plot is from the *difference matrix* (tumour gene expression values subtracted from normal gene expression values; section 4.2.5.3) used in the paired analysis. Thus the barcode plot cannot be read at face value to extract an indication of the directionality of gene and pathway expression. A heatmap was generated to obtain an indication of the direction of gene and pathway expression (Figure 5.11). The heatmap depicts a separation of the pathway gene expression differences into two groups where they tend to be negative (green colouring in the heatmap) in one group and positive (red) in the other. The negative pathway gene expression differences indicate lower levels of expression in the normal tissue compared to the tumour tissue. Likewise, the positive differences signify higher levels of expression in the normal when compared to the tumour. The grouping into these two categories is not a division based strictly on the treatment and placebo groups. The

positive and negative pathway gene expression difference subgroups in the heatmap were checked against the clinical data but no associations were found.

There was no difference in the level of 'Ribosome pathway' metagene expression in the paired analysis when the treatment and control groups were compared (Figure 5.10). This signifies that the pattern of pathway gene expression was not consistent across the patients in the treated group compared to the untreated group.

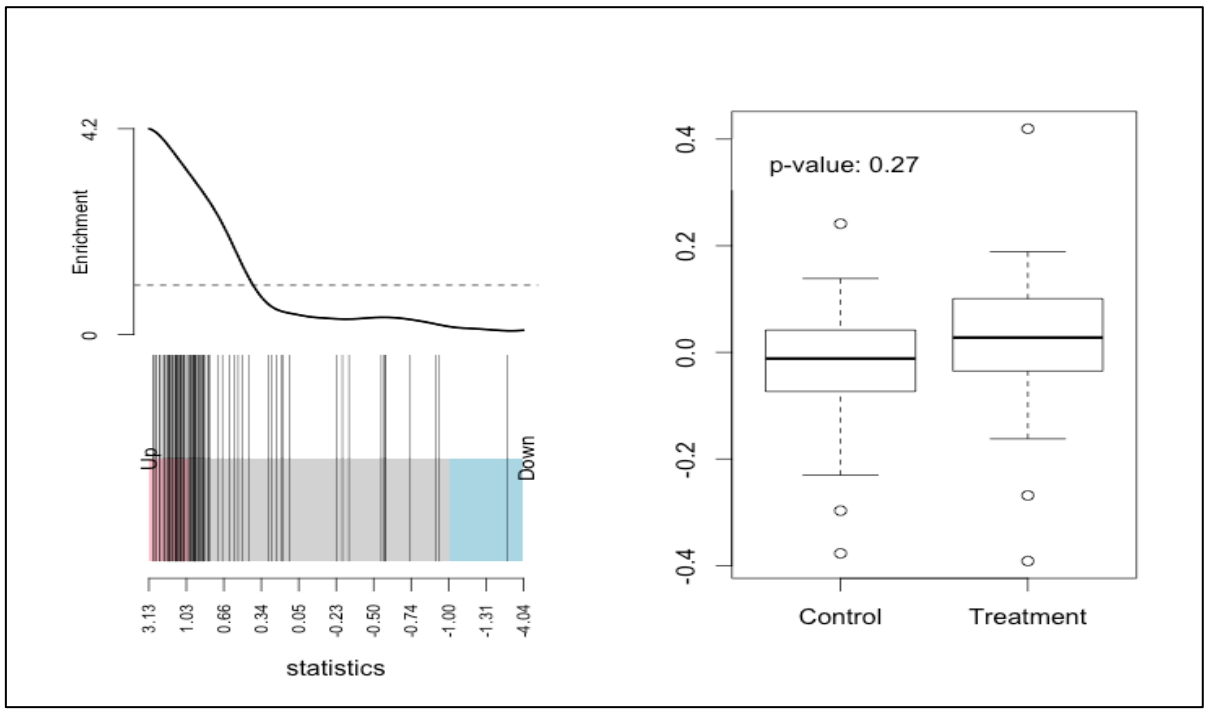


Figure 5.10. Barcode plot for the Ribosome pathway and boxplot of 'Ribosome pathway' metagene expression. The barcode plot presents the t-statistics for the genes of the Ribosome pathway. The enrichment curve is the local regression of the pathway t-statistics. The pathway has 86 genes. The majority of the genes in the pathway have undergone a change in expression. The pathway overall appears up-regulated but this is an erroneous inference in the context of paired analysis. A heatmap was generated to examine gene expression directionality. The boxplot presents the level of the metagene in the paired analysis compared between the control and treatment arms of the study. There is no difference in metagene expression between the groups.

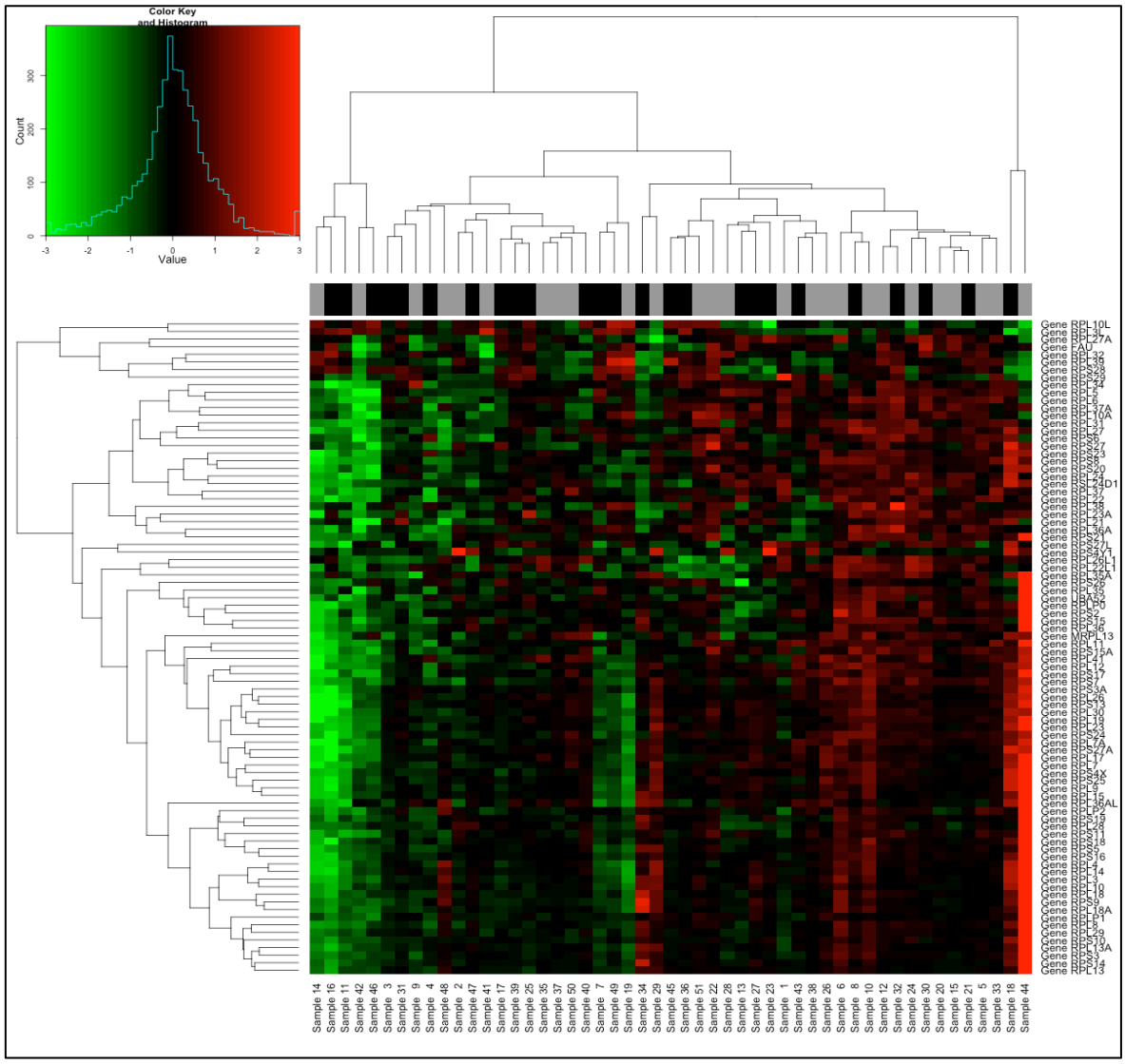


Figure 5.11. Heatmap of Ribosome pathway standardised gene expression differences in the paired analysis across the treatment and control groups. Black in the colour bar indicates the treated group. The pathway gene expression differences were negative in almost half of the patients, equating to lower levels of expression in the normal compared to the tumour tissue, and positive in the other half, depicting higher levels of expression in the normal compared to the tumour. The division into two subgroups is not strictly related to the treatment and control groups.

The Ribosome pathway featured in the paired analysis without incorporating the effect of treatment with vitamin D. It manifested as a highly significant pathway with an adjusted p-value of <0.001, compared with an adjusted p-value of 0.0012 in the paired analysis with the incorporation of the vitamin D treatment effect. This indicates that the expression of the pathway is significantly different between the normal and tumour tissue, which was expected. Vitamin D also appears to alter the expression of the pathway, causing reduced Ribosome pathway expression in the normal tissue of one subgroup of patients and enhanced pathway expression in the normal tissue of the other

when compared with the tumour tissue or conversely, enhanced expression of the pathway in the tumour tissue of one subgroup and reduced in the other when compared to the normal tissue.

5.5.6 *Gene-set analysis - paired analysis - GO*

The geneSetTest in limma using GO project data was utilised to determine differential expression of pathways in the paired analysis. Pathway expression was compared between the treated and untreated groups. Significant results are displayed in Table 5.11. An adjusted p-value of 0.05 was chosen as the cut-off for establishing pathway/process significance. Thirteen pathways were identified with significant enrichment of highly ranked genes. The mRNA (218 genes) and RNA metabolism (240 genes) pathways share 214 genes. These pathways also share genes with the Nuclear-transcribed mRNA Catabolic Process pathway (112 genes). The intercept between these three pathways is 108 genes. The translation pathways share 80 genes. The RNA and translation pathways also share genes in the region of 80 genes. These pathways having genes in common may signify that some of these pathways feature here because they share genes with an authentically significant pathway. A barcode plot for the Translational Termination pathway can be seen in Figure 5.12. A boxplot depicting metagene expression level in the groups can be seen alongside.

GO id	GO Term	p-value	Adjusted p-value (fdr)	Pathway size
GO:0016071	mRNA metabolic process	<0.001	0.012	218
GO:0005080	Protein kinase C binding	<0.001	0.017	43
GO:0006415	Translational termination	<0.001	0.017	83
GO:0016070	RNA metabolic process	<0.001	0.017	240
GO:0019058	Viral life cycle	<0.001	0.017	113
GO:0019083	Viral transcription	<0.001	0.017	80
GO:0045095	Keratin filament	<0.001	0.017	81
GO:0030054	Cell junction	<0.001	0.017	377
GO:0000502	Proteasome complex	<0.001	0.019	55
GO:0006418	tRNA aminoacylation for protein translation	<0.001	0.019	42
GO:0000184	Nuclear -transcribed mRNA catabolic process	<0.001	0.024	112
GO:0006413	Translational initiation	<0.001	0.024	114
GO:0007268	Synaptic transmission	<0.001	0.024	380

Table 5.11. Significant GO terms featured in GSA of the paired analysis compared between the study groups. GO_id = GO term identification number. Pathway size = the number of genes in the pathway. p-value adjusted using the Benjamini and Hochberg FDR correction method. The presented pathways are not highly significant with adjusted p-values of <0.05.

Translation is the process of mRNA being 'read' into proteins and has three components, initiation, elongation and termination. Two of these translational pathways, initiation and termination, are present amongst the significant pathways returned in the present analysis. Translational termination of protein synthesis and the release of the pre-folded protein are signaled by one of three stop codons and is mediated by protein release factors. The pathway seems to have undergone an alteration in expression.

A heatmap was produced to determine the directionality of gene and pathway expression (Figure 5.13). The pathway gene expression differences tend to be negative (green) in almost half of the patients and positive (red) in the remainder. The negative differences indicate lower levels of expression in the normal tissue compared to the tumour tissue and the positive differences signify higher levels of expression in the normal tissue compared to the tumour (the *difference matrix* was generated by subtracting the tumour gene expression values from the normal values). This separation into two groups is not strictly related to the treatment and control groups. The subgroups in the heatmap were compared against clinical data but no associations were found.

No difference in 'Translational Termination pathway' metagene expression level was observed in the paired analysis when comparisons were made between the treatment and control groups (Figure 5.12). This signifies that the pathway gene expression pattern was not consistent across the patients in the treated group.

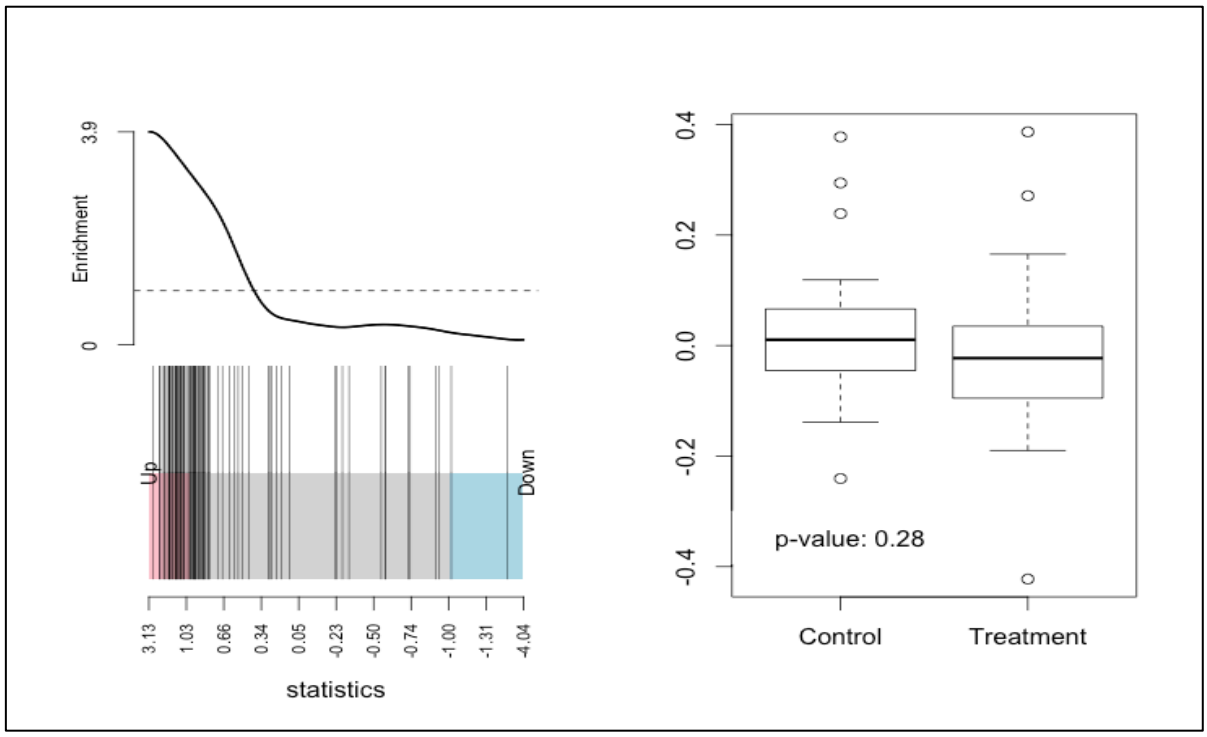


Figure 5.12. Barcode plot for the Translational Termination pathway and boxplot of 'Translational Termination pathway' metagene expression. The barcode plot presents the t-statistics for the genes of the Translational Termination pathway. The enrichment curve is the local regression of the pathway t-statistics. The pathway has 129 genes. Most of the genes for the pathway have undergone a change in expression. The boxplot presents the level of the metagene in the paired analysis compared between the control and treatment groups. There is no difference in metagene expression between the groups.

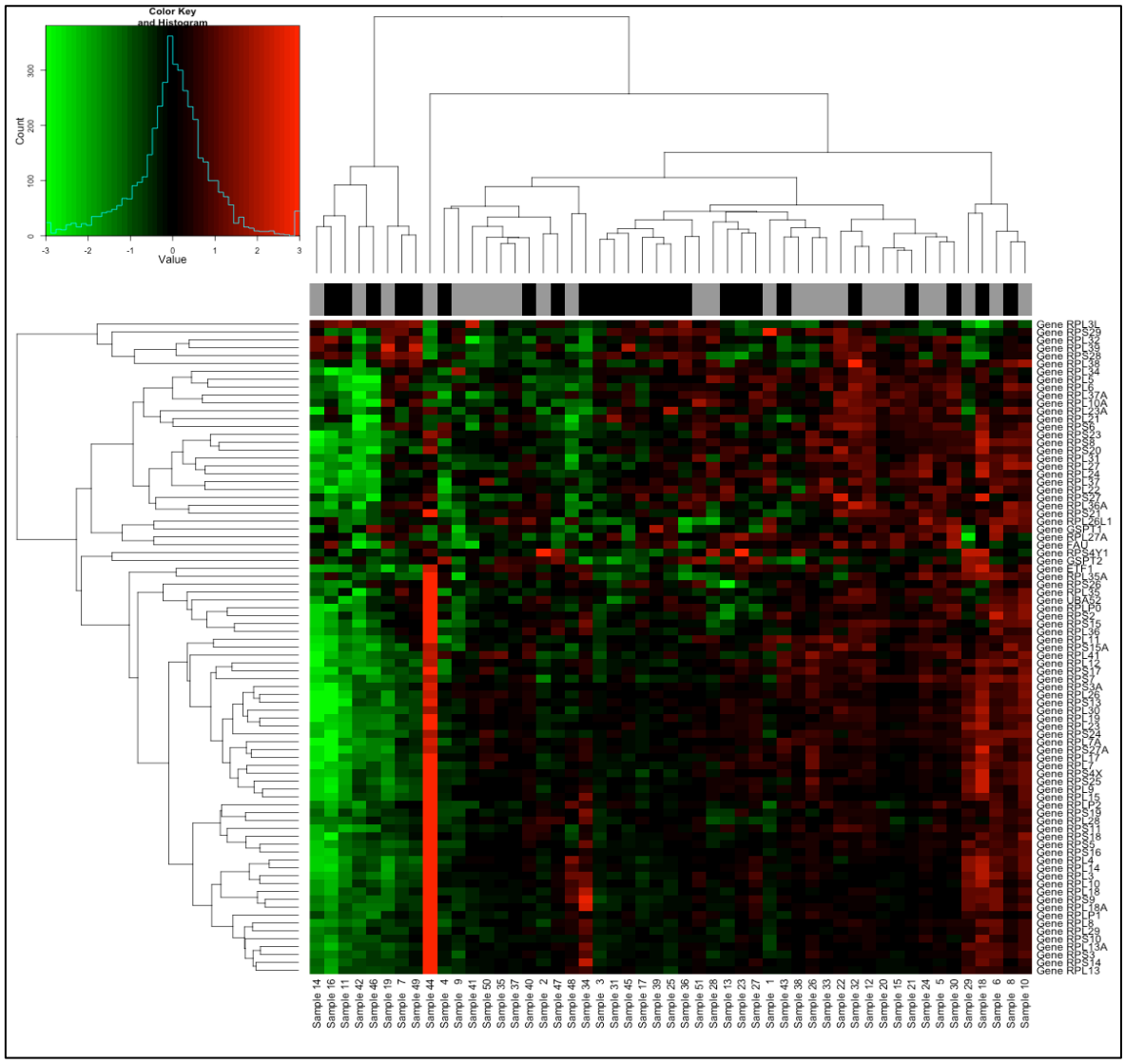


Figure 5.13. Heatmap of Translational Termination pathway standardised gene expression differences in the paired analysis across the treatment and control groups. Black in the colour bar indicates the treated group. The pathway gene expression differences were negative in almost half of the patients, equating to lower levels of expression in the normal compared to the tumour tissue, and positive in the other half, depicting higher levels of expression in the normal compared to the tumour. The division into two subgroups is not strictly related to the study group.

The Translational Termination pathway featured in the paired analysis without inclusion of the vitamin D effect. It had an adjusted p-value of <0.001. The pathway appeared less significant with an adjusted p-value of 0.017 in the paired analysis incorporating vitamin D effect. There are significant differences in expression of the pathway between the normal and tumour tissue as expected. Vitamin D appears to alter the pathway expression, causing reduced Translational Termination pathway expression in the normal tissue of one subgroup and elevated expression in the normal tissue of the other when compared to the tumour tissue. The converse is also true. The

pathway expression is enhanced in the tumour tissue of one subgroup and reduced in the other when compared to the normal tissue.

Over-representation analysis identified pathways with gene enrichment including the MAPK Signaling Pathway and cellular response to stress pathways in the treatment group compared with the placebo group. These pathways did not feature in GSA using KEGG and GO data. Paired analysis performed by GSA identified primarily translation pathways with differences of expression between the treatment and placebo groups.

Heatmaps generated for the Ribosome and Translational Termination pathway gene expression differences depicted separation into two subgroups. The negative gene expression differences relate to lower levels of expression in the normal tissue (and higher levels of expression in the tumour) when compared to the tumour tissue, while the positive gene expression differences relate to higher levels of expression in the tumour tissue (and lower expression levels in the normal) when compared to the normal tissue. Vitamin D appears to reduce Ribosome and Translational Termination pathway gene expression in the normal tissue of one subgroup and to enhance the pathway expression in the other when compared to the tumour tissue. These subgroups were not strictly related to the treatment and control groups or other clinical data. Translation pathway metagene expression between the groups did not yield any significant results suggesting that the gene expression patterns of the translation pathways were not being consistently expressed in the patients of the treatment arm compared with the placebo arm.

5.6 Discussion

This study was a randomised, double-blind, placebo-controlled trial of vitamin D supplementation in patients undergoing elective surgery for CRC at Dunedin Hospital. The primary aim of this study was to determine whether a single large dose of vitamin D administered to CRC patients in the window between diagnosis and surgery could have a measurable impact on vitamin D responsive genes in the colorectal tumour and corresponding normal colonic tissue. Serum calcidiol concentrations were elevated to above recommended optimal levels in the treatment group, with a mean of 87 nmol/L compared to the placebo group, with a mean of 49 nmol/L. The post-treatment mean

calcidiol levels for both groups may be within the 'normal range' (section 1.2.5). However, the treatment group underwent a 60% increase in mean calcidiol serum level. This difference may have led to observable changes in the expression of vitamin D responsive genes in the tumour and corresponding normal tissue.

In single gene analyses performed in the normal, tumour and paired analyses, no individual genes were identified as being differentially expressed between the treatment and control arms. This was also true for the VDRE gene list. Gene-set and over-representation analyses were performed to identify pathways with differences of expression between the study groups in the normal, tumour and paired analyses. In the normal tissue treated patients had down-regulation of a number of pathways compared to untreated patients including the Fatty Acid Metabolism, Drug Metabolism-Cytochrome P450, Metabolism of Xenobiotics by Cytochrome P450, Vitamin Digestion and Absorption and Negative Regulation of Growth pathways. In the tumour tissue treated patients also had down-regulation of several pathways compared to untreated patients including the Fatty Acid Metabolism, Fatty Acid beta-Oxidation and Oxidative Phosphorylation pathways. In the paired analysis, the expression of the Ribosome and Translational Termination pathways were enhanced by vitamin D in the tumour tissue of a subgroup of patients and reduced in another subgroup, compared to normal tissue. Over-representation analyses identified a number of other pathways that may be different between the study groups.

Individual gene expression analysis was performed for normal, tumour and paired analyses using the limma package, which utilises a linear models method in combination with an empirical Bayes modified t-statistic. After adjustment for multiple hypothesis testing, no genes were found to be significantly differentially expressed in any of the single gene analyses.

Upon examination of the VDRE gene list for differentially expressed genes, none were found. The observation of differentially expressed VDRE genes would depend on the infiltration of vitamin D into the tumour tissue. Results from GSA in the tumour tissue may indicate that vitamin D given to the treatment group did infiltrate the tumour but the fact that genes from the VDRE list of over 3,000 genes were not differentially expressed induces some doubt. The VDRE gene list is not exhaustive, however. The two

pathways that were identified by GSA as having significant differences in the tumour tissue of the treated patients when compared to the untreated patients may contain genes that are regulated by vitamin D but that have not yet been identified as such. Continuing research will further elucidate the activity of vitamin D in the body and thus other genes that are regulated by vitamin D are likely to be found. What would be interesting to see would be the impact of a prolonged period of supplementation with a dose of vitamin D that keeps the serum calcidiol level at recommended optimal levels of > 80nmol/L. Prolonged supplementation with vitamin D may also allow the observation of further significantly differentially expressed pathways. These could be used to enable the discovery of further genes modulated by vitamin D.

Over-representation analyses performed using differentially expressed genes from each of the normal, tumour and paired analyses were carried out using the Wikipathways, GeneSetDB and Reactome databases and the GATHER online tool. For a pathway to be identified as significant in over-representation analysis it was required to have more genes in it from the presented pathway than would be expected by chance. The Wikipathways database generated too many 'significant' pathways to be useful. They were presented without p-values and FDRs, so that it was not possible to even focus on a most highly significant subgroup, which might have proven to be more useful. The Reactome and GeneSetDB databases provided some more useful signals in that the pathways were returned with FDR values. The drawback of the information presented by these databases, however, is an inability to construct visualisations of the data that help with interpretation of the activity of the pathways and the genes within.

Gene-set analysis using the limma package was useful in this respect and the results from these methods are discussed below. In the normal tissue, the GeneSetDB enrichment analysis identified three vitamin processing pathways in the treatment group compared with the placebo group. When GSA was performed for the normal tissue, a vitamin processing pathway was also identified in the treated group when compared with the control group. Vitamin processing may be modulated by vitamin D during baseline activity in the normal tissue if calcidiol concentrations are at proposed optimal levels. Metabolism pathways featured from over-representation analysis carried out in the tumour tissue of the treated patients using the GeneSetDB and Reactome databases when compared with the untreated group. The Reactome database also

identified ‘cellular response to stress’ pathways. Metabolism pathways were the mainstay of significant pathways identified in the GSA performed in the tumour tissue of the treatment group compared with the control group. Two of these pathways, discussed below, may prove to be informative regarding vitamin D activity in the tumour tissue of treated participants.

In the paired analysis, the MAPK Signaling Pathway was identified by GeneSetDB analysis and cellular response to stress pathways were identified by Reactome database analysis. These pathways did not have visualisations generated for them (none of the pathways identified by over-representation had visualisations generated for them) thus comment on directionality of expression was not possible. These pathways did not feature in the GSA undertaken in the paired analysis.

Gene-set analyses using DEG from the single gene analyses in normal, tumour and paired analyses were performed using the *geneSetTest* function in limma. Pathways were identified as being significantly enriched for highly ranked genes between the study groups in the normal, tumour and paired analyses. In the normal tissue the treated patients had down-regulation of pathways compared with the untreated group including fatty acid metabolism, xenobiotic metabolism, vitamin digestion and absorption and negative regulation of growth pathways. Examination of each of the pathway metagene expression levels between the study groups indicated that the pattern of pathway gene expression was not strongly consistent (in terms of a subset of the pathway genes being tightly correlated) across the normal tissue of the treated participants. However, these pathways may still provide an illustration of baseline vitamin D activity in the normal tissue when calcidiol concentrations are at the recommended optimal level.

In the tumour tissue treated patients had down-regulation of metabolism pathways compared with the untreated group. Two of the metabolism pathways with significant differences in expression were the Fatty Acid Metabolism and Fatty Acid beta-Oxidation pathways. They were demonstrated to have consistent pathway gene expression patterns across the treated participants. These results were determined via metagene analysis and were significant at the 0.05 significance threshold. However, metagene analysis was not corrected for multiple testing and thus although the finding was a

corroborating observation it cannot be strictly referred to as being a significant result. Examination of the pathway genes in a heatmap demonstrated that the majority of the pathway genes in the treated group were down-regulated compared to the untreated group. This is also a confirmatory finding.

In the paired analysis the Ribosome and Translational Termination pathways were enhanced by vitamin D in the tumour tissue of one subgroup of participants and reduced in another subgroup, when compared to normal tissue. The two subgroups were not strictly aligned to the treatment and placebo groups. Examining clinical data for the participants did not provide any elucidation. This is an intriguing finding because vitamin D appears to be having different impacts on the same pathway at a single time point. This is occurring not just in tumour tissue, where one might expect differences due to the presence of different clonal populations but also in the normal tissue. It would be interesting to ascertain why the two subgroups respond differently to vitamin D.

The differences could simply be due to the inclusion of several different cell types in the tissue samples. The samples primarily consisted of mucosal tissue but the mucosa comprises stem, goblet, some paneth, transit amplifying and enteroendocrine cells. Vitamin D activity is tissue- and cell-type specific. The Ribosome and Translational Termination pathways were also identified in the paired analysis independent of vitamin D treatment and were found to be more significant. This signifies that there are many differences in pathway expression between the normal and tumor tissue as expected. In addition to these differences vitamin D appears to have a measurable impact in the normal and tumour tissue, hence the identification of significant pathways.

Energy metabolism is frequently de-regulated in many human cancers. The reprogramming of energy metabolism is prevalent enough that it is now considered an emerging hallmark of tumours. The most frequently observed change is the uptake of increased amounts of glucose for increased cell proliferation in the tumour. Interestingly, the metabolic pathway used for energy release from glucose is not the highly energy producing mitochondrial oxidative phosphorylation (MOP) pathway but rather aerobic glycolysis (AG). Aerobic glycolysis produces approximately 18-fold less energy than the MOP pathway. Cancer cells appear to select the AG pathway even in the presence of oxygen. It has been proposed that glycolytic intermediates of AG may be

diverted to biosynthetic pathways facilitating the synthesis of macromolecules and organelles required for cell proliferation. At present researchers are unsure whether the accepted cancer hallmark of cell proliferation and the emergent cancer hallmark of reprogramming of energy metabolism are independent. It may be that increased cell proliferation leads to altered energy metabolism (Hanahan and Weinberg 2011).
The down-regulation of the energy metabolism pathways, Fatty Acid Metabolism and Fatty Acid beta-Oxidation, is not in accordance with the de-regulated glucose metabolism often observed in tumours. However, fatty acids are the readily available energy source utilised by the epithelium of the colon (Gervaz *et al.* 2004). Short chain fatty acids are produced as by-products of digestion of dietary fibre by members of the colonic microbiota. Acetate is the main energy source for colonocytes in the proximal colon and is produced by *Bacteroides,* which reside there. In the distal colon, *Clostridium*, *Eubacterium* and *Butyrivibrio,* among other bacteria, all produce butyrate, the main energy source for colonocytes in the distal colon. Colonic tumour cells are reported to oxidise butyrate inefficiently due to their predominantly anaerobic metabolism. This allows excess butyrate to inhibit histone deacetylation (HDAC) (reviewed in (Leonel and Alvarez-Leite 2012)). The impact of butyrate's HDAC inhibitory activity leads to reduced cell proliferation and enhanced differentiation and apoptosis (Waldecker *et al.* 2008).

Serpa et al. (2010) found that aggressive tumour cells that retain their ability to metabolise butyrate are selected for in the tumour microenvironment. In their mouse xenograft model, butyrate-preselected human CRC cells produced faster growing tumours that were more angiogenic than tumours derived from untreated cells (Serpa *et al.* 2010). Vitamin D up-regulated protein 1 (VDUP1) reportedly has the ability to regulate fatty acid utilisation (Kim *et al.* 2007). Perhaps vitamin D infiltrates the tumour and is able to down-regulate fatty acid metabolism pathways via VDUP1. This activity may lead to slowing of tumour growth, firstly, by reducing the energy available for cell proliferation and secondly, by allowing un-metabolised butyrate to enter the nucleus and function as an HDAC inhibitor. A proliferation pathway metagene was generated to examine cell proliferation between the study groups to determine whether proliferation may have also been down-regulated in the tumour tissue. There was no significant difference in the expression of the proliferation metagene between the study groups (Appendix 7).

The primary aim of the study was to investigate whether a large dose of vitamin D given to CRC patients prior to surgery could cause a measurable change in activity of the vitamin D transcriptome in the normal and tumour tissue. Elevated serum calcidiol levels were achieved in the treatment group compared with the placebo group. The primary endpoint of the study was to observe a 1.5-fold change in gene expression between the normal and tumour tissue in response to vitamin D. Although this endpoint was not observed, vitamin D may have infiltrated the tumour and caused significant modulation of fatty acid metabolism pathways. This observation of possible vitamin D infiltration of the tumour, and the fact that the biology seems reasonable, may be useful in furthering the understanding of the activity of vitamin D in tumour tissue.

The apparent inability of a large dose of vitamin D to impact the vitamin D transcriptome in a more extensive way could be due to several possibilities. Exposure to the recommended optimal levels of vitamin D may have been too brief. Patients were given 200,000 IU of vitamin D, which raised the serum calcidiol concentration to recommended optimal levels in the treatment group compared with the control group. According to the literature, T_{max} occurs for calcidiol by day seven (section 2.4.1). Further elucidation of the activity of vitamin D in the tumour tissue may be afforded by prolonged exposure to proposed optimal levels of calcidiol. In two studies where patients were supplemented with vitamin D for six months, changes were reported in the tissue of the treatment groups compared with the control groups. In one study the expression of an apoptosis regulator and a differentiation inducer were enhanced in response to vitamin D in the normal tissue of the treated group compared with the placebo group (Fedirko *et al.* 2009a, Fedirko *et al.* 2009b). Proliferation of normal rectal mucosa and polyps were suppressed in response to vitamin D in the treated group compared with the control group. Also the expression of a pro-apoptotic protein-encoding gene was enhanced in the polyps of the treated group (Holt *et al.* 2006) (section 1.3.2).

Resistance of the tumour cells to vitamin D may have occurred (section 1.3.4.4). Resistance to vitamin D by tumour cells can be precipitated by either the suppression of *CYP27B1*, the elevation of *CYP24A1* or the down-regulation of VDR or a combination of the three. *CYP27B1*, *CYP24A1* and *VDR* were included on the VDRE list. They were not identified as differentially expressed when the list was examined for differential

expression between the groups. Although good expression data can be obtained from lower quality RNA (section 5.2.1, Appendix 5b), it is possible that alterations in gene expression were not detected due to this factor. To demonstrate whether resistance to vitamin D may have been an issue in the patients of this study, remaining stored normal and tumour tissue could be analysed for levels of *CYP27B1* and *CYP24A1* gene product and for VDR protein.

In summary, CRC patients were supplemented with a single dose of 200,000 IU of vitamin D administered prior to surgery. Serum calcidiol levels were elevated to within the range regarded as optimal for broader health effects. Vitamin D down-regulated pathways in the normal and tumour tissue compared with the placebo group. Of particular interest are the Fatty Acid Metabolism and Fatty Acid beta-Oxidation pathways that were identified in the tumour tissue of the treated patients compared with the untreated group. The down-regulation of these pathways may precipitate a reduction in tumour cell proliferation and via butyrate a further reduction in proliferation and induction of tumour cell apoptosis. These findings may describe one of the activities of vitamin D in colorectal tumour tissue.

6. CONCLUSION

Colorectal cancer is the third most frequently diagnosed malignancy and the fourth leading cause of death from cancer worldwide (IARC 2015). Although the incidence of CRC is projected to continue to decline in New Zealand over the next few years, decreasing overall by approximately one-quarter in the 45-74 age group, the overall burden is projected to continue increasing as a result of population growth and an aging population (MOH 2010).

Vitamin D is essential for bone health. Recently, it has been postulated that vitamin D may be important for a number of other health outcomes. To maintain bone health and obtain benefits for other health effects, many experts advocate serum calcidiol concentrations of >80nmol/L (Holick 2005a). In NZ, vitamin D deficiency levels are frequently below this threshold, particularly during the winter months (Rockell *et al.* 2006).

Epidemiological studies have reported an inverse association between vitamin D status and incidence of CRC (Ma *et al.* 2011). Higher serum vitamin D levels at the time of diagnosis or post-surgery have been associated with improved long-term outcome in CRC, but there is no proven causal link (Ng *et al.* 2008, Zgaga *et al.* 2014). In patients supplemented with vitamin D, enhancement of apoptosis and differentiation regulators and inhibition of proliferation was reported in normal rectal mucosa, and enhancement of apoptosis and inhibition of proliferation was reported in polyps, in comparison to the placebo group (Fedirko *et al.* 2009a, Fedirko *et al.* 2009b, Holt *et al.* 2006). Suppression of tumour growth was reported in human CRC xenografts when animals were supplemented with vitamin D (Ordonez-Moran *et al.* 2005). Researchers have demonstrated that vitamin D administration alters the gene expression of genes involved in cell proliferation, cell differentiation and apoptosis *in vitro* also (Barnes *et al.* 2005, Díaz *et al.* 2000, Palmer *et al.* 2001, Palmer *et al.* 2003).

The primary aim of the study described in this thesis was to investigate whether a single dose of vitamin D administered to CRC patients in the window between diagnosis and surgery could have a measurable effect on vitamin D regulated genes in the tumour and

corresponding normal colon. The study was a randomised, double-blind, placebo-controlled trial of 200,000 IU (5mg) vitamin D administered at least 7 days prior to surgery. The study was powered to detect a 1.5-fold difference in gene expression, with 25 patients required in the treatment and control arms. This recruitment target was met. No adverse events were reported in relation to administration of the study drug. There were no significant baseline differences between the treatment and control groups, although baseline calcidiol levels were significantly higher in patients diagnosed in the summer and autumn months. Serum calcidiol concentrations were elevated in the treatment group, with a mean of 87nmol/L, compared to the placebo group, with a mean of 49nmol/L. No differences in post-operative recovery, cancer recurrence or survival were expected in this trial and none were observed, however the study was not powered to measure differences in these secondary outcomes.

In single gene analyses performed in the normal tissue, tumour and in paired tissue analyses, no individual genes were identified as being significantly differentially expressed between the treatment and placebo arms. There were also no significant differences in gene expression between the groups when the analysis was restricted to only include vitamin D modulated genes.

Over-representation analyses were performed using differentially expressed genes from each of the normal, tumour and paired analyses. Pathways with differences in expression were identified using the multiple online databases. No pathways with differences in expression were identified by the GATHER online tool. The Wikipathways database identified pathways with differences in expression but these were presented without adjusted p-values and therefore it is difficult to comment on their significance. Pathways identified by the GeneSetDB and Reactome databases were presented with FDR adjusted p-values. The GeneSetDB database identified genes involved in vitamin processing pathways as being significantly over-represented when looking at changes in expression in the normal tissue of the treated patients relative to the placebo group. A vitamin processing pathway was also identified using gene-set analysis in the normal tissue of the treated patients when compared to the untreated group. The GeneSetDB and Reactome databases identified enrichment of genes involved in metabolism pathways in the tumour tissue of the treatment group when compared with the control group, as did gene-set analysis using the limma package in R. The pathways identified in

the paired analysis were not identified via gene-set analysis and thus were not explored further.

Gene-set analyses were performed to identify pathways with differences of expression between the study groups using the differential expression data from each of the normal, tumour and paired analyses. In the normal tissue treated patients had down-regulation of a number of pathways compared with the untreated group including the Fatty Acid Metabolism, Drug Metabolism- Cytochrome P450, Metabolism of Xenobiotics by Cytochrome P450, Vitamin Digestion and Absorption and Negative Regulation of Growth pathways. The down-regulation of these pathways may represent the housekeeping activity of vitamin D in the normal tissue when calcidiol concentrations are adequate.

In the tumour tissue, treated patients also had down-regulation of several pathways compared to the untreated group including the Fatty Acid Metabolism, Fatty Acid beta-Oxidation and Oxidative Phosphorylation pathways. The fatty acid metabolism pathway metagenes exhibited relatively consistent differences in expression across the tumour tissue of the treated patients. This finding may indicate greater significance of these pathways compared with other pathways identified in the tumour tissue analysis. Vitamin D may have infiltrated the tumour tissue and inhibited the expression of these pathways via the vitamin D up-regulated protein 1 (VDUP1). Cell proliferation and energy metabolism are generally increased in tumours compared to normal tissue and down-regulation of fatty acid metabolism in the tumour may impede tumour growth. Unmetabolised butyrate is free to enter the cell nucleus and function as an HDAC inhibitor leading to decreased proliferation and increased differentiation and apoptosis. This may be a mechanism by which vitamin D could inhibit cell growth and promote cell death in the tumour.

In the paired tissue analysis, vitamin D appeared to enhance the expression of the Ribosome and Translational Termination pathways in the tumour tissue of some patients and inhibit expression of these pathways in others, relative to the normal tissue. These sub-groups were not distinguishable in terms of demographic or clinico-pathological characteristics. The reasons for these observed differences and their biological significance are unclear.

The main strength of this study was the study design. The absence of significant baseline differences between the groups and the blinded analysis means that the risk of bias was minimised and significant differences in outcome between the groups is likely to be real.

A limitation of the study may be variability in the quality of the RNA. There was a variable time delay between collecting resected bowel specimens from theatre and immersion in RNAlater® or direct freezing at -80^{o}C, due to pathologist availability. The sample size was also small, and although adequately powered for molecular primary endpoint, the study was not intended and not able to address the more important clinical endpoints.

In summary, a single dose of 200,000 IU of vitamin D, administered at least seven days prior to patients undergoing surgery for CRC resulted in serum levels within the range regarded as optimal for broader health effects. Vitamin D down-regulated pathways in the normal and tumour tissue compared with the placebo group in particular those related to Fatty Acid Metabolism and Fatty Acid beta-Oxidation. The down-regulation of these pathways may induce reduced proliferation and increased apoptosis via butyrate induced HDAC inhibition. These findings indicate that some biological effects can be detected in the tumour, even after short term administration of vitamin D, and may indicate one mechanism by which vitamin D can have an impact in tumour tissue.

REFERENCES

Abe, E., Miyaura, C., Sakagami, H., Takeda, M., Konno, K., Yamazaki, T., Yoshiki, S. and Suda, T. (1981) 'Differentiation of mouse myeloid leukemia cells induced by 1 alpha, 25-dihydroxyvitamin D3', *Proceedings of the National Academy of Sciences*, 78(8), 4990-4994.

Affymetrix (2005) *Exon Array Background Correction* [online], available: http://media.affymetrix.com/support/technical/whitepapers/exon_background_correction_whitepaper.pdf [accessed 2015].

Affymetrix (2007) *Technical Note, GeneChip® Gene 1.0 ST Array Design* [online], available: http://static1.1.sqspcdn.com/static/f/1438485/21486042/1357241021290/gene_1_0_st_technote.pdf?token=cx5KljY3o%2BN4j3yxeDIQSxiq7dk%3D [accessed 2015].

Affymetrix (2015) *GeneChip® PrimeView™ Human Gene Expression Array* [online], available: http://www.affymetrix.com/estore/catalog/prod530005/AFFY/PrimeView%26%23153%3B+Human+Gene+Expression+Array+ - 1_1 [accessed 2013].

Aguilera, O., Peña, C., García, J. M., Larriba, M. J., Ordóñez-Morán, P., Navarro, D., Barbáchano, A., de Silanes, I. L., Ballestar, E. and Fraga, M. F. (2007) 'The Wnt antagonist DICKKOPF-1 gene is induced by 1 α , 25-dihydroxyvitamin D3 associated to the differentiation of human colon cancer cells', *Carcinogenesis*, 28(9), 1877-1884.

Ajibade, D., Benn, B. S. and Christakos, S. (2010) 'Mechanism of action of 1, 25-dihydroxyvitamin D3 on intestinal calcium absorption and renal calcium transport' in *vitamin D*Springer, 175-187.

Alter, O., Brown, P. O. and Botstein, D. (2000) 'Singular value decomposition for genome-wide expression data processing and modeling', *Proceedings of the National Academy of Sciences*, 97(18), 10101-10106.

Anderson, M. G., Nakane, M., Ruan, X., Kroeger, P. E. and Wu-Wong, J. R. (2006) 'Expression of VDR and CYP24A1 mRNA in human tumors', *Cancer chemotherapy and pharmacology*, 57(2), 234-240.

Araki, H., Knapp, C., Tsai, P. and Print, C. (2012) 'GeneSetDB: A comprehensive meta-database, statistical and visualisation framework for gene set analysis', *FEBS Open Bio*, 2, 76-82.

Arner, P., Bernard, S., Salehpour, M., Possnert, G., Liebl, J., Steier, P., Buchholz, B. A., Eriksson, M., Arner, E. and Hauner, H. (2011) 'Dynamics of human adipose lipid turnover in health and metabolic disease', *Nature*, 478(7367), 110-113.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L. and Kasarskis, A. (2000) 'Gene Ontology: tool for the unification of biology', *Nature Genetics*, 25(1), 25-29.

Autier, P., Boniol, M., Pizot, C. and Mullie, P. (2014) 'Vitamin D status and ill health: a systematic review', *The Lancet Diabetes & Endocrinology*, 2(1), 76-89.

Bareis, P., Bises, G., Bischof, M. G., Cross, H. S. and Peterlik, M. (2001) '25-hydroxy-vitamin d metabolism in human colon cancer cells during tumor progression', *Biochemical and Biophysical Research Communications*, 285(4), 1012-1017.

Barger-Lux, M. J., Heaney, R. P., Dowell, S., Chen, T. C. and Holick, M. F. (1998) 'Vitamin D and its major metabolites: serum levels after graded oral dosing in healthy men', *Osteoporosis International*, 8(3), 222-30.

Barnes, J., Arhel, N., Lee, S., Sharp, A., Al-Okail, M., Packham, G., Hague, A., Paraskeva, C. and Williams, A. (2005) 'Nuclear BAG-1 expression inhibits apoptosis in colorectal adenoma-derived epithelial cells', *Apoptosis*, 10(2), 301-311.

Barry, W. T., Nobel, A. B. and Wright, F. A. (2005) 'Significance analysis of functional categories in gene expression studies: a structured permutation approach', *Bioinformatics*, 21(9), 1943-1949.

Ben-Shoshan, M., Amir, S., Dang, D. T., Dang, L. H., Weisman, Y. and Mabjeesh, N. J. (2007) '1alpha,25-dihydroxyvitamin D3 (Calcitriol) inhibits hypoxia-inducible factor-1/vascular endothelial growth factor pathway in human cancer cells', *Molecular Cancer Therapeutics*, 6(4), 1433-9.

Benjamini, Y. and Hochberg, Y. (1995) 'Controlling the false discovery rate: a practical and powerful approach to multiple testing', *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300.

Bikle, D. D., Teichert, A., Arnold, L. A., Uchida, Y., Elias, P. M. and Oda, Y. (2010) 'Differential regulation of epidermal function by VDR coactivators', *Journal of Steroid Biochemistry & Molecular Biology*, 121(1-2), 308-13.

Bises, G., Kallay, E., Weiland, T., Wrba, F., Wenzl, E., Bonner, E., Kriwanek, S., Obrist, P. and Cross, H. S. (2004) '25-hydroxyvitamin D3-1alpha-hydroxylase expression in normal and malignant human colon', *Journal of Histochemistry & Cytochemistry*, 52(7), 985-9.

Bolland, M. J., Grey, A. B., Ames, R. W., Mason, B. H., Horne, A. M., Gamble, G. D. and Reid, I. R. (2006) 'Determinants of vitamin D status in older men living in a subtropical climate', *Osteoporosis International*, 17(12), 1742-8.

Bolstad, B. (2002) 'Comparing the effects of background, normalization and summarization on gene expression estimates', *Unpublished Manuscript*.

Bolstad, B. M., Irizarry, R. A., Åstrand, M. and Speed, T. P. (2003) 'A comparison of normalization methods for high density oligonucleotide array data based on variance and bias', *Bioinformatics*, 19(2), 185-193.

Bouillon, R. (2011) 'Why modest but widespread improvement of the vitamin D status is the best strategy?', *Best Practice & Research Clinical Endocrinology & Metabolism*, 25(4), 693-702.

Bouillon, R., Carmeliet, G., Verlinden, L., van Etten, E., Verstuyf, A., Luderer, H. F., Lieben, L., Mathieu, C. and Demay, M. (2008) 'Vitamin D and human health: lessons from vitamin D receptor null mice', *Endocrine Reviews*, 29(6), 726-76.

Box, J. F. (1987) 'Guinness, Gosset, Fisher, and small samples', *Statistical Science*, 45-52.

Brandi, M. L. and Minisola, S. (2013) 'Calcidiol [25 (OH) D3]: from diagnostic marker to therapeutical agent', *Current Medical Research & Opinion*, 29(11), 1565-1572.

Braun, M. M., Helzlsouer, K. J., Hollis, B. W. and Comstock, G. W. (1995) 'Colon cancer and serum vitamin D metabolite levels 10–17 years prior to diagnosis', *American Journal of Epidemiology*, 142(6), 608-608.

Brehier, A. and Thomasset, M. (1988) 'Human colon cell line HT-29: Characterisation of 1,25-dihydroxyvitamin D3 receptor and induction of differentiation by the hormone', *Journal of Steroid Biochemistry*, 29(2), 265-270.

Brenner, M. and Hearing, V. J. (2008) 'The Protective Role of Melanin Against UV Damage in Human Skin†', *Photochemistry and photobiology*, 84(3), 539-549.

Carlberg, C. and Campbell, M. J. (2012) 'Vitamin D receptor signaling mechanisms: Integrated actions of a well-defined transcription factor', *Steroids*, 78(2), 127-136.

Carlberg, C. and Dunlop, T. W. (2006) 'The impact of chromatin organization of vitamin D target genes', *Anticancer Research*, 26(4A), 2637-45.

Carlberg, C. and Seuter, S. (2009) 'A genomic perspective on vitamin D signaling', *Anticancer Research*, 29(9), 3485-3493.

Chakravarti, I. and Laha, R. (1967) 'Handbook of Methods of Applied Statistics.', 1, 392-394.

Chang, J. T. and Nevins, J. R. (2006) 'GATHER: a systems approach to interpreting genomic signatures', *Bioinformatics*, 22(23), 2926-2933.

Chen, A., Davis, B. H., Sitrin, M. D., Brasitus, T. A. and Bissonnette, M. (2002) 'Transforming growth factor-beta 1 signaling contributes to Caco-2 cell growth inhibition induced by 1,25(OH)(2)D(3)', *American Journal of Physiology - Gastrointestinal & Liver Physiology*, 283(4), G864-74.

Chen, W., Clements, M., Rahman, B., Zhang, S., Qiao, Y. and Armstrong, B. K. (2010) 'Relationship between cancer mortality/incidence and ambient ultraviolet B irradiance in China', *Cancer Causes & Control*, 21(10), 1701-9.

Cianferotti, L., Cox, M., Skorija, K. and Demay, M. B. (2007) 'Vitamin D receptor is essential for normal keratinocyte stem cell function', *Proceedings of the National Academy of Sciences of the United States of America*, 104(22), 9428-33.

Cleveland, W. S. (1979) 'Robust locally weighted regression and smoothing scatterplots', *Journal of the American Statistical Association*, 74(368), 829-836.

Clevers, H. (2006) 'Wnt/beta-catenin signaling in development and disease', *Cell*, 127(3), 469-80.

Creighton University, S. o. M. (2015) *Vitamin D: How long is vitamin D stored in the body?* [online], available: http://medschool.creighton.edu/search/?cx=015839161432768919274%3Aaxvf6sw9siu&cof=FORID%3A10&ie=UTF-8&q=vitamin+D+adipose+ [accessed 28 July 2015].

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M. R., Jassal, B., Jupe, S., Matthews, L., May, B., Palatnik, S., Rothfels, K., Shamovsky, V., Song, H., Williams, M., Birney, E., Hermjakob, H., Stein, L. and D'Eustachio, P. (2014) 'The Reactome pathway knowledgebase', *Nucleic Acids Research*, 42(D1), D472-D477.

Cross, H. S., Bareis, P., Hofer, H., Bischof, M. G., Bajna, E., Kriwanek, S., Bonner, E. and Peterlik, M. (2001) '25-Hydroxyvitamin D(3)-1alpha-hydroxylase and vitamin D receptor gene expression in human colonic mucosa is elevated during early cancerogenesis', *Steroids*, 66(3-5), 287-92.

Cross, H. S., Kallay, E., Farhan, H., Weiland, T. and Manhardt, T. (2003) 'Regulation of extrarenal vitamin D metabolism as a tool for colon and prostate cancer prevention', *Recent Results in Cancer Research*, 164, 413-25.

DeBarros, M. and Steele, S. R. (2013) 'Perioperative protocols in colorectal surgery', *Clinics in colon and rectal surgery*, 26(3), 139.

Deeb, K. K., Trump, D. L. and Johnson, C. S. (2007) 'Vitamin D signalling pathways in cancer: potential for anticancer therapeutics', *Nature Reviews. Cancer*, 7(9), 684-700.

Díaz, G. D., Paraskeva, C., Thomas, M. G., Binderup, L. and Hague, A. (2000) 'Apoptosis is induced by the active metabolite of vitamin D3 and its analogue EB1089 in colorectal adenoma and carcinoma cells: possible implications for prevention and therapy', *Cancer Research*, 60(8), 2304-2312.

Dindo, D., Demartines, N. and Clavien, P. A. (2004) 'Classification of surgical complications: a new proposal with evaluation in a cohort of 6336 patients and results of a survey', *Annals of Surgery*, 240(2), 205-13.

Ding, E. L., Mehta, S., Fawzi, W. W. and Giovannucci, E. L. (2008) 'Interaction of estrogen therapy with calcium and vitamin D supplementation on colorectal cancer risk: reanalysis of Women's Health Initiative randomized trial', *International Journal of Cancer*, 122(8), 1690-1694.

Dripps, R. D., Lamont, A. and Eckenhoff, J. E. (1961) 'The role of anesthesia in surgical mortality', *Jama*, 178(3), 261-266.

Edge, S. B. and Compton, C. C. (2010) 'The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM', *Annals of surgical oncology*, 17(6), 1471-1474.

Engelsen, O. (2010) 'The relationship between ultraviolet radiation exposure and vitamin D status', *Nutrients*, 2(5), 482-95.

Fedirko, V., Bostick, R. M., Flanders, W. D., Long, Q., Shaukat, A., Rutherford, R. E., Daniel, C. R., Cohen, V. and Dash, C. (2009a) 'Effects of vitamin D and calcium supplementation on markers of apoptosis in normal colon mucosa: a randomized, double-blind, placebo-controlled clinical trial', *Cancer Prevention Research*, 2(3), 213-23.

Fedirko, V., Bostick, R. M., Flanders, W. D., Long, Q., Sidelnikov, E., Shaukat, A., Daniel, C. R., Rutherford, R. E. and Woodard, J. J. (2009b) 'Effects of vitamin d and calcium on proliferation and differentiation in normal colon mucosa: a randomized clinical trial', *Cancer Epidemiology, Biomarkers & Prevention*, 18(11), 2933-41.

Feldman, D., Krishnan, A. V., Swami, S., Giovannucci, E. and Feldman, B. J. (2014) 'The role of vitamin D in reducing cancer risk and progression', *Nature Reviews. Cancer*, 14(5), 342-57.

Feskanich, D., Ma, J., Fuchs, C. S., Kirkner, G. J., Hankinson, S. E., Hollis, B. W. and Giovannucci, E. L. (2004) 'Plasma vitamin D metabolites and risk of colorectal cancer in women', *Cancer Epidemiology, Biomarkers & Prevention*, 13(9), 1502-8.

Fisher, R. (1955) 'Statistical Methods and Scientific Induction', *Journal of the Royal Statistical Society. Series B (Methodological)*, 17(1), 69-78.

Flossmann, E., Rothwell, P. M., British Doctors Aspirin, T. and the, U. K. T. I. A. A. T. (2007) 'Effect of aspirin on long-term risk of colorectal cancer: consistent evidence from randomised and observational studies', *Lancet*, 369(9573), 1603-13.

Freudenberg, J. M. (2005) 'Comparison of background correction and normalization procedures for high-density oligonucleotide microarrays', *Institut fur Informatik*, 120(11).

Fridley, B. L., Jenkins, G. D. and Biernacka, J. M. (2010) 'Self-contained gene-set analysis of expression data: an evaluation of existing and novel methods', *PLoS ONE [Electronic Resource]*, 5(9), e12693.

Gandini, S., Boniol, M., Haukka, J., Byrnes, G., Cox, B., Sneyd, M. J., Mullie, P. and Autier, P. (2011) 'Meta-analysis of observational studies of serum 25-hydroxyvitamin D levels and colorectal, breast and prostate cancer and colorectal adenoma', *International Journal of Cancer*, 128(6), 1414-24.

Garland, C. F. and Garland, F. C. (1980) 'Do sunlight and vitamin D reduce the likelihood of colon cancer?', *International journal of epidemiology*, 9(3), 227.

Garland, C. F., Garland, F. C., Shaw, E., Comstock, G., Helsing, K. and Gorham, E. (1989) 'Serum 25-hydroxyvitamin D and colon cancer: eight-year prospective study', *The Lancet*, 334(8673), 1176-1178.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. and Zhang, J. (2004) 'Bioconductor: open software development for computational biology and bioinformatics', *Genome Biology*, 5(10), R80.

Gervaz, P., Bucher, P. and Morel, P. (2004) 'Two colons-two cancers: paradigm shift and clinical implications', *Journal of Surgical Oncology*, 88(4), 261-6.

Giovannucci, E. (2011) 'Vitamin D, how much is enough and how much is too much?', *Public Health Nutrition*, 14(4), 740-1.

Giovannucci, E. (2013) 'Epidemiology of vitamin D and colorectal cancer', *Current Medicinal Chemistry - Anti-Cancer Agents*, 13(1), 11-9.

Girardin, A., McCall, J., Black, M. A., Edwards, F., Phillips, V., Taylor, E. S., Reeve, A. E. and Kemp, R. A. (2013) 'Inflammatory and regulatory T cells contribute to a unique immune microenvironment in tumor tissue of colorectal cancer patients', *International Journal of Cancer*, 132(8), 1842-50.

GitHub (2015) *GitHub, Inc.* [online], available: https://github.com/ [accessed 2015].

Giuliano, A. R., Franceschi, R. T. and Wood, R. J. (1991) 'Characterization of the vitamin D receptor from the Caco-2 human colon carcinoma cell line: Effect of cellular differentiation', *Archives of Biochemistry and Biophysics*, 285(2), 261-269.

Goeman, J. J. and Bühlmann, P. (2007) 'Analyzing gene expression data in terms of gene sets: methodological issues', *Bioinformatics*, 23(8), 980-987.

Gozdzik, A., Barta, J. L., Wu, H., Wagner, D., Cole, D. E., Vieth, R., Whiting, S. and Parra, E. J. (2008) 'Low wintertime vitamin D levels in a sample of healthy young adults of diverse ancestry living in the Toronto area: associations with vitamin D intake and skin pigmentation', *BMC Public Health*, 8, 336.

Grant, W. B. (2002) 'An estimate of premature cancer mortality in the U.S. due to inadequate doses of solar ultraviolet-B radiation', *Cancer*, 94(6), 1867-75.

Halline, A., Davidson, N., Skarosi, S., Sitrin, M., Tietze, C., Alpers, D. and Brasitus, T. (1994) 'Effects of 1, 25-dihydroxyvitamin D3 on proliferation and differentiation of Caco-2 cells', *Endocrinology*, 134(4), 1710-1717.

Hanahan, D. and Weinberg, Robert A. (2011) 'Hallmarks of Cancer: The Next Generation', *Cell*, 144(5), 646-674.

Hansen, C. M., Binderup, L., Hamberg, K. J. and Carlberg, C. (2001) 'Vitamin D and cancer: effects of 1, 25 (OH) 2D3 and its analogs on growth control and tumorigenesis', *Frontiers in Bioscience*, 6, D820-D848.

Hewison, M. (2011) 'Vitamin D and innate and adaptive immunity', *Vitamins & Hormones*, 86, 23-62.

Hochberg, Y. and Tamhane, A. C. (1987) *Multiple Comparison Procedures.*, Wiley and Sons: New York.

Hock, A. D. (2014) 'Review: Vitamin D3 deficiency results in dysfunctions of immunity with severe fatigue and depression in a variety of diseases', *In Vivo*, 28(1), 133-45.

Hoeffding, W. (1948) 'A Non-Parametric Test of Independence', *The Annals of Mathematical Statistics*, 19(4), 546-557.

Hogg, R. V. (1979) 'Statistical robustness: One view of its use in applications today', *The American Statistician*, 33(3), 108-115.

Holick, M. F. (1995) 'Environmental factors that influence the cutaneous production of vitamin D.(Nutrition and Environmental Health)', *American Journal of Clinical Nutrition*, 61(3), 638S.

Holick, M. F. (2000) 'Sunlight and vitamin D: The bone and cancer connections', *Radiation protection dosimetry*, 91(1-3), 65-71.

Holick, M. F. (2003) 'Evolution and function of vitamin D', *Recent Results in Cancer Research*, 164, 3-28.

Holick, M. F. (2005a) 'The influence of vitamin D on bone health across the life cycle', *Journal of Nutrition*, 135(11), 2726S-7S.

Holick, M. F. (2005b) 'The vitamin D epidemic and its health consequences', *Journal of Nutrition*, 135(11), 2739S-48S.

Holick, M. F. (2006) 'High prevalence of vitamin D inadequacy and implications for health', *Mayo Clinic Proceedings*, 81(3), 353-73.

Holick, M. F. (2007) 'Vitamin D deficiency', *New England Journal of Medicine*, 357(3), 266-81.

Holick, M. F. (2011) 'Photobiology of vitamin D', *vitamin D*, 1, 13-22.

Holick, M. F. (2013) 'Vitamin D, sunlight and cancer connection', *Current Medicinal Chemistry - Anti-Cancer Agents*, 13(1), 70-82.

Holick, M. F., Binkley, N. C., Bischoff-Ferrari, H. A., Gordon, C. M., Hanley, D. A., Heaney, R. P., Murad, M. H. and Weaver, C. M. (2011) 'Evaluation, treatment, and prevention of vitamin D deficiency: an Endocrine Society clinical practice guideline', *The Journal of Clinical Endocrinology & Metabolism*, 96(7), 1911-1930.

Hollis, B. W. and Wagner, C. L. (2013) 'Clinical review: The role of the parent compound vitamin D with respect to metabolism and function: Why clinical dose intervals can affect clinical outcomes', *Journal of Clinical Endocrinology & Metabolism*, 98(12), 4619-28.

Holt, P. R., Bresalier, R. S., Ma, C. K., Liu, K. F., Lipkin, M., Byrd, J. C. and Yang, K. (2006) 'Calcium plus vitamin D alters preneoplastic features of colorectal adenomas and rectal mucosa', *Cancer*, 106(2), 287-96.

Hopkins, M. H., Owen, J., Ahearn, T., Fedirko, V., Flanders, W. D., Jones, D. P. and Bostick, R. M. (2011) 'Effects of supplemental vitamin D and calcium on biomarkers of inflammation in colorectal adenoma patients: a randomized, controlled clinical trial', *Cancer Prevention Research*, 4(10), 1645-54.

Hossein-nezhad, A. and Holick, M. F. (2013) 'Vitamin D for Health: A Global Perspective', *Mayo Clinic Proceedings*, 88(7), 720-755.

Huerta, S., Irwin, R. W., Heber, D., Go, V. L., Koeffler, H. P., Uskokovic, M. R. and Harris, D. M. (2002) '1alpha,25-(OH)(2)-D(3) and its synthetic analogue decrease tumor load in the Apc(min) Mouse', *Cancer Research*, 62(3), 741-6.

Iacopetta, B. (2003) 'TP53 mutation in colorectal cancer', *Human mutation*, 21(3), 271-276.

IARC (2008) *Vitamin D and cancer* [online], available: http://www.iarc.fr/en/publications/pdfs-online/wrk/wrk5/Report_VitD.pdf [accessed 2014].

IARC (2015) *GLOBOCAN 2012:Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012* [online], available: http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx [accessed 2015].

Ilahi, M., Armas, L. A. and Heaney, R. P. (2008) 'Pharmacokinetics of a single, large dose of cholecalciferol', *American Journal of Clinical Nutrition*, 87(3), 688-91.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B. and Speed, T. P. (2003a) 'Summaries of Affymetrix GeneChip probe level data', *Nucleic Acids Research*, 31(4), e15-e15.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer‐Barclay, Y. D., Antonellis, K. J., Scherf, U. and Speed, T. P. (2003b) 'Exploration, normalization, and summaries of high density oligonucleotide array probe level data', *Biostatistics*, 4(2), 249-264.

Jenab, M., Bueno-de-Mesquita, H. B., Ferrari, P., van Duijnhoven, F. J., Norat, T., Pischon, T., Jansen, E. H., Slimani, N., Byrnes, G., Rinaldi, S., Tjonneland, A., Olsen, A., Overvad, K., Boutron-Ruault, M. C., Clavel-Chapelon, F., Morois, S., Kaaks, R., Linseisen, J., Boeing, H., Bergmann, M. M., Trichopoulou, A., Misirli, G., Trichopoulos, D., Berrino, F., Vineis, P., Panico, S., Palli, D., Tumino, R., Ros, M. M., van Gils, C. H., Peeters, P. H., Brustad, M., Lund, E., Tormo, M. J., Ardanaz, E., Rodriguez, L., Sanchez, M. J., Dorronsoro, M., Gonzalez, C. A., Hallmans, G., Palmqvist, R., Roddam, A., Key, T. J., Khaw, K. T., Autier, P., Hainaut, P. and Riboli, E. (2010) 'Association between pre-diagnostic circulating vitamin D concentration and risk of colorectal cancer in European populations:a nested case-control study', *BMJ*, 340, b5500.

Jones, G. (2008) 'Pharmacokinetics of vitamin D toxicity', *The American journal of clinical nutrition*, 88(2), 582S-586S.

Jones, G. and Dwyer, T. (1998) 'Bone Mass in Prepubertal Children: Gender Differences and the Role of Physical Activity and Sunlight Exposure', *The Journal of Clinical Endocrinology & Metabolism*, 83(12), 4274-4279.

Judd, S. E., Raiser, S. N., Kumari, M. and Tangpricha, V. (2010) '1,25-dihydroxyvitamin D3 reduces systolic blood pressure in hypertensive adults: a pilot feasibility study', *Journal of Steroid Biochemistry & Molecular Biology*, 121(1-2), 445-7.

Kaminsky, L. S. and Zhang, Q.-Y. (2003) 'The small intestine as a xenobiotic-metabolizing organ', *Drug metabolism and disposition*, 31(12), 1520-1525.

Kanehisa, M. and Goto, S. (2000) 'KEGG: Kyoto Encyclopedia of Genes and Genomes', *Nucleic Acids Research*, 28(1), 27-30.

Kang, W., Lee, S., Jeon, E., Yun, Y.-R., Kim, K.-H. and Jang, J.-H. (2011) 'Emerging role of vitamin D in colorectal cancer', *World journal of gastrointestinal oncology*, 3(8), 123.

Kaplan, E. L. and Meier, P. (1958) 'Nonparametric Estimation from Incomplete Observations', *Journal of the American Statistical Association*, 53(282), 457-481.

Kauffmann, A., Gentleman, R. and Huber, W. (2009) 'arrayQualityMetrics—a bioconductor package for quality assessment of microarray data', *Bioinformatics*, 25(3), 415-416.

Kelder, T., van Iersel, M. P., Hanspers, K., Kutmon, M., Conklin, B. R., Evelo, C. T. and Pico, A. R. (2012) 'WikiPathways: building research communities on biological pathways', *Nucleic Acids Research*, 40(D1), D1301-D1307.

Kerner, S. A., Scott, R. A. and Pike, J. W. (1989) 'Sequence elements in the human osteocalcin gene confer basal activation and inducible response to hormonal vitamin D3', *Proceedings of the National Academy of Sciences*, 86(12), 4455-4459.

Kim, S. Y., Suh, H. W., Chung, J. W., Yoon, S. R. and Choi, I. (2007) 'Diverse functions of VDUP1 in cell proliferation, differentiation, and diseases', *Cellular & Molecular Immunology*, 4(5), 345-51.

Klaus, A. and Birchmeier, W. (2008) 'Wnt signalling and its impact on development and cancer', *Nature Reviews. Cancer*, 8(5), 387-98.

Koutkia, P., Chen, T. C. and Holick, M. F. (2001) 'Vitamin D intoxication associated with an over-the-counter supplement', *New England Journal of Medicine*, 345(1), 66-67.

Kriebitzsch, C., Verlinden, L., Eelen, G., Tan, B. K., Van Camp, M., Bouillon, R. and Verstuyf, A. (2009) 'The impact of 1,25(OH)2D3 and its structural analogs on gene expression in cancer cells--a microarray approach', *Anticancer Research*, 29(9), 3471-83.

Krishnan, A. V. and Feldman, D. (2011) 'Mechanisms of the anti-cancer and anti-inflammatory actions of vitamin D', *Annual Review of Pharmacology & Toxicology*, 51, 311-36.

Kupferschmidt, K. (2012) 'Uncertain verdict as vitamin D goes on trial', *Science*, 337(6101), 1476-8.

Lagishetty, V., Misharin, A. V., Liu, N. Q., Lisse, T. S., Chun, R. F., Ouyang, Y., McLachlan, S. M., Adams, J. S. and Hewison, M. (2010) 'Vitamin D deficiency in mice impairs colonic antibacterial activity and predisposes to colitis', *Endocrinology*, 151(6), 2423-2432.

Lamprecht, S. A. and Lipkin, M. (2003) 'Chemoprevention of colon cancer by calcium, vitamin D and folate: molecular mechanisms', *Nature Reviews. Cancer*, 3(8), 601-14.

Lappe, J. M., Travers-Gustafson, D., Davies, K. M., Recker, R. R. and Heaney, R. P. (2007) 'Vitamin D and calcium supplementation reduces cancer risk: results of a randomized trial.[Erratum appears in Am J Clin Nutr. 2008 Mar;87(3):794]', *American Journal of Clinical Nutrition*, 85(6), 1586-91.

Larriba, M. J., Bonilla, F. and Muñoz, A. (2010) 'The transcription factors Snail1 and Snail2 repress vitamin D receptor during colon cancer progression', *The Journal of steroid biochemistry and molecular biology*, 121(1), 106-109.

Larriba, M. J., Martin-Villar, E., Garcia, J. M., Pereira, F., Pena, C., de Herreros, A. G., Bonilla, F. and Munoz, A. (2009) 'Snail2 cooperates with Snail1 in the repression of vitamin D receptor in colon cancer', *Carcinogenesis*, 30(8), 1459-68.

Larriba, M. J. and Munoz, A. (2005) 'SNAIL vs vitamin D receptor expression in colon cancer: therapeutics implications', *British Journal of Cancer*, 92(6), 985-989.

Larriba, M. J., Ordonez-Moran, P., Chicote, I., Martin-Fernandez, G., Puig, I., Munoz, A. and Palmer, H. G. (2011) 'Vitamin D receptor deficiency enhances Wnt/beta-catenin signaling and tumor burden in colon cancer', *PLoS ONE [Electronic Resource]*, 6(8), e23524.

Larriba, M. J., Valle, N., Alvarez, S. and Munoz, A. (2008) 'Vitamin D3 and colorectal cancer', *Advances in Experimental Medicine & Biology*, 617, 271-80.

Larriba, M. J., Valle, N., Palmer, H. G., Ordonez-Moran, P., Alvarez-Diaz, S., Becker, K. F., Gamallo, C., de Herreros, A. G., Gonzalez-Sancho, J. M. and Munoz, A. (2007) 'The inhibition of Wnt/beta-catenin signalling by 1alpha,25-dihydroxyvitamin D3 is abrogated by Snail1 in human colon cancer cells', *ENDOCRINE-RELATED CANCER*, 14(1), 141-51.

Lee, J. E., Li, H., Chan, A. T., Hollis, B. W., Lee, I. M., Stampfer, M. J., Wu, K., Giovannucci, E. and Ma, J. (2011) 'Circulating levels of vitamin D and colon and rectal cancer: the Physicians' Health Study and a meta-analysis of prospective studies', *Cancer Prevention Research*, 4(5), 735-43.

Leonel, A. J. and Alvarez-Leite, J. I. (2012) 'Butyrate: implications for intestinal function', *Current Opinion in Clinical Nutrition & Metabolic Care*, 15(5), 474-479.

Leyssens, C., Verlinden, L. and Verstuyf, A. (2013) 'Antineoplastic effects of 1,25(OH)(2)D-3 and its analogs in breast, prostate and colorectal cancer', *ENDOCRINE-RELATED CANCER*, 20(2), R31-R47.

Lieberman, S. and Bruning, N. (1990) *The Real Vitamin & Mineral Book. NY: Avery Group, 3*ISBN 0-89529-769-8.

Liel, Y., Ulmer, E., Shary, J., Hollis, B. W. and Bell, N. H. (1988) 'Low circulating vitamin D in obesity', *Calcified tissue international*, 43(4), 199-201.

Liu, P. and Hwang, J. T. (2007) 'Quick calculation for sample size while controlling false discovery rate with application to microarray analysis.[Erratum appears in Bioinformatics. 2008 Jan 1;24(1):149]', *Bioinformatics*, 23(6), 739-46.

Livesey, J., Elder, P., Ellis, M. J., McKenzie, R., Liley, B. and Florkowski, C. (2007) 'Seasonal variation in vitamin D levels in the Canterbury, New Zealand population in relation to available UV radiation', *New Zealand Medical Journal*, 120(1262), U2733.

Luxwolda, M. F., Kuipers, R. S., Kema, I. P., Dijck-Brouwer, D. A. and Muskiet, F. A. (2012) 'Traditionally living populations in East Africa have a mean serum 25-hydroxyvitamin D concentration of 115 nmol/l', *British Journal of Nutrition*, 108(9), 1557-61.

Ma, Y., Zhang, P., Wang, F., Yang, J., Liu, Z. and Qin, H. (2011) 'Association between vitamin D and risk of colorectal cancer: a systematic review of prospective studies', *J Clin Oncol*, 29(28), 3775-82.

Makishima, M., Lu, T. T., Xie, W., Whitfield, G. K., Domoto, H., Evans, R. M., Haussler, M. R. and Mangelsdorf, D. J. (2002) 'Vitamin D receptor as an intestinal bile acid sensor', *Science*, 296(5571), 1313-6.

Mallet, E., Philippe, F., Castanet, M. and Basuyau, J. P. (2010) '[Administration of a single Winter oral dose of 200,000 IU of vitamin D3 in adolescents in Normandy: evaluation of the safety and vitamin D status obtained]', *Archives de Pediatrie*, 17(7), 1042-6.

Mantel, N. (1966) 'Evaluation of survival data and two new rank order statistics arising in its consideration', *Cancer chemotherapy reports. Part 1*, 50(3), 163-170.

Markowitz, S. D. and Bertagnolli, M. M. (2009) 'Molecular origins of cancer: Molecular basis of colorectal cancer', *New England Journal of Medicine*, 361(25), 2449-60.

Marx, V. (2004) 'RNA quality: defining the good, the bad and the ugly', *Genomics Proteomics*, 4, 14-21.

Matsuoka, L., Wortsman, J., Haddad, J. and Hollis, B. (1989) 'In vivo threshold for cutaneous synthesis of vitamin D3', *The Journal of laboratory and clinical medicine*, 114(3), 301-305.

Matsuoka, L. Y., Wortsman, J., Haddad, J. G., Kolm, P. and Hollis, B. W. (1991) 'Racial pigmentation and the cutaneous synthesis of vitamin D', *Archives of dermatology*, 127(4), 536-538.

Matusiak, D., Murillo, G., Carroll, R. E., Mehta, R. G. and Benya, R. V. (2005) 'Expression of vitamin D receptor and 25-hydroxyvitamin D3-1{alpha}-hydroxylase in normal and malignant human colon', *Cancer Epidemiology, Biomarkers & Prevention*, 14(10), 2370-6.

Meyer, M. B., Goetsch, P. D. and Pike, J. W. (2012) 'VDR/RXR and TCF4/-catenin cistromes in colonic cells of colorectal tumor origin: impact on c-FOS and c-MYC gene expression', *Molecular Endocrinology*, 26(1), 37-51.

Milacic, M., Haw, R., Rothfels, K., Wu, G., Croft, D., Hermjakob, H., D'Eustachio, P. and Stein, L. (2012) 'Annotating Cancer Variants and Anti-Cancer Therapeutics in Reactome', *Cancers*, 4(4), 1180-1211.

Mizoue, T. (2004) 'Ecological study of solar radiation and cancer mortality in Japan', *Health Physics*, 87(5), 532-8.

MOH, N. (2010) *Cancer Projections: Incidence 2004-08 to 2014-18* [online], available: http://www.health.govt.nz/publication/cancer-projections-incidence-2004-08-2014-18 [accessed 29 July 2015].

MOH, N. (2011a) *Cancer: Major sites by DHB region of residence 2006-2008* [online], available: http://www.health.govt.nz/system/files/documents/publications/cancer-major-sites-dhb2006-2008.pdf [accessed 2015].

MOH, N. (2011b) *Management of Early Colorectal Cancer* [online], available: http://www.health.govt.nz/system/files/documents/publications/early-management-colorectal-cancer-guideline.pdf [accessed 2015].

MOH, N. (2013) *Cancer: New registrations and deaths 2010* [online], available: http://www.health.govt.nz/system/files/documents/publications/cancer-new-registrations-deaths-2010-aug13.pdf [accessed 2014].

MOH, N. (2014) *Cancer: New registrations and deaths 2011* [online], available: http://www.health.govt.nz/publication/cancer-new-registrations-and-deaths-2011 [accessed 2014].

MOH, N. (2015) *Mortality 2012 online tables* [online], available: http://www.health.govt.nz/publication/mortality-2012-online-tables [accessed 2015].

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P., Elbourne, D., Egger, M. and Altman, D. G. (2010) 'CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials', *Journal of Clinical Epidemiology*, 63(8), e1-e37.

Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M. and Laurila, E. (2003) 'PGC-1 α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes', *Nature Genetics*, 34(3), 267-273.

Munro, F. (2015) *Thesis_Code* [online], available: https://github.com/FrancescaMunro/Thesis_Code [accessed 2015].

Nam, D. and Kim, S.-Y. (2008) 'Gene-set approach for expression pattern analysis', *Briefings in Bioinformatics*, 9(3), 189-197.

Ng, K., Meyerhardt, J. A., Wu, K., Feskanich, D., Hollis, B. W., Giovannucci, E. L. and Fuchs, C. S. (2008) 'Circulating 25-hydroxyvitamin d levels and survival in patients with colorectal cancer', *Journal of Clinical Oncology*, 26(18), 2984-91.

NHMRC, A. G. and MOH, N. (2005) *Nutrient Reference Values for Australia and New Zealand* [online], available: https://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/n35.pdf [accessed 2014].

Norman, A. W., Roth, J. and Orci, L. (1982) 'The Vitamin D Endocrine System: Steroid Metabolism, Hormone Receptors, and Biological Response (Calcium Binding Proteins)*', *Endocrine Reviews*, 3(4), 331-366.

Norval, M., Bjorn, L. O. and de Gruijl, F. R. (2010) 'Is the action spectrum for the UV-induced production of previtamin D3 in human skin correct?', *Photochemical & Photobiological Sciences*, 9(1), 11-7.

O'Connell, J. B., Maggard, M. A. and Ko, C. Y. (2004) 'Colon cancer survival rates with the new American Joint Committee on Cancer sixth edition staging', *Journal Of The National Cancer Institute*, 96(19), 1420-1425.

Ordonez-Moran, P., Larriba, M. J., Palmer, H. G., Valero, R. A., Barbachano, A., Dunach, M., de Herreros, A. G., Villalobos, C., Berciano, M. T., Lafarga, M. and Munoz, A. (2008) 'RhoA-ROCK and p38MAPK-MSK1 mediate vitamin D effects on gene expression, phenotype, and Wnt pathway in colon cancer cells', *Journal of Cell Biology*, 183(4), 697-710.

Ordonez-Moran, P., Larriba, M. J., Pendas-Franco, N., Aguilera, O., Gonzalez-Sancho, J. M. and Munoz, A. (2005) 'Vitamin D and cancer: an update of in vitro and in vivo data', *Frontiers in Bioscience*, 10, 2723-49.

Otani, T., Iwasaki, M., Sasazuki, S., Inoue, M., Tsugane, S. and Japan Public Health Center-Based Prospective Study, G. (2007) 'Plasma vitamin D and risk of colorectal cancer: the Japan Public Health Center-Based Prospective Study', *British Journal of Cancer*, 97(3), 446-51.

Palmer, H. G., Gonzalez-Sancho, J. M., Espada, J., Berciano, M. T., Puig, I., Baulida, J., Quintanilla, M., Cano, A., de Herreros, A. G., Lafarga, M. and Munoz, A. (2001) 'Vitamin D(3) promotes the differentiation of colon carcinoma cells by the induction of E-cadherin and the inhibition of beta-catenin signaling', *Journal of Cell Biology*, 154(2), 369-87.

Pálmer, H. G., Larriba, M. J., García, J. M., Ordóñez-Morán, P., Peña, C., Peiró, S., Puig, I., Rodríguez, R., de la Fuente, R. and Bernad, A. (2004) 'The transcription factor SNAIL represses vitamin D receptor expression and responsiveness in human colon cancer', *Nature medicine*, 10(9), 917-919.

Palmer, H. G., Sanchez-Carbayo, M., Ordonez-Moran, P., Larriba, M. J., Cordon-Cardo, C. and Munoz, A. (2003) 'Genetic signatures of differentiation induced by 1alpha,25-dihydroxyvitamin D3 in human colon cancer cells', *Cancer Research*, 63(22), 7799-806.

Pálmer, H. G., Sánchez-Carbayo, M., Ordóñez-Morán, P., Larriba, M. J., Cordón-Cardó, C. and Muñoz, A. (2003) 'Genetic signatures of differentiation induced by 1 α, 25-dihydroxyvitamin D3 in human colon cancer cells', *Cancer Research*, 63(22), 7799-7806.

Pan, Y.-Z., Gao, W. and Yu, A.-M. (2009) 'MicroRNAs regulate CYP3A4 expression via direct and indirect targeting', *Drug metabolism and disposition*, 37(10), 2112-2117.

Pearson, K. (1901) 'On lines and planes of closest fit to systems of points in space.', *Philosophical Magazine*, 2(11), 559-572.

Pendas-Franco, N., Aguilera, O., Pereira, F., González-Sancho, J. M. and Munoz, A. (2008a) 'Vitamin D and Wnt/ β -catenin pathway in colon cancer: role and regulation of DICKKOPF genes', *Anticancer Research*, 28(5A), 2613-2623.

Pendas-Franco, N., Garcia, J. M., Pena, C., Valle, N., Palmer, H. G., Heinaniemi, M., Carlberg, C., Jimenez, B., Bonilla, F., Munoz, A. and Gonzalez-Sancho, J. M. (2008b) 'DICKKOPF-4 is induced by TCF/beta-catenin and upregulated in human colon cancer, promotes tumour cell invasion and angiogenesis and is repressed by 1alpha,25-dihydroxyvitamin D3', *Oncogene*, 27(32), 4467-77.

Pereira, F., Larriba, M. J. and Munoz, A. (2012) 'Vitamin D and colon cancer', *Endocr Relat Cancer*, 19(3), R51-71.

Pike, J. W. (2011) 'Genome-wide principles of gene regulation by the vitamin D receptor and its activating ligand', *Molecular and Cellular Endocrinology*, 347(1–2), 3-10.

Pike, J. W., Meyer, M. B. and Bishop, K. A. (2012) 'Regulation of target gene expression by the vitamin D receptor - an update on mechanisms', *Reviews in Endocrine and Metabolic Disorders*, 13(1), 45-55.

Prentice, A. (2013) 'Nutritional rickets around the world', *Journal of Steroid Biochemistry & Molecular Biology*, 136, 201-6.

R-Core-Team (2014) *R: A language and environment for statistical computing.* [online], available: http://www.R-project.org/. [accessed 2014].

Rebhan, M., Chalifa-Caspi, V., Prilusky, J. and Lancet, D. (1997) 'GeneCards: integrating information about genes, proteins and diseases', *Trends in Genetics*, 13(4), 163.

Reichel, H., Koeffler, H. P. and Norman, A. W. (1989) 'The role of the vitamin D endocrine system in health and disease', *The New England journal of medicine*, 320(15), 980-991.

Reid, I. R., Gallagher, D. J. and Bosworth, J. (1986) 'Prophylaxis against vitamin D deficiency in the elderly by regular sunlight exposure', *Age Ageing*, 15(1), 35-40.

Riedt, C. S., Cifuentes, M., Stahl, T., Chowdhury, H. A., Schlussel, Y. and Shapses, S. A. (2005) 'Overweight postmenopausal women lose bone with moderate weight reduction and 1 g/day calcium intake', *Journal of Bone & Mineral Research*, 20(3), 455-63.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W. and Smyth, G. K. (2015) 'limma powers differential expression analyses for RNA-sequencing and microarray studies', *Nucleic Acids Research*, 1-13.

Rockell, J., Skeaff, C., Williams, S. and Green, T. (2006) 'Serum 25-hydroxyvitamin D concentrations of New Zealanders aged 15 years and older', *Osteoporosis International*, 17(9), 1382-1389.

Rosenstreich, S. J., Rich, C. and Volwiler, W. (1971) 'Deposition in and release of vitamin D3 from body fat: evidence for a storage site in the rat', *Journal of Clinical Investigation*, 50(3), 679-87.

Ross, A. C., Manson, J. E., Abrams, S. A., Aloia, J. F., Brannon, P. M., Clinton, S. K., Durazo-Arvizu, R. A., Gallagher, J. C., Gallo, R. L., Jones, G., Kovacs, C. S., Mayne, S. T., Rosen, C. J. and Shapses, S. A. (2011) 'The 2011 report on dietary reference intakes for calcium and vitamin D from the Institute of Medicine: what clinicians need to know', *Journal of Clinical Endocrinology & Metabolism*, 96(1), 53-8.

Schmiedlin-Ren, P., Thummel, K. E., Fisher, J. M., Paine, M. F. and Watkins, P. B. (2001) 'Induction of CYP3A4 by 1 alpha,25-dihydroxyvitamin D3 is human cell line-specific and is unlikely to involve pregnane X receptor', *Drug Metabolism & Disposition*, 29(11), 1446-53.

Schulz, K. F., Altman, D. G. and Moher, D. (2010) 'CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials', *BMC medicine*, 8(1), 18.

Schurch, M. A., Rizzoli, R., Slosman, D., Vadas, L., Vergnaud, P. and Bonjour, J. P. (1998) 'Protein supplements increase serum insulin-like growth factor-I levels and attenuate proximal femur bone loss in patients with recent hip fracture. A randomized, double-blind, placebo-controlled trial', *Annals of Internal Medicine*, 128(10), 801-9.

Serpa, J., Caiado, F., Carvalho, T., Torre, C., Gonçalves, L. G., Casalou, C., Lamosa, P., Rodrigues, M., Zhu, Z. and Lam, E. W. (2010) 'Butyrate-rich colonic microenvironment is a relevant selection factor for metabolically adapted tumor cells', *Journal of Biological Chemistry*, 285(50), 39211-39223.

Sheinin, Y., Kaserer, K., Wrba, F., Wenzl, E., Kriwanek, S., Peterlik, M. and Cross, H. S. (2000) 'In situ mRNA hybridization analysis and immunolocalization of the vitamin D receptor in normal and carcinomatous human colonic mucosa: relation to epidermal growth factor receptor expression', *Virchows Archiv*, 437(5), 501-507.

Shrapnel, W. and Truswell, S. (2006) 'Vitamin D deficiency in Australia and New Zealand: what are the dietary options?', *Nutrition & Dietetics*, 63(4), 206-212.

Smith, J. E. and Goodman, D. S. (1971) 'The turnover and transport of vitamin D and of a polar metabolite with the properties of 25-hydroxycholecalciferol in human plasma', *Journal of Clinical Investigation*, 50(10), 2159.

Smyth, G. K. (2004) 'Linear models and empirical bayes methods for assessing differential expression in microarray experiments', *Statistical applications in genetics and molecular biology*, 3(1), 1-25.

Spang, R., Zuzan, H., West, M., Nevins, J., Blanchette, C. and Marks, J. R. (2002) 'Prediction and uncertainty in the analysis of gene expression profiles', *In silico biology*, 2(3), 369-381.

Stambolsky, P., Tabach, Y., Fontemaggi, G., Weisz, L., Maor-Aloni, R., Siegfried, Z., Shiff, I., Kogan, I., Shay, M., Kalo, E., Blandino, G., Simon, I., Oren, M. and Rotter, V. (2010) 'Modulation of the vitamin D3 response by cancer-associated mutant p53.[Erratum appears in Cancer Cell. 2010 May 18;17(5):523 Note: Sigfried, Zahava [corrected to Siegfried, Zehava]]', *Cancer Cell*, 17(3), 273-85.

Su, L.-K., Kinzler, K. W., Vogelstein, B., Preisinger, A. C., Moser, A. R., Luongo, C., Gould, K. A. and Dove, W. F. (1992) 'Multiple intestinal neoplasia caused by a mutation in the murine homolog of the APC gene', *Science*, 256(5057), 668-670.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R. and Lander, E. S. (2005) 'Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles', *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), 15545-15550.

Tangrea, J., Helzlsouer, K., Pietinen, P., Taylor, P., Hollis, B., Virtamo, J. and Albanes, D. (1997) 'Serum levels of vitamin D metabolites and the subsequent risk of colon and rectal cancer in Finnish men', *Cancer Causes & Control*, 8(4), 615-625.

Thompson, P. D., Jurutka, P. W., Kerr Whitfield, G., Myskowski, S. M., Eichhorst, K. R., Encinas Dominguez, C., Haussler, C. A. and Haussler, M. R. (2002) 'Liganded VDR induces CYP3A4 in small intestinal and colon cancer cells via DR3 and ER6 vitamin D responsive elements', *Biochemical and Biophysical Research Communications*, 299(5), 730-738.

Thorne, J. and Campbell, M. J. (2008) 'The vitamin D receptor in cancer', *Proceedings of the Nutrition Society*, 67(02), 115-127.

Thummel, K. E., Brimer, C., Yasuda, K., Thottassery, J., Senn, T., Lin, Y., Ishizuka, H., Kharasch, E., Schuetz, J. and Schuetz, E. (2001) 'Transcriptional control of intestinal cytochrome P-4503A by 1alpha,25-dihydroxy vitamin D3', *Molecular Pharmacology*, 60(6), 1399-406.

Touvier, M., Chan, D. S., Lau, R., Aune, D., Vieira, R., Greenwood, D. C., Kampman, E., Riboli, E., Hercberg, S. and Norat, T. (2011) 'Meta-analyses of vitamin D intake, 25-hydroxyvitamin D status, vitamin D receptor polymorphisms, and colorectal cancer risk', *Cancer Epidemiology, Biomarkers & Prevention*, 20(5), 1003-16.

Tzotzas, T., Papadopoulou, F. G., Tziomalos, K., Karras, S., Gastaris, K., Perros, P. and Krassas, G. E. (2010) 'Rising serum 25-hydroxy-vitamin D levels after weight loss in obese women correlate with improvement in insulin resistance', *Journal of Clinical Endocrinology & Metabolism*, 95(9), 4251-7.

University of Auckland, N. *Enhanced Recovery After Surgery: What is it and is it worth the trouble?* [online], available: https://www.google.co.nz/?gws_rd=ssl -q=length+of+day+stay+for+bowel+surgery+patients+in+NZ [accessed 2015].

University of Auckland, N. (2012) *Vitamin D assessment (ViDA) study* [online], available: https://www.fmhs.auckland.ac.nz/en/soph/about/our-

departments/epidemiology-and-biostatistics/research/vida-study.html [accessed 2014].

US Department of Health, H. S. (2009) *Common terminology criteria for adverse events (CTCAE) version 4.0* [online], available: http://evs.nci.nih.gov/ftp1/CTCAE/CTCAE_4.03_2010-06-14_QuickReference_5x7.pdf [accessed 2014].

Verstuyf, A., Carmeliet, G., Bouillon, R. and Mathieu, C. (2010) 'Vitamin D: a pleiotropic hormone', *Kidney International*, 78(2), 140-5.

von Hurst, P. R., Stonehouse, W. and Coad, J. (2010) 'Vitamin D status and attitudes towards sun exposure in South Asian women living in Auckland, New Zealand', *Public Health Nutrition*, 13(4), 531-6.

Wactawski-Wende, J., Kotchen, J. M., Anderson, G. L., Assaf, A. R., Brunner, R. L., O'Sullivan, M. J., Margolis, K. L., Ockene, J. K., Phillips, L., Pottern, L., Prentice, R. L., Robbins, J., Rohan, T. E., Sarto, G. E., Sharma, S., Stefanick, M. L., Van Horn, L., Wallace, R. B., Whitlock, E., Bassford, T., Beresford, S. A., Black, H. R., Bonds, D. E., Brzyski, R. G., Caan, B., Chlebowski, R. T., Cochrane, B., Garland, C., Gass, M., Hays, J., Heiss, G., Hendrix, S. L., Howard, B. V., Hsia, J., Hubbell, F. A., Jackson, R. D., Johnson, K. C., Judd, H., Kooperberg, C. L., Kuller, L. H., LaCroix, A. Z., Lane, D. S., Langer, R. D., Lasser, N. L., Lewis, C. E., Limacher, M. C., Manson, J. E. and Women's Health Initiative, I. (2006) 'Calcium plus vitamin D supplementation and the risk of colorectal cancer.[Erratum appears in N Engl J Med. 2006 Mar 9;354(10):1102]', *New England Journal of Medicine*, 354(7), 684-96.

Waldecker, M., Kautenburger, T., Daumann, H., Busch, C. and Schrenk, D. (2008) 'Inhibition of histone-deacetylase activity by short-chain fatty acids and some polyphenol metabolites formed in the colon', *The Journal of Nutritional Biochemistry*, 19(9), 587-593.

Wang, E. (2010) *Cancer systems biology*, CRC Press: Florida.

Wang, T. T., Tavera-Mendoza, L. E., Laperriere, D., Libby, E., MacLeod, N. B., Nagai, Y., Bourdeau, V., Konstorum, A., Lallemant, B., Zhang, R., Mader, S. and White, J. H. (2005) 'Large-scale in silico and microarray-based identification of direct 1,25-dihydroxyvitamin D3 target genes', *Molecular Endocrinology*, 19(11), 2685-95.

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Liaw, W. H. A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M. and Venables, B. (2015) *gplots: Various R Programming Tools for Plotting Data* [online], available: http://CRAN.R-project.org/package=gplots [accessed 2013].

Webb, A. R. and Holick, M. F. (1988) 'The role of sunlight in the cutaneous production of vitamin D3', *Annual Review of Nutrition*, 8, 375-99.

Welch, B. L. (1947) 'The Generalization of `Student's' Problem when Several Different Population Variances are Involved', *Biometrika*, 34(1/2), 28-35.

Welch, C., Santra, M. K., El-Assaad, W., Zhu, X., Huber, W. E., Keys, R. A., Teodoro, J. G. and Green, M. R. (2009) 'Identification of a protein, G0S2, that lacks Bcl-2 homology domains and interacts with and antagonizes Bcl-2', *Cancer Research*, 69(17), 6782-6789.

Welsh, E. A., Eschrich, S. A., Berglund, A. E. and Fenstermacher, D. A. (2013) 'Iterative rank-order normalization of gene expression microarray data', *BMC Bioinformatics*, 14(1), 153.

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J. A., Marks, J. R. and Nevins, J. R. (2001) 'Predicting the clinical status of human breast cancer by using gene expression profiles', *Proceedings of the National Academy of Sciences*, 98(20), 11462-11467.

WHO (2000) *Pathology and genetics of tumours of the digestive system*, IARC press: Lyon.

Wilcoxon, F. (1945) 'Individual Comparisons by Ranking Methods', *Biometrics Bulletin*, 1(6), 80-83.

Witham, M. D., Dove, F. J., Dryburgh, M., Sugden, J. A., Morris, A. D. and Struthers, A. D. (2010) 'The effect of different doses of vitamin D(3) on markers of vascular health in patients with type 2 diabetes: a randomised controlled trial', *Diabetologia*, 53(10), 2112-9.

Wood, R. J., Tchack, L., Angelo, G., Pratt, R. E. and Sonna, L. A. (2004) 'DNA microarray analysis of vitamin D-induced gene expression in a human colon carcinoma cell line', *Physiological Genomics*, 17(2), 122-9.

Woolcott, C. G., Wilkens, L. R., Nomura, A. M., Horst, R. L., Goodman, M. T., Murphy, S. P., Henderson, B. E., Kolonel, L. N. and Le Marchand, L. (2010) 'Plasma 25-hydroxyvitamin D levels and the risk of colorectal cancer: the multiethnic cohort study', *Cancer Epidemiology, Biomarkers & Prevention*, 19(1), 130-4.

Wortsman, J., Matsuoka, L. Y., Chen, T. C., Lu, Z. and Holick, M. F. (2000) 'Decreased bioavailability of vitamin D in obesity', *The American journal of clinical nutrition*, 72(3), 690-693.

Wu, K., Feskanich, D., Fuchs, C. S., Willett, W. C., Hollis, B. W. and Giovannucci, E. L. (2007) 'A nested case control study of plasma 25-hydroxyvitamin D concentrations and risk of colorectal cancer', *Journal Of The National Cancer Institute*, 99(14), 1120-9.

Wu-Wong, J. R. (2007) 'The potential for vitamin D receptor activation in cardiovascular research', *Expert Opinion on Investigational Drugs*, 16(4), 407-411.

Wu-Wong, J. R. (2009) 'Potential for vitamin D receptor agonists in the treatment of cardiovascular disease', *British Journal of Pharmacology*, 158(2), 395-412.

Xu, H., Posner, G. H., Stevenson, M. and Campbell, F. C. (2010) 'Apc(MIN) modulation of vitamin D secosteroid growth control', *Carcinogenesis*, 31(8), 1434-41.

Yaari, G., Bolen, C. R., Thakar, J. and Kleinstein, S. H. (2013) 'Quantitative set analysis for gene expression: a method to quantify gene set differential expression including gene-gene correlations', *Nucleic Acids Research*, gkt660.

Zgaga, L., Theodoratou, E., Farrington, S. M., Din, F. V. N., Ooi, L. Y., Glodzik, D., Johnston, S., Tenesa, A., Campbell, H. and Dunlop, M. G. (2014) 'Plasma vitamin D concentration influences survival outcome after a diagnosis of colorectal cancer', *Journal of Clinical Oncology*, 32(23), 2430-2439.

Zheng, W., Wong, K. E., Zhang, Z., Dougherty, U., Mustafi, R., Kong, J., Deb, D. K., Zheng, H., Bissonnette, M. and Li, Y. C. (2012) 'Inactivation of the vitamin D receptor in APC(min/+) mice reveals a critical role for the vitamin D receptor in intestinal tumor growth', *International Journal of Cancer*, 130(1), 10-9.

Zittermann, A., Frisch, S., Berthold, H. K., Gotting, C., Kuhn, J., Kleesiek, K., Stehle, P., Koertke, H. and Koerfer, R. (2009) 'Vitamin D supplementation enhances the beneficial effects of weight loss on cardiovascular disease risk markers', *American Journal of Clinical Nutrition*, 89(5), 1321-7.

APPENDIX 1

Randomisation list

Study Code	Patient Identifier	Study Medication	Date Capsules Prepared
VDS	01	Group 1	
VDS	02	Group 2	
VDS	03	Group 1	
VDS	04	Group 2	
VDS	05	Group 1	
VDS	06	Group 2	
VDS	07	Group 1	
VDS	08	Group 2	
VDS	09	Group 2	
VDS	10	Group 1	31/03/11
VDS	11	Group 2	
VDS	12	Group 2	
VDS	13	Group 1	
VDS	14	Group 1	
VDS	15	Group 2	
VDS	16	Group 1	
VDS	17	Group 2	
VDS	18	Group 1	
VDS	19	Group 2	
VDS	20	Group 1	20/07/11
VDS	21	Group 1	
VDS	22	Group 1	
VDS	23	Group 2	
VDS	24	Group 2	
VDS	25	Group 2	
VDS	26	Group 1	
VDS	27	Group 2	
VDS	28	Group 2	
VDS	29	Group 1	
VDS	30	Group 1	20/07/11

Study Code	Patient Identifier	Study Medication	Date Capsules Prepared
VDS	31	Group 2	
VDS	32	Group 1	
VDS	33	Group 2	
VDS	34	Group 2	
VDS	35	Group 2	
VDS	36	Group 1	
VDS	37	Group 1	
VDS	38	Group 1	
VDS	39	Group 1	
VDS	40	Group 2	21/03/12
VDS	41	Group 1	
VDS	42	Group 1	
VDS	43	Group 2	
VDS	44	Group 1	
VDS	45	Group 2	
VDS	46	Group 2	
VDS	47	Group 1	
VDS	48	Group 2	
VDS	49	Group 2	
VDS	50	Group 1	21/03/12
VDS	51	Group 2	
VDS	52	Group 1	
VDS	53	Group 2	
VDS	54	Group 2	
VDS	55	Group 2	
VDS	56	Group 1	
VDS	57	Group 1	
VDS	58	Group 1	
VDS	59	Group 1	
VDS	60	Group 2	7/08/12

APPENDIX 2

VDRE gene list

AAMP	(Palmer et al., 2003)
ABCA11P	(Haussler et al., 2013)
ABCB1	(Haussler et al., 2013, Meyer et al., 2012)
ABCD1	(Haussler et al., 2013)
ACTA2	(Mapes et al., 2014)
ACTB	(Palmer et al., 2003)
ACTG2	(Mapes et al., 2014)
ACTN1	(Palmer et al., 2003)
ADAM17	(Palmer et al., 2003)
AIG1	(Wang et al., 2005)
AKAP12	(Mapes et al., 2014, Palmer et al., 2003)
AKR1C1	(Wang et al., 2005)
AKR1C2	(Wang et al., 2005)
AKT1	(Palmer et al., 2003)
ALOX5	(Carlberg and Campbell, 2012)
ALPI	(Haussler et al., 2013)
ALPL	(Ebert et al., 2006)
ALPPL2	(Haussler et al., 2013)
ALS2	(Wang et al., 2005)
AMOTL1	(Wang et al., 2005)
ANKMY2	(Wang et al., 2005)
ANXA6	(Palmer et al., 2003)
AP1G1	(Wood et al., 2004)
AP1S1	(Palmer et al., 2003)
APOBEC1	(Palmer et al., 2003)
APOE	(Mapes et al., 2014)
AREG	(Wood et al., 2004)
ARF6	(Palmer et al., 2003)
ARH	(Wang et al., 2005)
ARHGAP5	(Palmer et al., 2003)
ARHGAP10	(Wang et al., 2005)
ARHGAP29	(Palmer et al., 2003)
ARHGAP35	(Palmer et al., 2003)
ARHGDIA	(Palmer et al., 2003)
ARHGEF2	(Palmer et al., 2003)
ARL8	(Wang et al., 2005)
ARL5B	(Mapes et al., 2014)
ARL2BP	(Wang et al., 2005)
ASCL2	(Palmer et al., 2003)
ATF3	(Palmer et al., 2003)
ATP8B1	(Wang et al., 2005)

ATP2C2	(Haussler et al., 2013)
AURKA	(Palmer et al., 2003)
AXL	(Palmer et al., 2003)
B4GALT6	(Wang et al., 2005)
BAG1	(Díaz et al., 2000, Barnes et al., 2005)
BAK1	(Díaz et al., 2000, Barnes et al., 2005)
BAX	(Palmer et al., 2003)
BCAR3	(Palmer et al., 2003)
BENE	(Wang et al., 2005)
BGLAP	(Ebert et al., 2006)
BIRC5	(Liu et al., 2010)
BMP6	(Wang et al., 2005)
BSG	(Palmer et al., 2003)
BTLA	(Carlberg and Campbell, 2012)
C1orf131	(Mapes et al., 2014)
C3orf52	(Mapes et al., 2014)
C9orf89	(Carlberg and Campbell, 2012)
C12orf36	(Mapes et al., 2014)
C14orf149	(Wang et al., 2005)
C20orf82	(Wang et al., 2005)
C1QB	(Mapes et al., 2014)
C1QC	(Mapes et al., 2014)
C1QTNF5	(Mapes et al., 2014)
CA2	(Wang et al., 2005, Palmer et al., 2003)
CA9	(Meyer et al., 2012)
CA12	(Kriebitzsch et al., 2009, Wood et al., 2004)
CALB1	(Meyer et al., 2012)
CALB2	(Palmer et al., 2003)
CAMP	(Gombart et al., 2005, Kriebitzsch et al., 2009))
CAPN7	(Wang et al., 2005)
CARHSP1	(Wang et al., 2005)
CaSR	(Fetahu et al., 2014)
CAV2	(Palmer et al., 2003)
CBLB	(Palmer et al., 2003)
CCDC85B	(Palmer et al., 2003)
CCL20	(Palmer et al., 2003)
CCNC	(Ebert et al., 2006)
CCNF	(Kriebitzsch et al., 2009, Palmer et al., 2003)
CD14	(Ebert et al., 2006)
CD55	(Palmer et al., 2003)
CD86	(Wang et al., 2005)
CD151	(Palmer et al., 2003)
CD200	(Carlberg and Campbell, 2012)
CDC34	(Haussler et al., 2013)
CDH1	(Palmer et al., 2003)
CDK16	(Palmer et al., 2003)
CDKN1A	(Carlberg and Campbell, 2012)

CDKN1B	(Carlberg and Campbell, 2012)
CDX-2	(Haussler et al., 2013)
CEACAM6	(Haussler et al., 2013)
CEBPA	(Haussler et al., 2013)
CEBPB	(Palmer et al., 2003, Ebert et al., 2006)
CEBPG	(Palmer et al., 2003)
CELF1	(Palmer et al., 2003)
CENTG3	(Wang et al., 2005)
CFL2	(Wang et al., 2005)
CHST7	(Wang et al., 2005)
CHST11	(Wang et al., 2005)
CITED2	(Palmer et al., 2003)
CKS1B	(Ebert et al., 2006)
CLCF1	(Mapes et al., 2014)
CLDN2	(Haussler et al., 2013, Fujita et al., 2008)
CLDN12	(Fujita et al., 2008)
CLDN23	(Mapes et al., 2014)
CLGN	(Wang et al., 2005)
CLU	(Palmer et al., 2003)
CMKOR1	(Wang et al., 2005)
CNGB1	(Wang et al., 2005)
CNOT3	(Palmer et al., 2003)
COL1A1	(Wang et al., 2005, Ebert et al., 2006)
COL3A1	(Mapes et al., 2014)
COL13A1	(Wang et al., 2005)
COL1A2	(Mapes et al., 2014)
COL5A2	(Mapes et al., 2014)
COL4A5	(Mapes et al., 2014)
CORO1C	(Pereira et al., 2012)
COX-2	(Hummel et al., 2014)
CPVL	(Mapes et al., 2014)
CRABP2	(Palmer et al., 2003)
CRLF1	(Palmer et al., 2003)
CSF1R	(Mapes et al., 2014)
CST1	(Palmer et al., 2003)
CST5	(Palmer et al., 2003)
CST6	(Wang et al., 2005)
CTAGE5	(Palmer et al., 2003)
CTDSPL	(Palmer et al., 2003)
CTSD	(Wang et al., 2005)
CXCL14	(Mapes et al., 2014, Wang et al., 2005)
CXorf6	(Wang et al., 2005)
CYP1A1	(Haussler et al., 2013)
CYP24A1	(Palmer et al., 2003, Haussler et al., 2013, Mapes et al., 2014, Meyer et al., 2012)
CYP2B	(Jurutka et al., 2005)
CYP26B1	(Wang et al., 2005)
CYP27B1	(Meyer et al., 2012)

CYP2B6	(Palmer et al., 2003, Meyer et al., 2012)
CYP2C9	(Ebert et al., 2006)
CYP2R1	(Ahn et al., 2010)
CYP2S1	(Haussler et al., 2013)
CYP3A	(Ebert et al., 2006)
CYP3A4	(Ebert et al., 2006, Thompson et al., 2002)
CYP3A5	(Palmer et al., 2003, Haussler et al., 2013)
CYP3A7	(Wang et al., 2005)
CYP3A43	(Palmer et al., 2003)
CYR61	(Palmer et al., 2003)
DACT2	(Haussler et al., 2013)
DAP	(Palmer et al., 2003)
DBC-1	(Wang et al., 2005)
DDIT4	(Mapes et al., 2014)
DDX39B	(Palmer et al., 2003)
DEFB32	(Haussler et al., 2013)
DEFB4A	(Wang et al., 2005)
DGAT1	(Palmer et al., 2003)
DGAT2	(Wang et al., 2005)
DHFR	(Wang et al., 2005, Palmer et al., 2003)
DKFZp761P0423	(Wang et al., 2005)
DKFZp762A217	(Wang et al., 2005)
DKFZp762C186	(Wang et al., 2005)
DKK1	(Meyer et al., 2012)
DKK4	(Pendas-Franco et al., 2008)
DLST	(Palmer et al., 2003)
DLX2	(Wang et al., 2005)
DMRT2	(Mapes et al., 2014)
DOCK4	(Wang et al., 2005)
DRCTNNB1A	(Wang et al., 2005)
DTR	(Wang et al., 2005)
DUSP5	(Mapes et al., 2014)
DUSP8	(Palmer et al., 2003)
DUSP10	(Palmer et al., 2003)
E2-EPF	(Palmer et al., 2003)
E4BP4	(Palmer et al., 2003)
EDNRB	(Mapes et al., 2014)
EFNA5	(Haussler et al., 2013)
EFTUD1	(Mapes et al., 2014)
EGFR	(Palmer et al., 2003)
EGR3	(Wang et al., 2005)
EID3	(Mapes et al., 2014)
EIF2S3	(Palmer et al., 2003)
EIF3J	(Palmer et al., 2003)
EIF4EBP1	(Kriebitzsch et al., 2009, Palmer et al., 2003)
ELL	(Heikkinen et al., 2011)
EMP1	(Palmer et al., 2003)

EMP3	(Palmer et al., 2003)
ENSG00000229470	(Kriebitzsch et al., 2009)
ENTPD3	(Wang et al., 2005)
EOMES	(Wang et al., 2005)
EPHB4	(Haussler et al., 2013)
EPHX1	(Palmer et al., 2003)
ESR1	(Wang et al., 2005)
EXT1	(Palmer et al., 2003)
EZR	(Pereira et al., 2012)
F3	(Mapes et al., 2014)
FAM38A	(Wang et al., 2005)
FBXO3	(Wang et al., 2005)
FCER1G	(Mapes et al., 2014)
FEM1B	(Wang et al., 2005)
FGD3	(Wang et al., 2005)
FGD6	(Mapes et al., 2014)
FGF23	(Pike et al., 2012)
FGFBP1	(Wang et al., 2005)
FICD	(Palmer et al., 2003)
FLJ12552	(Wood et al., 2004)
FLJ20668	(Wang et al., 2005)
FLJ35024	(Mapes et al., 2014)
FLJ38607	(Wang et al., 2005)
FLNA	(Palmer et al., 2003)
FOS	(Meyer et al., 2012)
FOXF1	(Wang et al., 2005, Palmer et al., 2003)
FOXN2	(Palmer et al., 2003)
FOXO1	(Wang et al., 2005)
FOXO1A	(Wang et al., 2005)
FOXQ1	(Wang et al., 2005)
FRAS1	(Wang et al., 2005)
FSTL1	(Wang et al., 2005)
FTH1	(Palmer et al., 2003)
G6PD	(Haussler et al., 2013)
GADD45A	(Wang et al., 2005, Palmer et al., 2003, Ebert et al., 2006)
GALNT5	(Wang et al., 2005)
GATA-2	(Palmer et al., 2003)
GC	(Ahn et al., 2010)
GEM	(Mapes et al., 2014)
GLT8D4	(Haussler et al., 2013)
GLUL	(Palmer et al., 2003)
GNA13	(Palmer et al., 2003)
GNAQ	(Palmer et al., 2003)
GOLPH2	(Wang et al., 2005)
GOS2	(Palmer et al., 2003)
GPRC5A	(Palmer et al., 2003)
GPRC5B	(Wang et al., 2005)

GPX1	(Palmer et al., 2003)
GPX2	(Ebert et al., 2006)
GPX4	(Palmer et al., 2003)
GREM1	(Wang et al., 2005)
GSN	(Palmer et al., 2003)
GSTA4	(Kriebitzsch et al., 2009)
H1FX	(Palmer et al., 2003)
H2AFX	(Palmer et al., 2003)
HIF1A	(Wang et al., 2005)
HIRA	(Palmer et al., 2003)
HIST4H4	(Palmer et al., 2003)
HIVEP3	(Wang et al., 2005)
HMGCS1	(Palmer et al., 2003)
HNRPH1	(Alvarez-Diaz et al., 2012)
HR	(Wang et al., 2005)
HRAS	(Palmer et al., 2003)
HRASLS2	(Mapes et al., 2014)
HSD3B1	(Palmer et al., 2003)
HSD17B2	(Wang et al., 2005)
HSD17B12	(Wang et al., 2005, Palmer et al., 2003)
HSPA13	(Palmer et al., 2003)
HSPB1	(Palmer et al., 2003)
HT011	(Wang et al., 2005)
HTR6	(Palmer et al., 2003)
HUMPPA	(Wang et al., 2005)
IFIT2	(Wang et al., 2005)
IFITM1	(Palmer et al., 2003)
IFITM2	(Palmer et al., 2003)
IFNG	(Ebert et al., 2006)
IGF2	(Pereira et al., 2012)
IGFBP1	(Carlberg and Dunlop, 2006)
IGFBP2	(Palmer et al., 2003, Carlberg and Dunlop, 2006)
IGFBP3	(Kriebitzsch et al., 2009, Palmer et al., 2003, Carlberg and Dunlop, 2006)
IGFBP4	(Palmer et al., 2003)
IGFBP5	(Carlberg and Dunlop, 2006)
IGFBP6	(Palmer et al., 2003, Carlberg and Dunlop, 2006)
IL2	(Ebert et al., 2006)
IL4	(Ebert et al., 2006)
IL5	(Ebert et al., 2006)
IL10	(Carlberg and Campbell, 2012)
IL12B	(Carlberg and Campbell, 2012)
IL1R1	(Wang et al., 2005)
IL1RL1	(Wang et al., 2005)
IL6ST	(Palmer et al., 2003)
IRF1	(Wang et al., 2005)
IRF8	(Haussler et al., 2013)
ITM2A	(Mapes et al., 2014)

ITPKB	(Wang et al., 2005)
JMJD2B	(Wang et al., 2005)
JUN	(Palmer et al., 2003)
JUNB	(Wang et al., 2005, Palmer et al., 2003, Haussler et al., 2013)
JUND	(Palmer et al., 2003)
KATNA1	(Palmer et al., 2003)
KCNK9	(Wang et al., 2005)
KCNMB3	(Wang et al., 2005)
KDM6B	(Pereira et al., 2011)
KHSRP	(Pereira et al., 2012)
KIAA0256	(Wang et al., 2005)
KIAA2018	(Wang et al., 2005)
KIF22	(Palmer et al., 2003)
KIF3C	(Wang et al., 2005)
KLHL4	(Wang et al., 2005)
KLK6	(Wang et al., 2005, Kriebitzsch et al., 2009, Palmer et al., 2003)
KLK9	(Meyer et al., 2012)
KNG1	(Haussler et al., 2013)
KPNA7	(Mapes et al., 2014)
KRT2	(Palmer et al., 2003)
KRT13	(Palmer et al., 2003)
KRT15	(Palmer et al., 2003)
LAMB2	(Palmer et al., 2003)
LAMB3	(Palmer et al., 2003)
LAMP3	(Palmer et al., 2003)
LFNG	(Wang et al., 2005)
LILRB1	(Palmer et al., 2003)
LINCR	(Mapes et al., 2014)
LMNA	(Palmer et al., 2003)
LOC83690	(Wang et al., 2005)
LOC126669	(Wang et al., 2005)
LOC153222	(Wang et al., 2005)
LOC200230	(Wang et al., 2005)
LOC340075	(Wang et al., 2005)
LOC340274	(Mapes et al., 2014)
LOC392871	(Mapes et al., 2014)
LOC400578	(Mapes et al., 2014)
LOC404636	(Wang et al., 2005)
LOC440388	(Wang et al., 2005)
LOC440398	(Wang et al., 2005)
LOC642113	(Mapes et al., 2014)
LOC647450	(Mapes et al., 2014)
LOC652102	(Mapes et al., 2014)
LOC730525	(Mapes et al., 2014)
LOC101929775	(Palmer et al., 2003)
LRP5	(Pike et al., 2007)
LRRFIP1	(Palmer et al., 2003)

LUM	(Mapes et al., 2014)
MAD	(Wang et al., 2005)
MAFB	(Wang et al., 2005)
MAFF	(Palmer et al., 2003)
MAN2C1	(Wang et al., 2005)
MAOA	(Palmer et al., 2003)
MAPK8	(Picotto et al., 2012)
MAPK14	(Picotto et al., 2012)
MAT2A	(Mapes et al., 2014)
MAZ	(Palmer et al., 2003)
MCAM	(Wang et al., 2005)
MCM3	(Palmer et al., 2003)
MCM4	(Palmer et al., 2003)
MCM7	(Carlberg and Campbell, 2012)
MCP	(Wang et al., 2005)
MDM4	(Palmer et al., 2003)
MED9	(Haussler et al., 2013)
MET	(Palmer et al., 2003)
MFI2	(Wang et al., 2005)
MGC10500	(Wang et al., 2005)
MGC26885	(Wang et al., 2005)
MGC45731	(Wang et al., 2005)
MGC102966	(Mapes et al., 2014)
MGLL	(Palmer et al., 2003)
MIR22	(Pereira et al., 2012)
MMP1	(Wang et al., 2005)
MPRG	(Wang et al., 2005)
MR-1	(Wang et al., 2005)
MRPS12	(Wang et al., 2005, Palmer et al., 2003)
MTA1	(Palmer et al., 2003)
MUC5B	(Palmer et al., 2003)
MUC6	(Palmer et al., 2003)
MUT	(Wang et al., 2005)
MX2	(Haussler et al., 2013)
MXRA5	(Mapes et al., 2014)
MYC	(Meyer et al., 2012)
MYH11	(Mapes et al., 2014)
MYO1C	(Palmer et al., 2003)
MYO9B	(Carlberg and Campbell, 2012)
NADSYN1	(Ahn et al., 2010)
NCoR	(Meyer and Pike, 2013)
NCX1	(Pike et al., 2012)
NDUFS8	(Palmer et al., 2003)
NEFL	(Wang et al., 2005)
NELL2	(Alvarez-Diaz et al., 2012)
NES1	(Palmer et al., 2003)
NET1	(Mapes et al., 2014)

NF2	(Palmer et al., 2003)
NFAT5	(Alvarez-Diaz et al., 2012)
NFE2	(Palmer et al., 2003)
NFE2L2	(Ebert et al., 2006)
NINJ1	(Carlberg and Campbell, 2012)
NKSF1	(Ebert et al., 2006)
NME1	(Palmer et al., 2003)
NOPD1	(Wang et al., 2005)
NOX1	(Haussler et al., 2013)
NR1D1	(Palmer et al., 2003)
NR3C1	(Palmer et al., 2003)
OASL	(Mapes et al., 2014)
OCLN	(Palmer et al., 2001)
OGDH	(Palmer et al., 2003)
OGN	(Alvarez-Diaz et al., 2012)
ORM1	(Haussler et al., 2013)
ORM2	(Haussler et al., 2013)
OSR2	(Wang et al., 2005)
OXTR	(Wang et al., 2005)
P4HB	(Palmer et al., 2003)
PADI1	(Meyer et al., 2012)
PAFAH1B3	(Palmer et al., 2003)
PAK6	(Wang et al., 2005)
PARP1	(Pereira et al., 2012)
PAWR	(Palmer et al., 2003)
PAX9	(Wang et al., 2005)
PCDH7	(Wang et al., 2005)
PCNX	(Wang et al., 2005)
PCP4	(Palmer et al., 2003)
PDE2A	(Wang et al., 2005)
PDE4D	(Palmer et al., 2003)
PDGFRA	(Mapes et al., 2014, Leyssens et al., 2013)
15-PGDH	(Hummel et al., 2014)
PGM2L1	(Wang et al., 2005)
PEMT	(Palmer et al., 2003)
PEX3	(Wang et al., 2005)
PIM1	(Mapes et al., 2014)
PINK1	(Wang et al., 2005)
PKM	(Palmer et al., 2003)
PKN1	(Palmer et al., 2003)
PLA2G10	(Wang et al., 2005)
PLAT	(Mapes et al., 2014)
PLEC	(Palmer et al., 2003)
PLEKHO2	(Wang et al., 2005)
PMAIP1	(Palmer et al., 2003)
PMS2	(Palmer et al., 2003)
PODXL	(Wang et al., 2005)

POLD1	(Palmer et al., 2003)
PPARG2	(Ebert et al., 2006)
PPM2C	(Wang et al., 2005)
PPP1R12A	(Wang et al., 2005)
PPP2R1A	(Palmer et al., 2003)
PPP2R5A	(Palmer et al., 2003)
PRK1	(Palmer et al., 2003)
PRKAR1A	(Palmer et al., 2003)
PRKCI	(Palmer et al., 2003)
PRPF19	(Palmer et al., 2003)
PRPS1	(Palmer et al., 2003)
PRSS3	(Palmer et al., 2003)
PRSS22	(Mapes et al., 2014)
PSMB6	(Palmer et al., 2003)
PSMF1	(Wang et al., 2005)
PTGER4	(Wang et al., 2005)
PTH	(Pike et al., 2012)
PTHLH	(Wang et al., 2005, Pike et al., 2007)
PTK7	(Palmer et al., 2003)
PVR	(Palmer et al., 2003)
QSOX1	(Palmer et al., 2003)
RAB2A	(Palmer et al., 2003)
RAB3IP	(Wang et al., 2005)
RACK7	(Palmer et al., 2003)
RARRES1	(Wood et al., 2004)
RASSF3	(Wang et al., 2005)
RDX	(Pereira et al., 2012)
RELA	(Leyssens et al., 2013)
REPS2	(Palmer et al., 2003)
RERE	(Alvarez-Diaz et al., 2012)
ROR1	(Wang et al., 2005)
RP5-1022P6.2	(Mapes et al., 2014)
RRBP1	(Palmer et al., 2003)
RYK	(Palmer et al., 2003)
S100A4	(Mapes et al., 2014, Haussler et al., 2013)
S100A11	(Palmer et al., 2003)
SEC24A	(Palmer et al., 2003)
SEC14L1	(Mapes et al., 2014)
SEL1L	(Palmer et al., 2003)
SEMA3B	(Wang et al., 2005)
SEMA3F	(Wang et al., 2005)
SEMA4B	(Wang et al., 2005)
SERPINA1	(Palmer et al., 2003)
SERPINB1	(Wang et al., 2005)
SFPQ	(Pereira et al., 2012)
SGCE	(Mapes et al., 2014)
SH2D1A	(Palmer et al., 2003)

SH3TC1	(Mapes et al., 2014)
SIGMAR1	(Palmer et al., 2003)
SIL1	(Palmer et al., 2003)
SIVA1	(Palmer et al., 2003)
SKP2	(Ebert et al., 2006)
SLC4A2	(Palmer et al., 2003)
SLC5A3	(Palmer et al., 2003)
SLC6A19	(Palmer et al., 2003)
SCL15A3	(Mapes et al., 2014)
SLC16A1	(Palmer et al., 2003)
SLC20A1	(Mapes et al., 2014)
SLC25A1	(Palmer et al., 2003)
SLC30A1	(Mapes et al., 2014)
SLC34A1	(Pike et al., 2012)
SLC34A2	(Pike et al., 2012)
SLC34A3	(Meyer et al., 2012)
SLC37A2	(Carlberg and Campbell, 2012)
SMAD6	(Palmer et al., 2003)
SMARCA5	(Wang et al., 2005)
SMARCE1	(Pereira et al., 2012)
SNORA1	(Palmer et al., 2003)
SOX9	(Meyer et al., 2012)
SP100	(Heikkinen et al., 2011)
SPP1	(Pike et al., 2012)
SPR-2	(Palmer et al., 2003)
SPRY2	(Palmer et al., 2003)
SRI	(Kriebitzsch et al., 2009)
SRSF1	(Palmer et al., 2003)
SRSF6	(Palmer et al., 2003)
SSH1	(Wang et al., 2005)
ST2	(Palmer et al., 2003)
STAM2	(Palmer et al., 2003)
STAMBP	(Wang et al., 2005)
STAT1	(Heikkinen et al., 2011)
STXBP2	(Palmer et al., 2003)
SULT1A2	(Haussler et al., 2013)
SULT2A1	(Echchgadda et al., 2004)
SULT1C2	(Haussler et al., 2013)
SYK	(Palmer et al., 2003)
SYNGR1	(Wang et al., 2005)
SYNJ2	(Wang et al., 2005)
SYT	(Palmer et al., 2003)
SYT12	(Wang et al., 2005)
T1A-2	(Wang et al., 2005)
TACSTD2	(Mapes et al., 2014)
TAF6L	(Palmer et al., 2003)
TARDBP	(Pereira et al., 2012)

TAX1BP1	(Palmer et al., 2003)
TCEA2	(Palmer et al., 2003)
TDE2L	(Wang et al., 2005)
TDRKH	(Wang et al., 2005)
TEAD1	(Palmer et al., 2003)
TFRC	(Palmer et al., 2003)
TGFB2	(Wang et al., 2005, Haussler et al., 2013)
TGFBR1	(Palmer et al., 2003)
TGIF	(Palmer et al., 2003)
TGM1	(Wang et al., 2005)
THBD	(Heikkinen et al., 2011, Carlberg and Campbell, 2012)
THBS1	(Fernandez-Garcia et al., 2005)
TIEG2	(Wang et al., 2005)
TIF1	(Wang et al., 2005)
TIMP2	(Haussler et al., 2013)
TIMP3	(Haussler et al., 2013)
TIPARP	(Mapes et al., 2014)
TJP1	(Pereira et al., 2012)
TJP2	(Pereira et al., 2012)
TK1	(Palmer et al., 2003)
TL132	(Palmer et al., 2003)
TM4SF1	(Wang et al., 2005)
TM4SF11	(Wang et al., 2005)
TM7SF3	(Wang et al., 2005)
TMEM57	(Mapes et al., 2014)
TMEM119	(Mapes et al., 2014)
TMEPAI	(Wang et al., 2005)
TNFA1P8LS	(Meyer et al., 2012)
TNFSF11	(Pike et al., 2012)
TNFRSF11B	(Pike et al., 2012)
TNFRSF21	(Wang et al., 2005)
TNK1	(Palmer et al., 2003)
TOB1	(Palmer et al., 2003)
TOM40	(Palmer et al., 2003)
TP53	(Palmer et al., 2003)
TPM4	(Mapes et al., 2014)
TPSB2	(Mapes et al., 2014)
TRAF4	(Wang et al., 2005)
TRIB1	(Wang et al., 2005)
TRIO	(Palmer et al., 2003)
TRIM16	(Wang et al., 2005)
TRIM23	(Wang et al., 2005)
TRIM56	(Wang et al., 2005)
TRPC1	(Palmer et al., 2003)
TRPV5	(Pike et al., 2012)
TRPV6	(Wang et al., 2005, Ebert et al., 2006, Haussler et al., 2013, Mapes et al., 2014, Meyer et al., 201
TSKU	(Mapes et al., 2014)

TUBA1A	(Palmer et al., 2003)
TUBB	(Palmer et al., 2003)
TUBB2A	(Palmer et al., 2003)
TULP3	(Palmer et al., 2003)
TXNRD1	(Mapes et al., 2014, Ebert et al., 2006)
TYK2	(Palmer et al., 2003)
TYMS	(Palmer et al., 2003)
UBA1	(Palmer et al., 2003)
UBE1	(Palmer et al., 2003)
UBE2H	(Palmer et al., 2003)
UBE2S	(Palmer et al., 2003)
UBR5	(Palmer et al., 2003)
UCA1	(Mapes et al., 2014)
UGCG	(Palmer et al., 2003)
UGT1A	(Kriebitzsch et al., 2009)
UGT1A2	(Palmer et al., 2003)
UGT1A4	(Palmer et al., 2003)
UPP1	(Mapes et al., 2014, Palmer et al., 2003)
UQCRC1	(Palmer et al., 2003)
UST	(Wang et al., 2005)
VCL	(Palmer et al., 2001)
VDR	(Wang et al., 2005)
VEGFA	(Fernandez-Garcia et al., 2005)
VLDLR	(Mapes et al., 2014)
WDSAM1	(Wang et al., 2005)
YOD1	(Mapes et al., 2014)
ZBTB7A	(Palmer et al., 2003)
ZFYVE9	(Palmer et al., 2003)
ZFAB	(Palmer et al., 2003)
ZKSCAN8	(Palmer et al., 2003)
ZNF43	(Palmer et al., 2003)
ZNF44	(Palmer et al., 2003)
ZNF185	(Palmer et al., 2003)
ZNF296	(Mapes et al., 2014)
ZNF573	(Mapes et al., 2014)
ZYX	(Palmer et al., 2003)

APPENDIX 3

arrayQualityMetrics Report for Microarray Expression Data

Section 1. Between array comparison

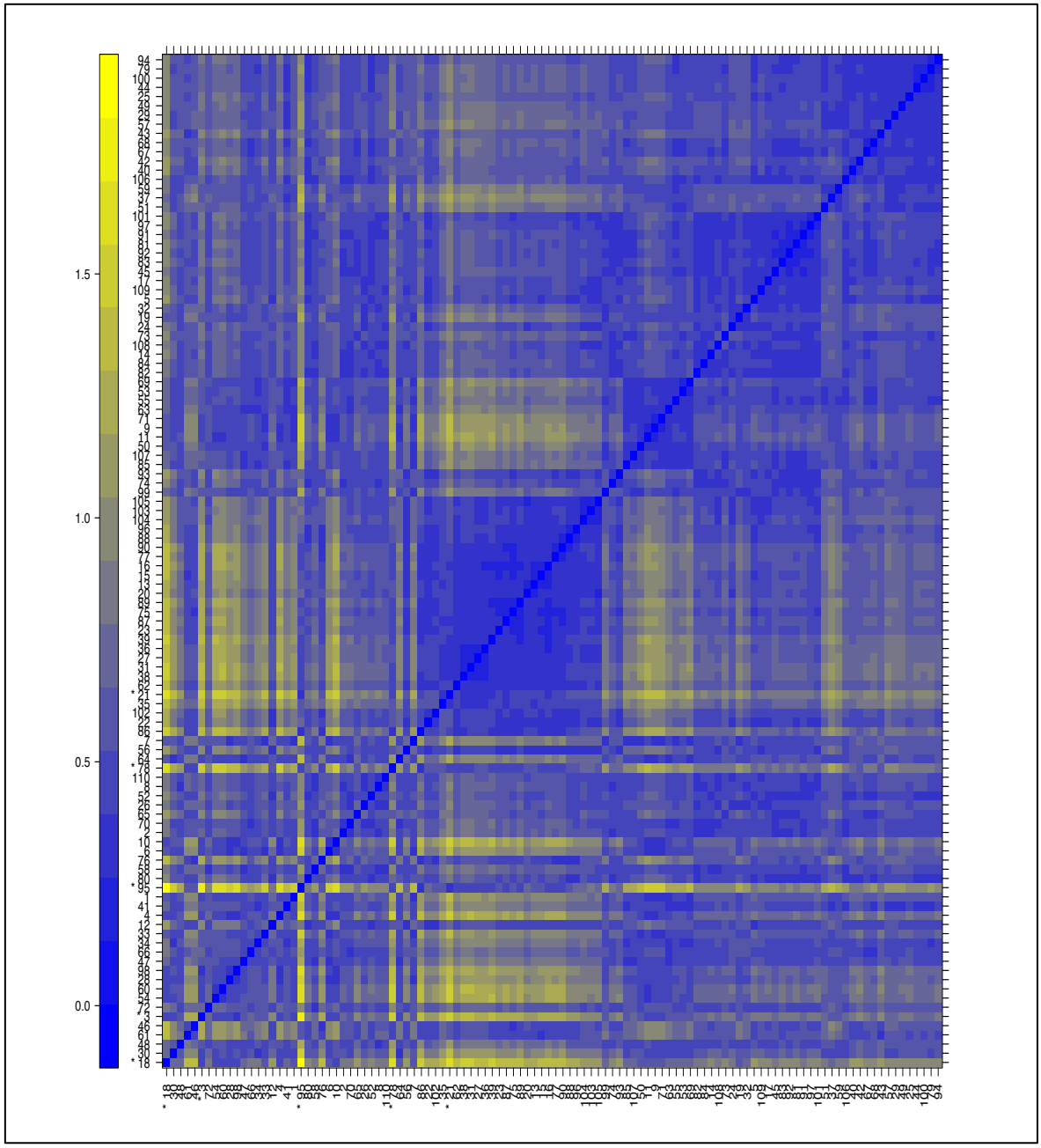


Figure A. Heatmap of the distances between the arrays. The array identification numbers are given on the x and y axes. The distance between two arrays is calculated as the mean absolute difference between the data of the arrays. The colour scale on the left of the heatmap depicts the range of distances encountered in the data. A threshold is established based on the distribution of the sum of distances across all arrays. Those above the threshold are considered to be outliers. Five arrays were considered to be outliers (Figure B) and are indicated by an asterisk on the x and y axes of the heatmap.

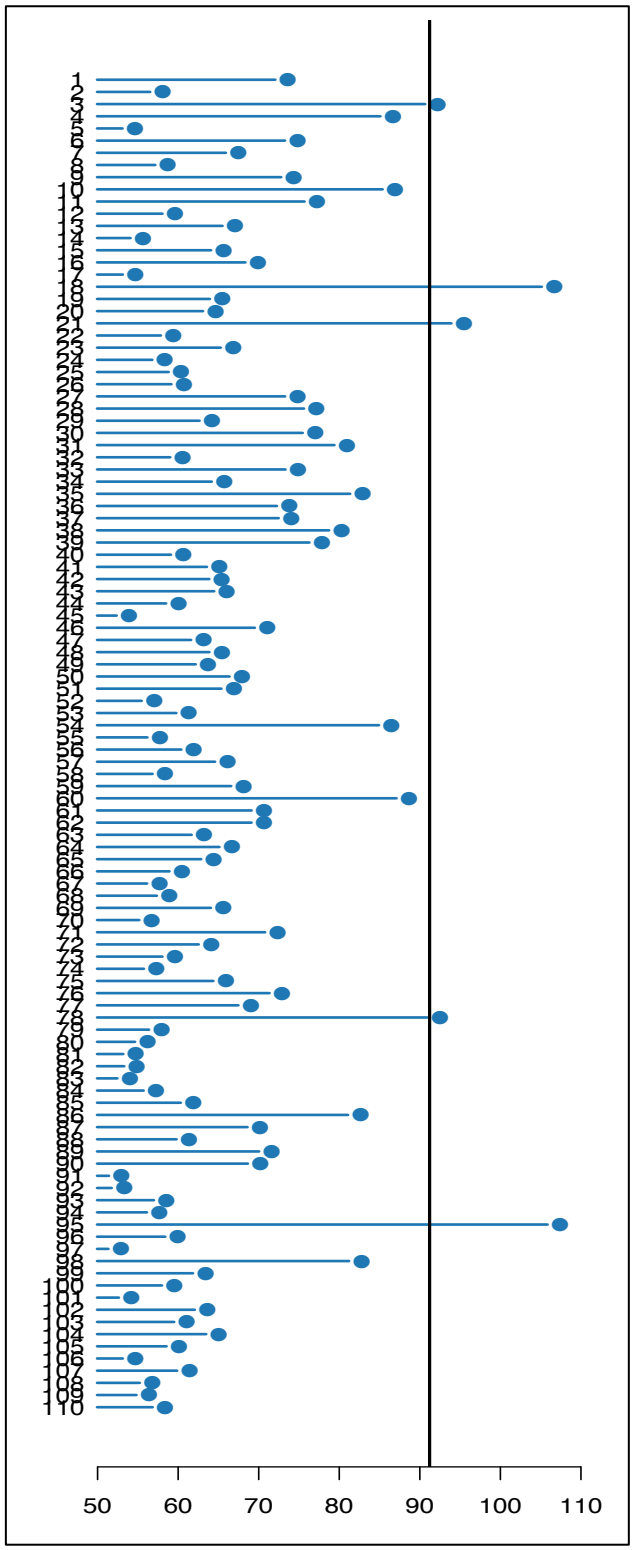


Figure B. Bar chart portraying outlier detection for distances between the arrays. The array identification numbers are given on the y axis. The distance between two arrays is calculated as the mean absolute difference between the data of the arrays. A threshold is established based on the distribution of the sum of distances across all arrays and represented as a vertical line. Outliers are found outside (to the right) of the line. A threshold of 91.2 was determined. Five arrays exceeded the threshold and were considered outliers.

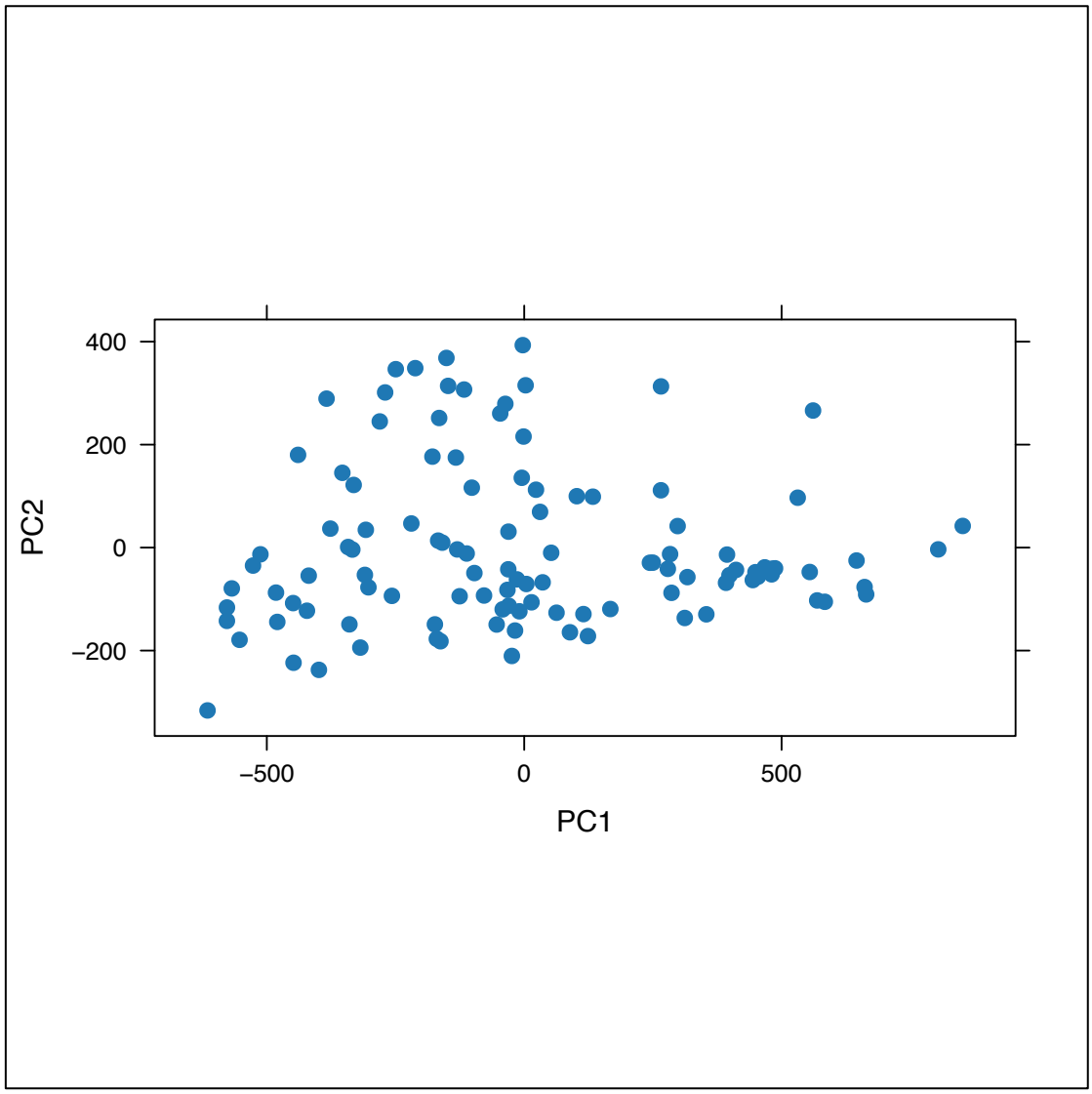


Figure C. Principal component analysis (PCA) plot representing the distribution of the arrays along the first two principal components. PCA is a dimension reduction and visualisation method. PCA is employed to translate the multivariate data vector of each array onto a two-dimensional plot. Overall array similarity (or dissimilarity) is displayed in the spatial arrangement of the points in the scatterplot. Outliers are portrayed by a larger dot.

Section 2. Array intensity distributions

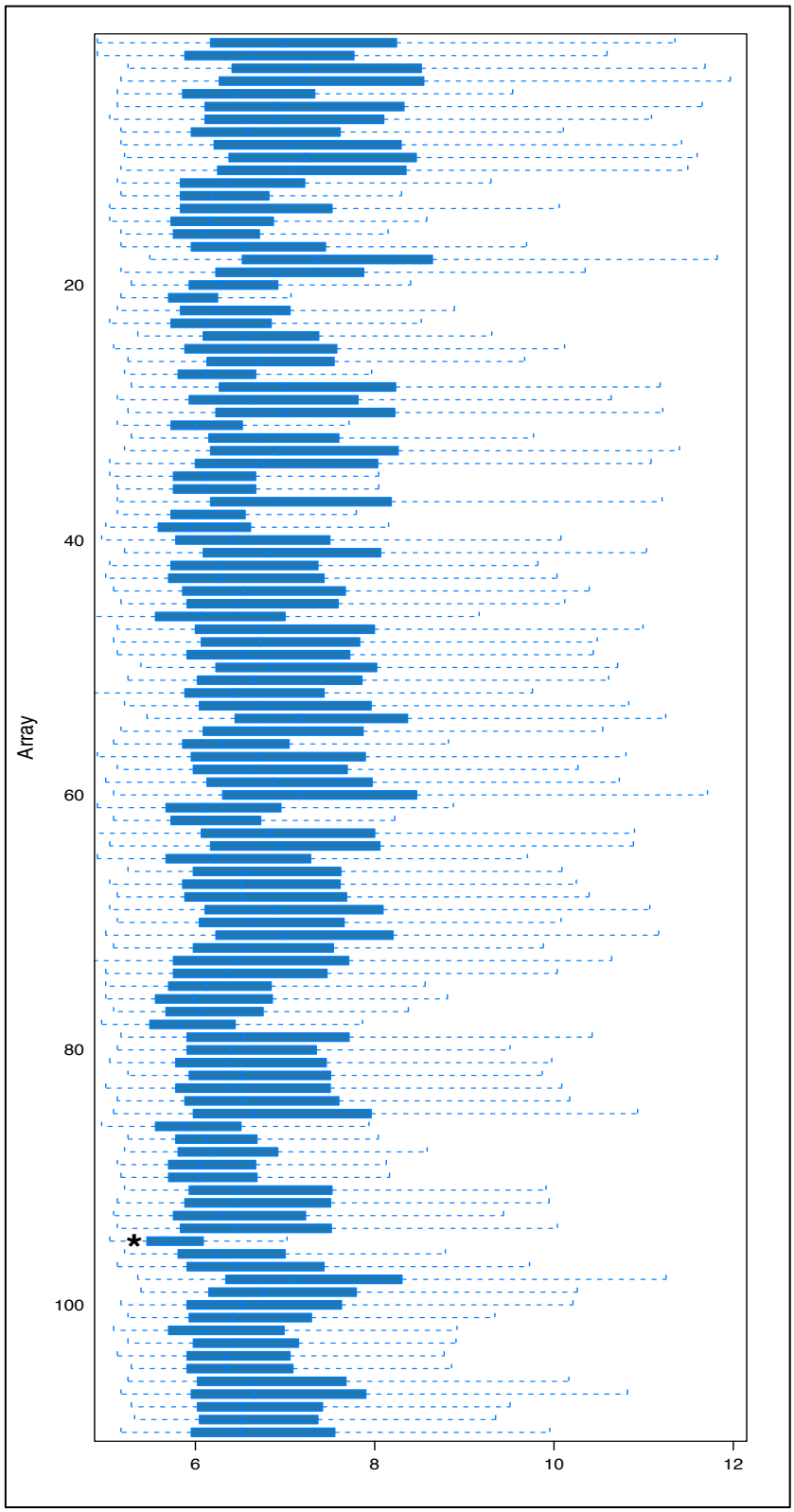


Figure D. Boxplots representing summaries of the signal intensity distributions for each array. The array identification numbers are given on the y axis. Generally, it is expected that the boxes will be in similar positions on the plot and have similar widths. One array was considered to be an outlier (Figure E) and is indicated by an asterisk.

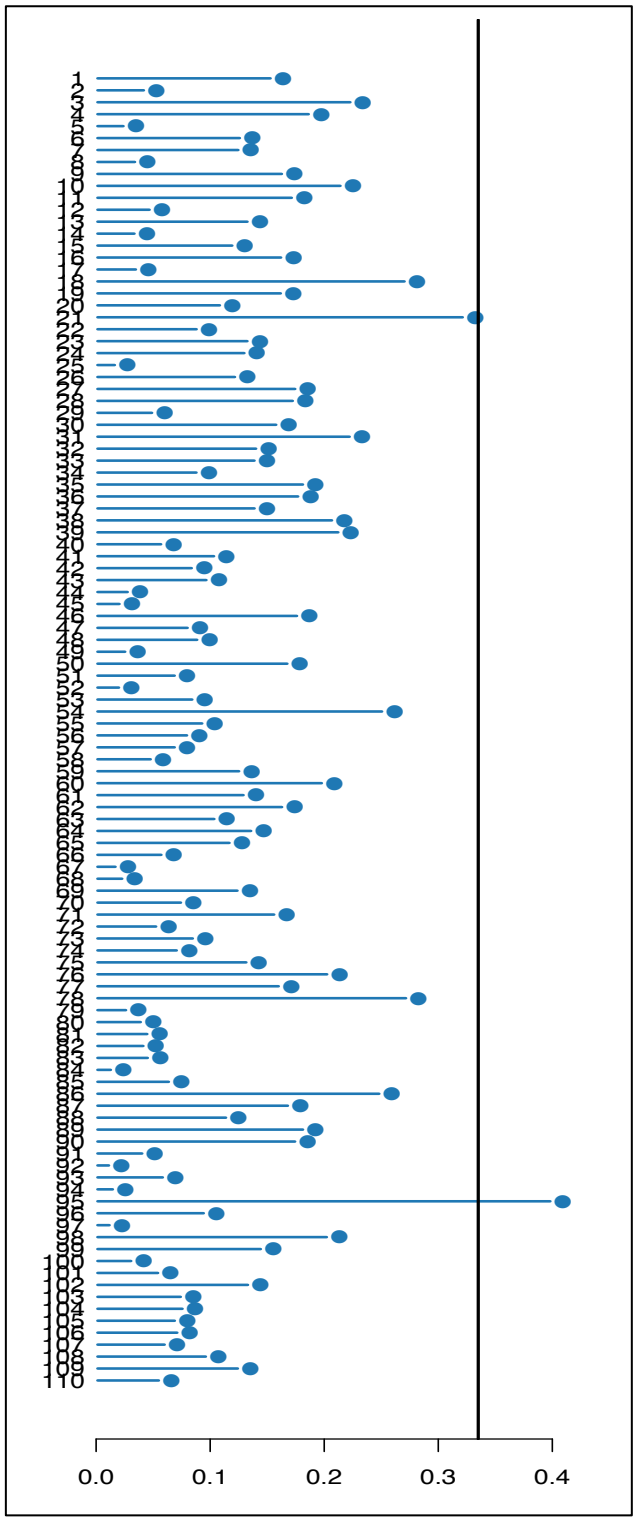


Figure E. Bar chart illustrating outlier detection for signal intensity distributions for each array. The array identification numbers are given on the y axis. Detection of outliers is achieved by calculating the Kolmogorov-Smirnov (KS) statistic between each array's distribution and the overall distribution of the combined data (Chakravarti and Laha 1967). A threshold is determined based on the distribution of the values across all arrays and represented as a vertical line. Outliers are found outside (to the right) of the threshold line. A threshold of 0.335 was determined. One array exceeded the threshold and was considered an outlier.

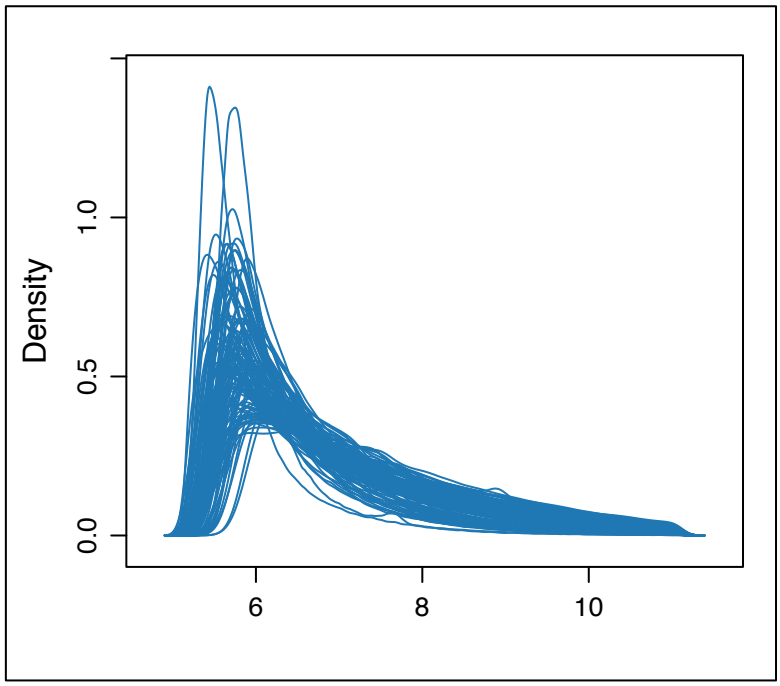


Figure F. Density histograms depicting the density estimates of the array intensity data. The plots are superimposed on a single graph enabling the observation of unusual distributions more readily. Generally, the histograms are expected to have similar shapes and ranges.

Section 3. Variance mean dependence

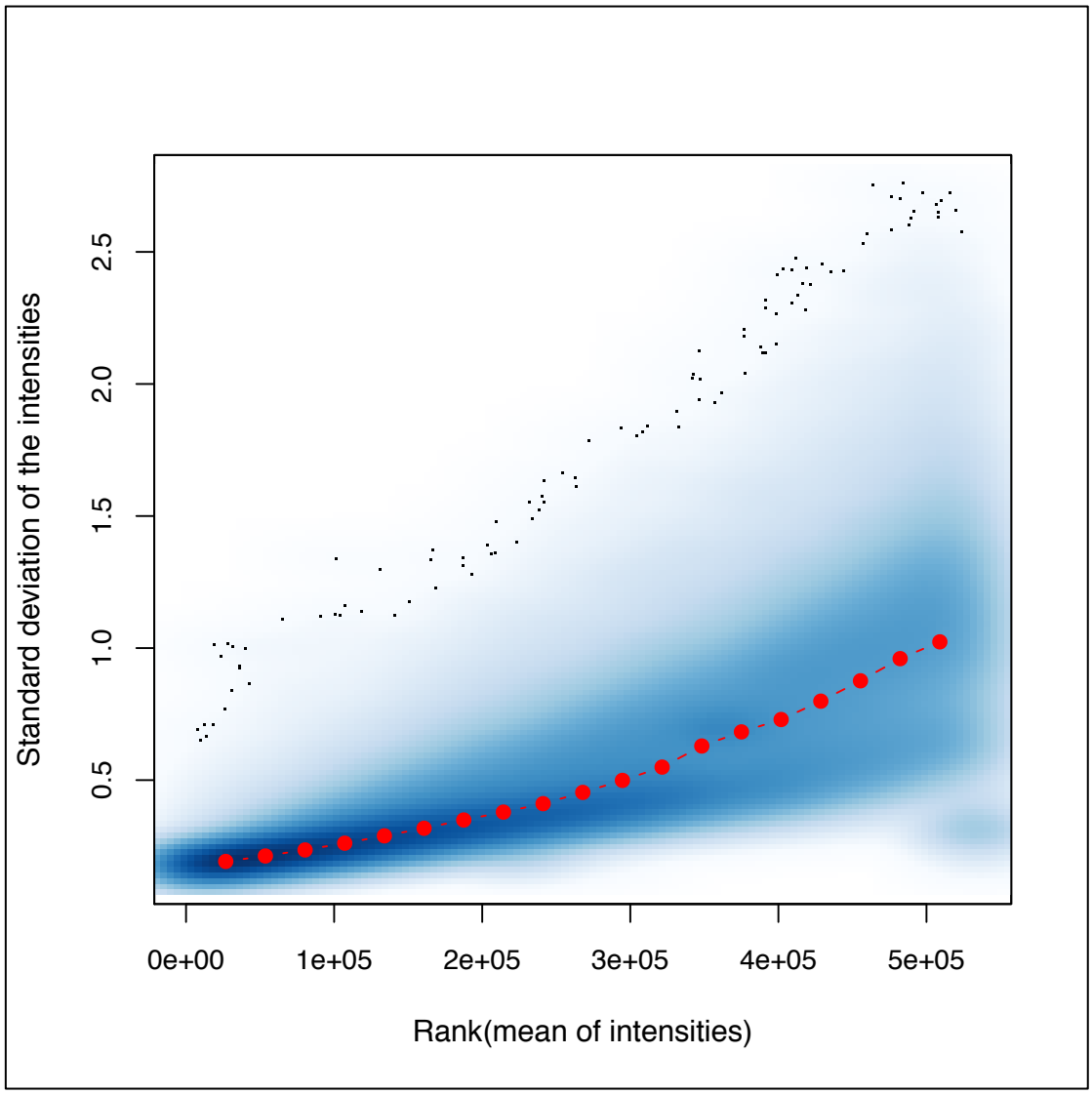


Figure G. Density plot representing the standard deviation of the signal intensities for the arrays versus the rank of their mean. The running median of the standard deviation is shown using red dots connected by lines. After the intensity data has been normalised and transformed to a logarithm-like scale, it is generally expected that the red line will be approximately horizontal, which would signify no significant trend.

Section 4. Affymetrix specific plots

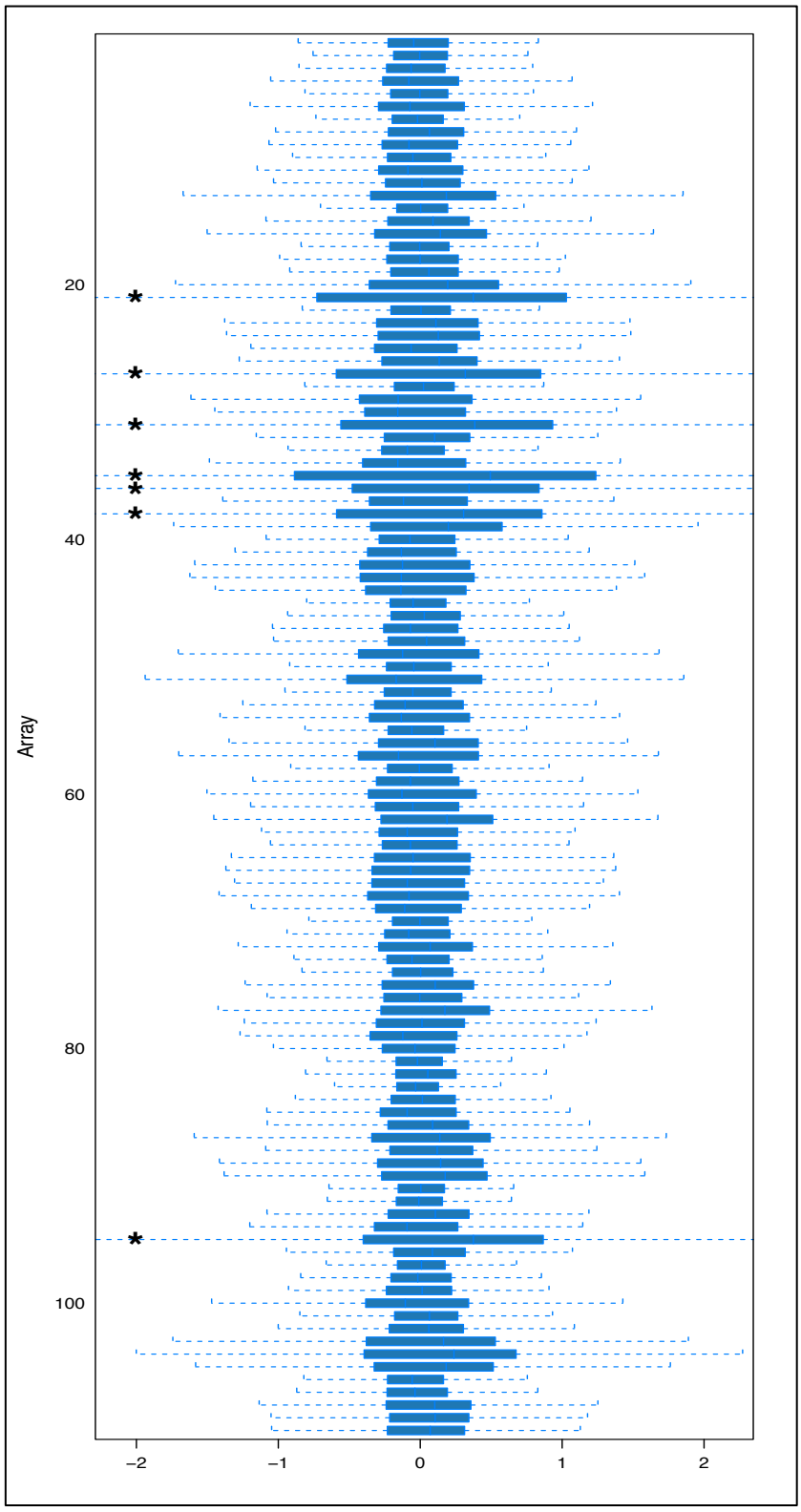


Figure H. Boxplots presenting the Relative Log Expression (RLE) for each array. The array identification numbers are given on the y axis. An array is considered to be potentially problematic in terms of quality if its boxplot is centred away from zero and/or it is more spread than the majority of the arrays. The seven arrays highlighted with an asterisk are considered to be outliers.

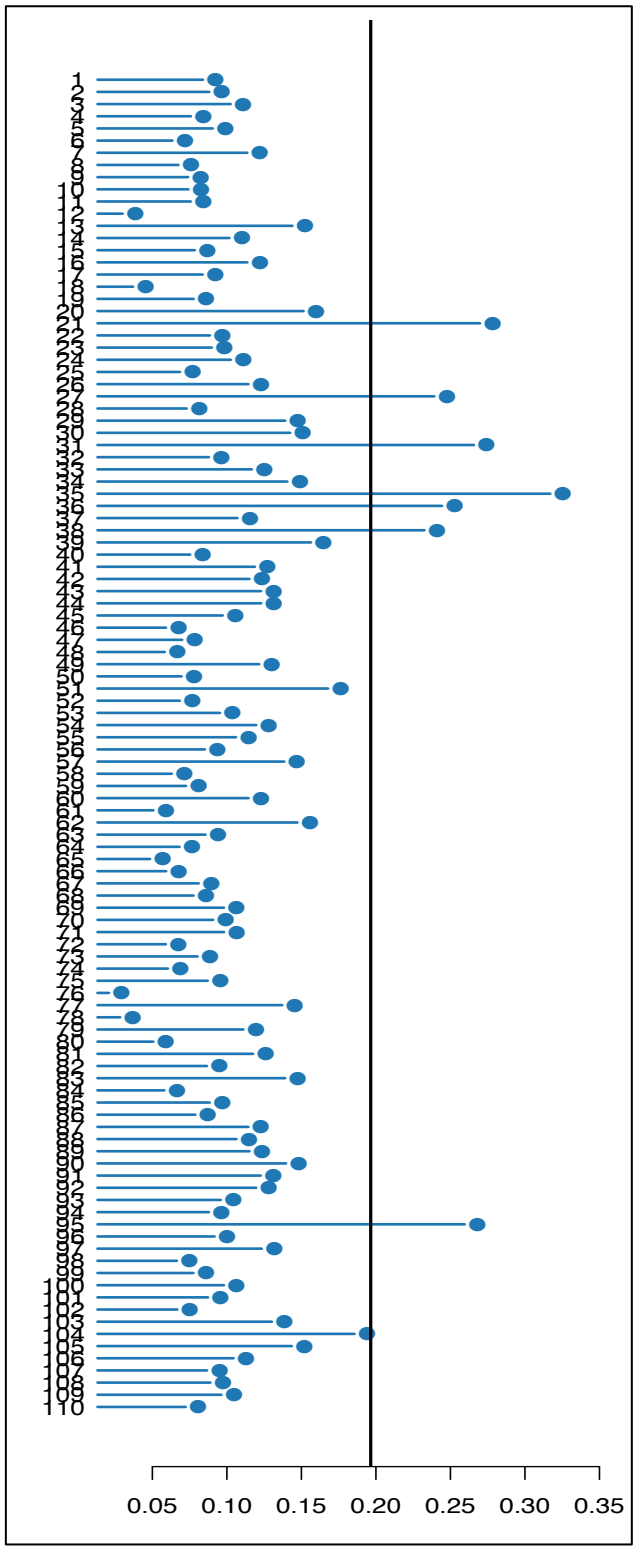


Figure I. Bar chart illustrating outlier detection for Relative Log Expression for each array. The array identification numbers are given on the y axis. Detection of outliers is achieved by calculating the Kolmogorov-Smirnov (KS) statistic between each array's RLE values and the pooled, overall distribution of the RLE values (Chakravarti and Laha 1967). A threshold is determined based on the distribution of the values across all arrays and represented as a vertical line. Outliers are found outside (to the right) of the threshold line. A threshold of 0.197 was determined. Seven arrays exceeded the threshold and were considered outliers.

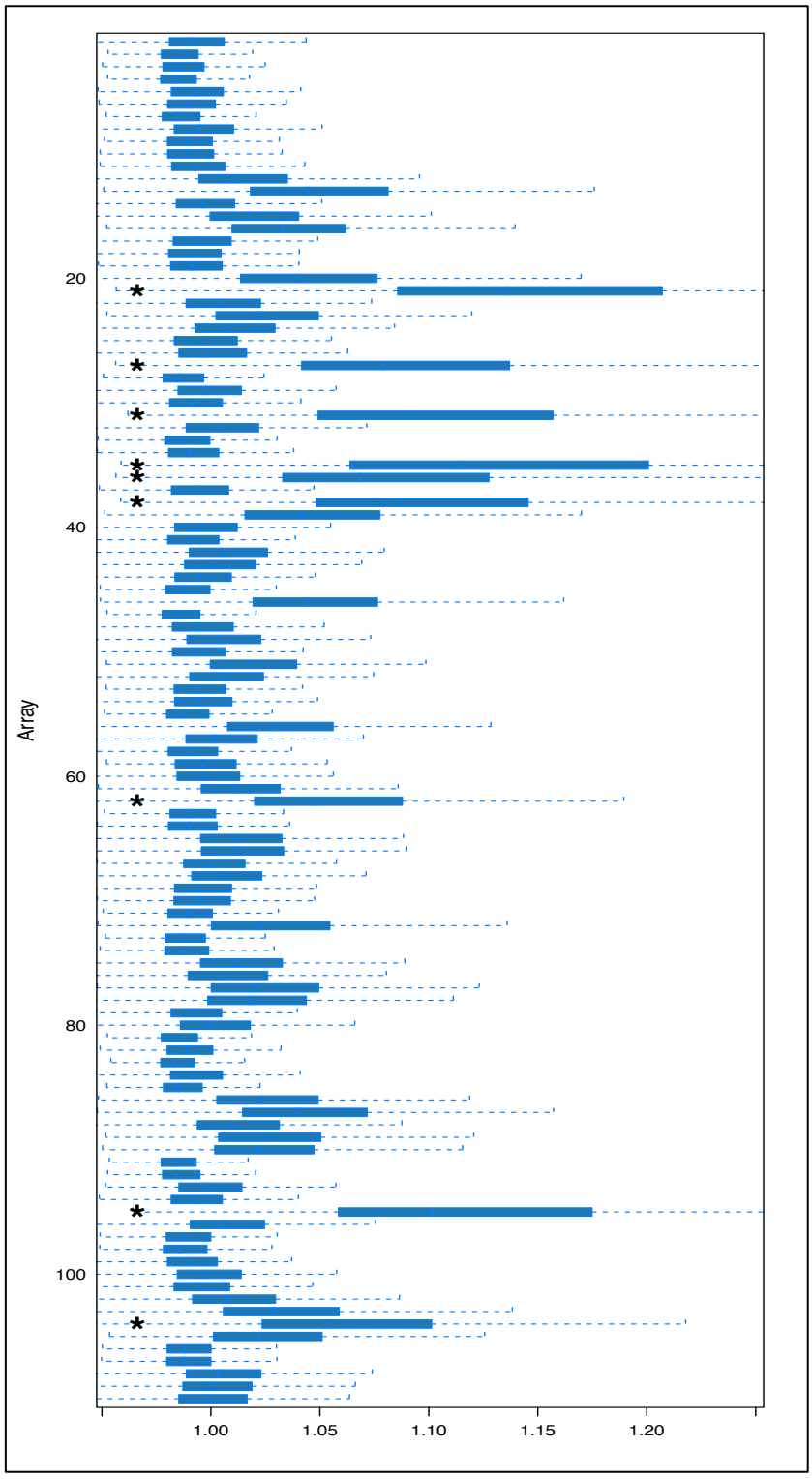


Figure J. Boxplots presenting the Normalised Unscaled Standard Error (NUSE) for each array. The array identification numbers are given on the y axis. For each array the boxplot should be centred around one. An array with elevated values relative to the other arrays is typically of lower quality. Nine of these are indicated in the plot by an asterisk.

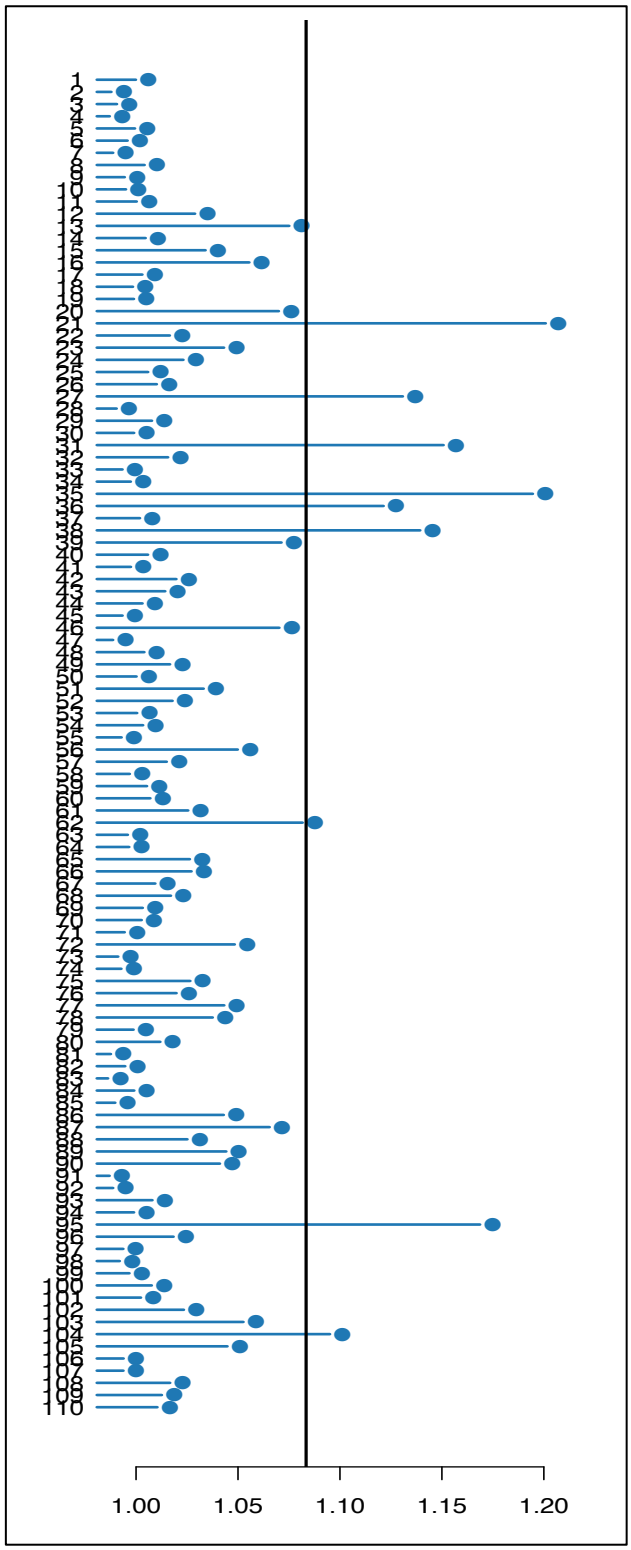


Figure K. Bar chart illustrating outlier detection for Normalised Unscaled Standard Error for each array. The array identification numbers are given on the y axis. Detection of outliers is achieved by calculating the 75% quantile, N_a, of each array's NUSE values. Outliers have a large N_a. A threshold is determined based on the distribution of the values across all arrays and represented as a vertical line. Outliers are found outside (to the right) of the threshold line. A threshold of 1.08 was determined. Nine arrays exceeded the threshold and were considered outliers.

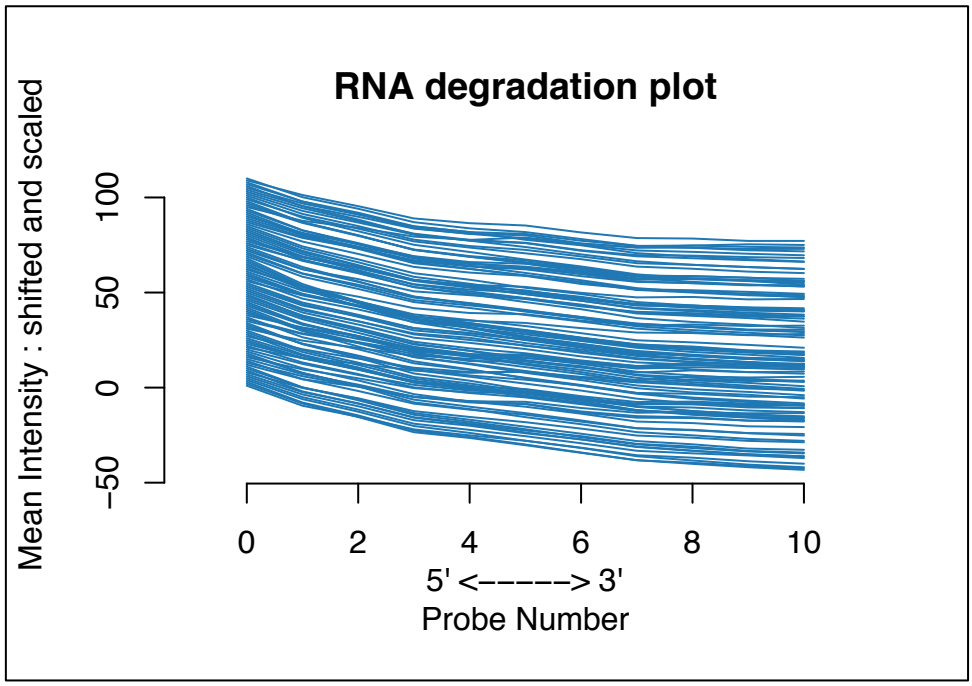


Figure L. RNA degradation plot. Each array is represented by a single line. The values in the plot are calculated from preprocessed data, after background correction and quantile normalisation. If an array in the plot has a different slope from the others it could indicate that the RNA on that array was dealt with differently than the RNA on the other arrays.

Section 5. Individual array quality

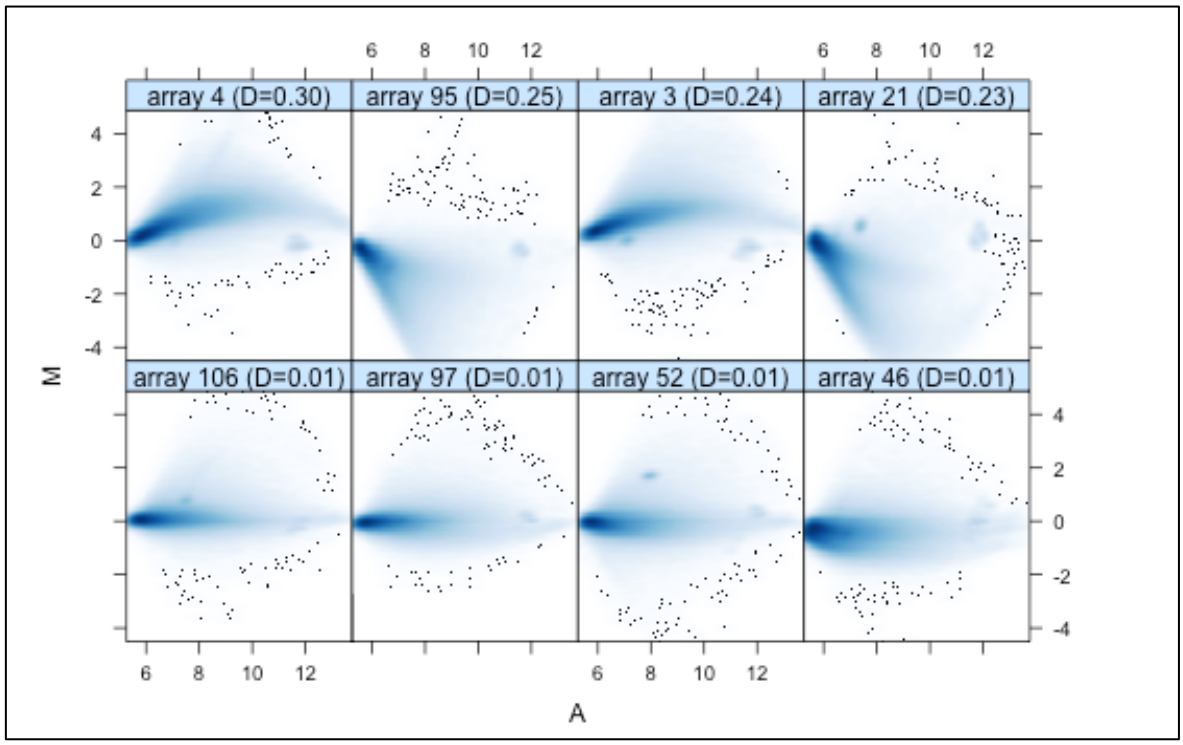


Figure M. MA plots to visualise the distribution of the difference between log intensities (M) for an array versus the average log intensities (A) for the array. The array of interest is compared to an 'array' consisting of the median log intensities across the set of experiment arrays. Generally, it is expected that the majority of the distribution would be focussed along the M=0 axis, signifying no trend. Arrays with the highest and lowest Hoeffding's D statistics are presented, along with a bar chart of the Hoeffding's D statistics (Figure I). The D statistic is shown in the panel headings.

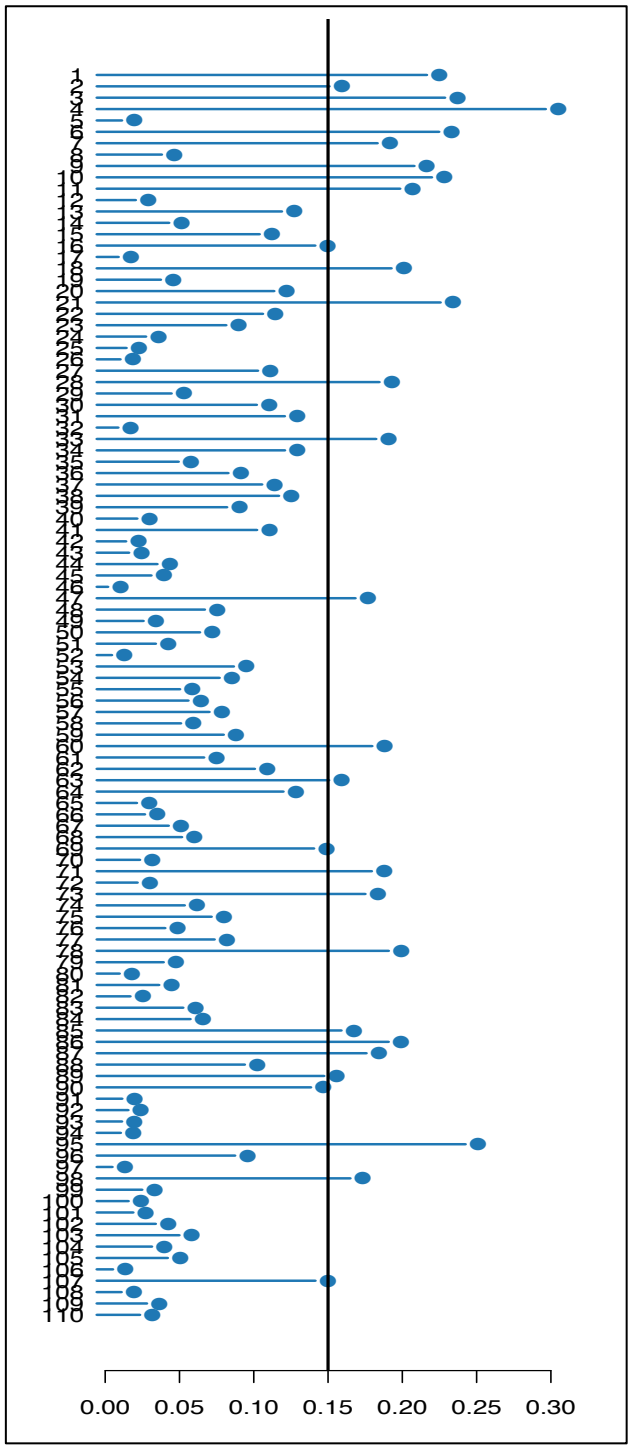


Figure N. Bar chart illustrating outlier detection for the distribution of the difference between log intensities (M) for an array versus the average log intensities (A) for the array. The array identification numbers are given on the y axis. Detection of outliers is achieved by calculating Hoeffding's D statistic on the paired distribution of M and A for each array (Hoeffding 1948). Hoeffding's D statistic is a non-parametric test of independence, which depends on the rank order of the observations, in this case the rank order of M for an array versus the rank order of A for the same array. The range for D in this context is -0.5 to 1.0. The higher the D statistic, the more dependent the variables being tested. A threshold of 0.15 is used in this context. Outliers are found outside (to the right) of the threshold line. Twenty five of the arrays exceeded the threshold and were considered to be outliers.

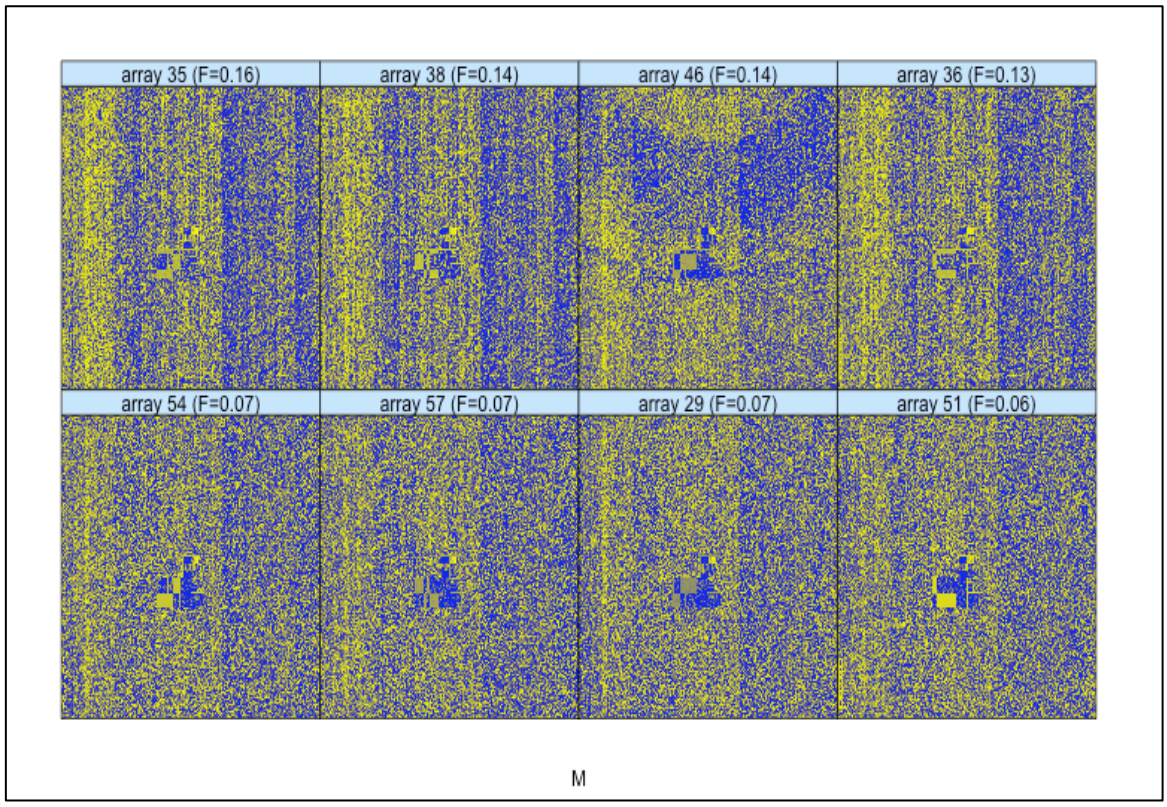


Figure O. False colour representations of the arrays' spatial distributions of feature intensities (M). When the features are distributed randomly on the arrays, the false colour plots appear uniform. Control features may stand out if they have particularly high or low intensities. The colour scale is proportional to the ranks of the probe intensities. The detection of outliers is achieved by calculating $F_{a,}$ the sum of the absolute value of low frequency Fourier coefficients. Above (in the top row) are the 4 arrays with the highest and 4 arrays with the lowest F_a values. The F_a value is shown in the panel heading.

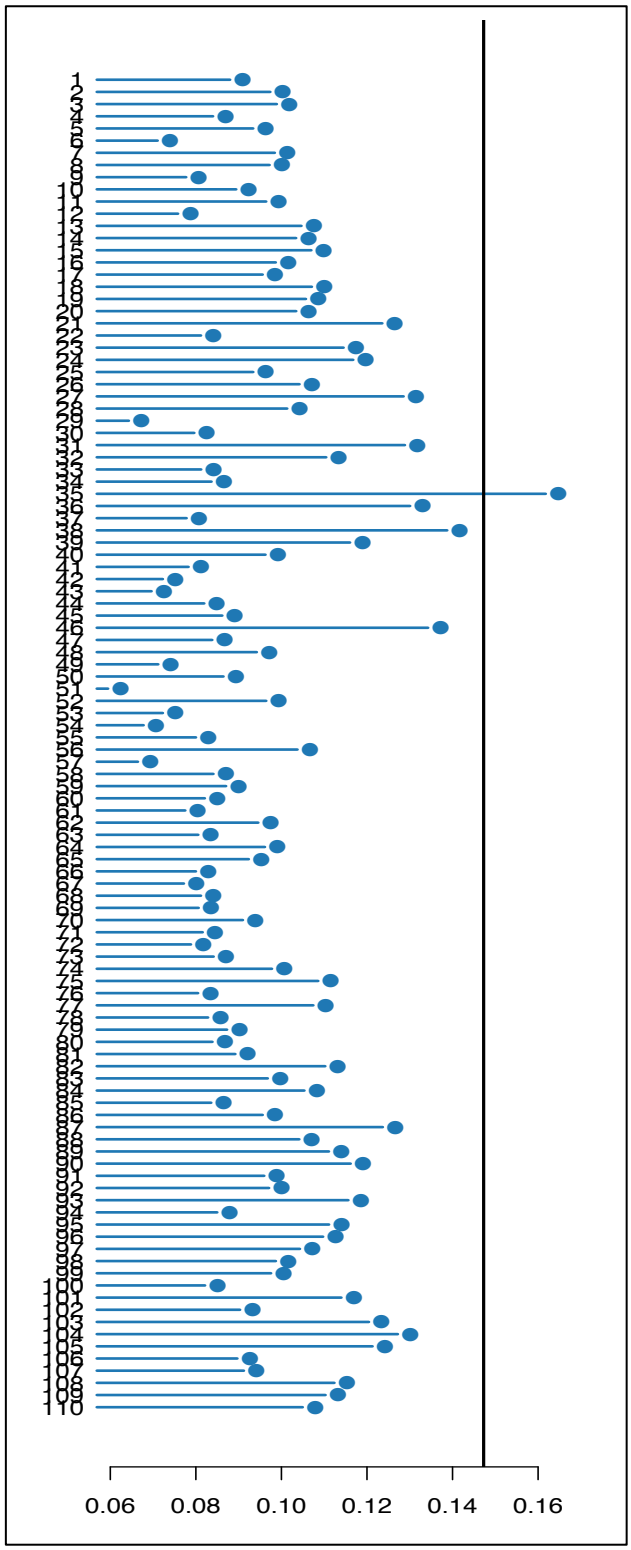


Figure P. Bar chart illustrating the F_a for each array. The array identification numbers are given on the y axis. A threshold is determined based on the distribution of the values across all arrays and represented as a vertical line. Outliers are found outside (to the right) of the threshold line. A threshold of 0.147 was determined. One array exceeded the threshold and was considered an outlier.

APPENDIX 4

arrayQualityMetrics Report for Normalised Microarray Expression Data

Section 1. Between array comparison

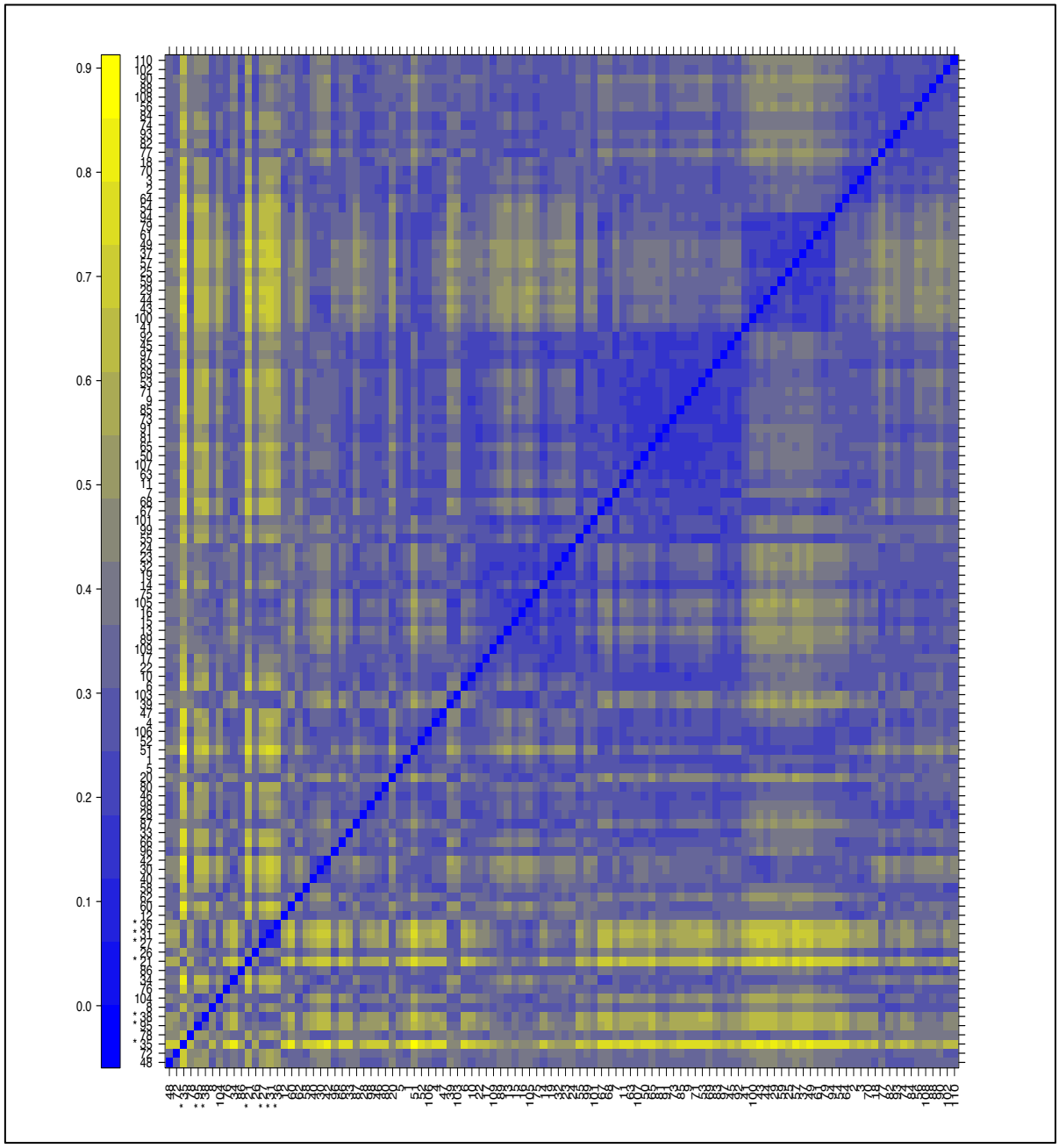


Figure A. Heatmap of the distances between the arrays. The array identification numbers are given on the x and y axes. The distance between two arrays is calculated as the mean absolute difference between the data of the arrays. The colour scale on the left of the heatmap depicts the range of distances encountered in the data. A threshold is established based on the distribution of the sum of distances across all arrays. Those above the threshold are considered to be outliers. Seven arrays were considered to be outliers (Figure B) and are indicated by an asterisk on the x and y axes of the heatmap.

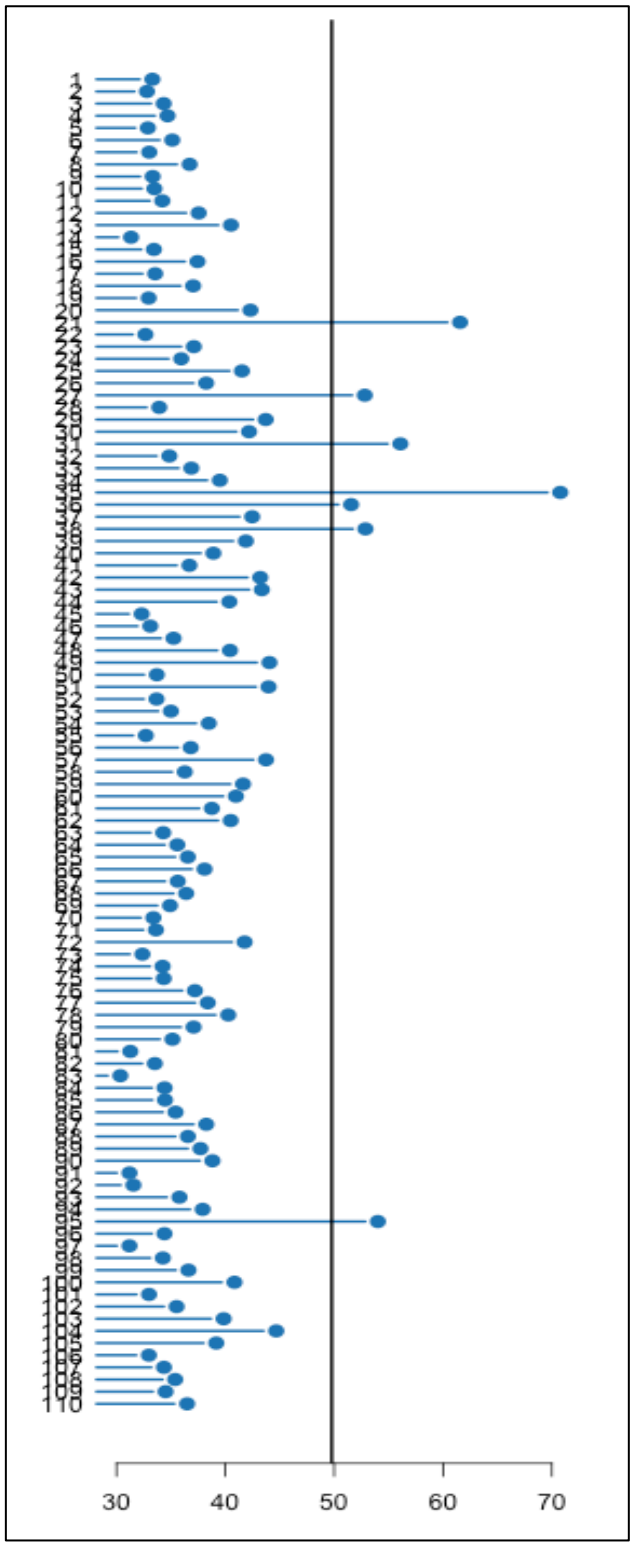


Figure B. Bar chart portraying outlier detection for distances between the arrays. The array identification numbers are given on the y axis. The distance between two arrays is calculated as the mean absolute difference between the data of the arrays. A threshold is established based on the distribution of the sum of distances across all arrays and represented as a vertical line. Outliers are found outside (to the right) of the line. A threshold of 49.8 was determined. Seven arrays exceeded the threshold and were considered outliers.

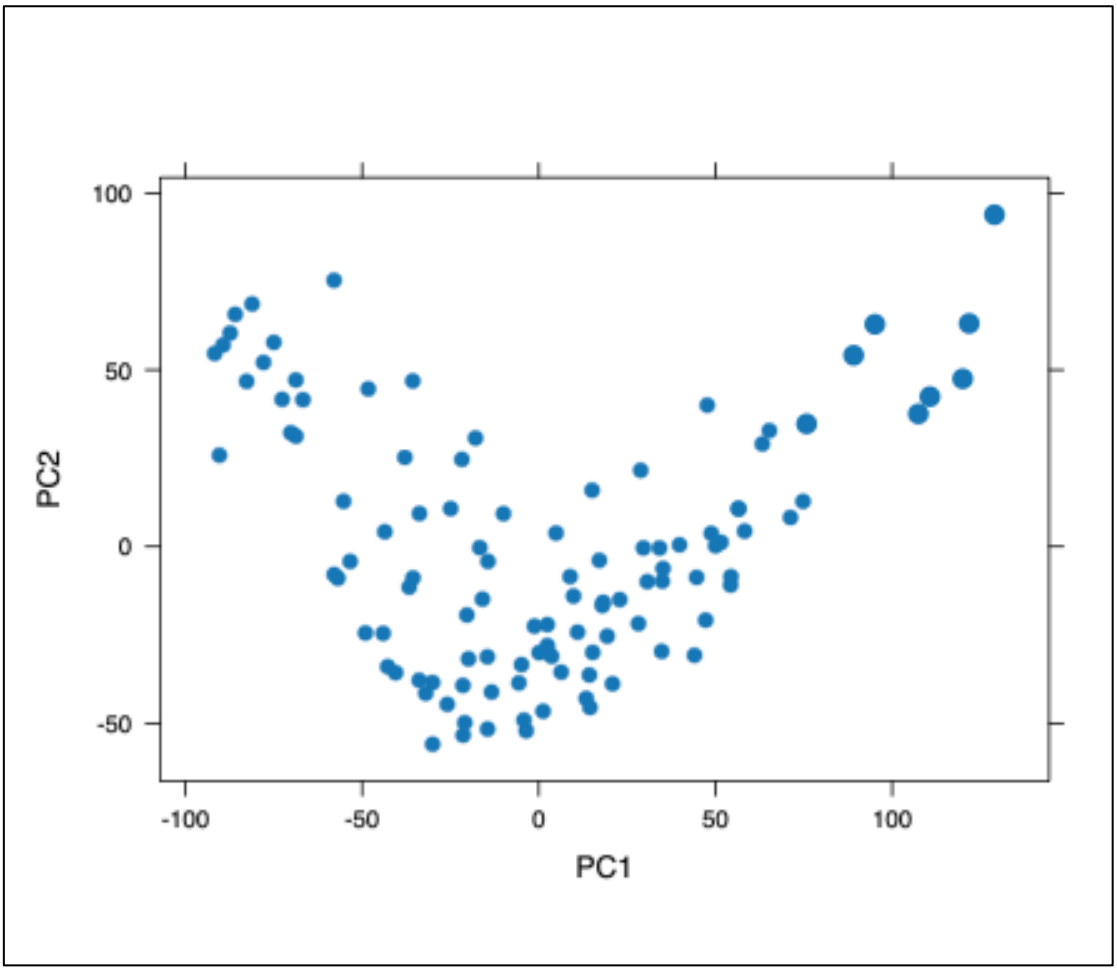


Figure C. Principal component analysis plot representing the distribution of the arrays along the first two principal components. PCA is a dimension reduction and visualisation method. PCA is employed to translate the multivariate data vector of each array onto a two-dimensional plot. Overall array similarity (or dissimilarity) is displayed in the spatial arrangement of the points in the scatterplot. Outliers are portrayed by a larger dot.

Section 2. Array intensity distributions

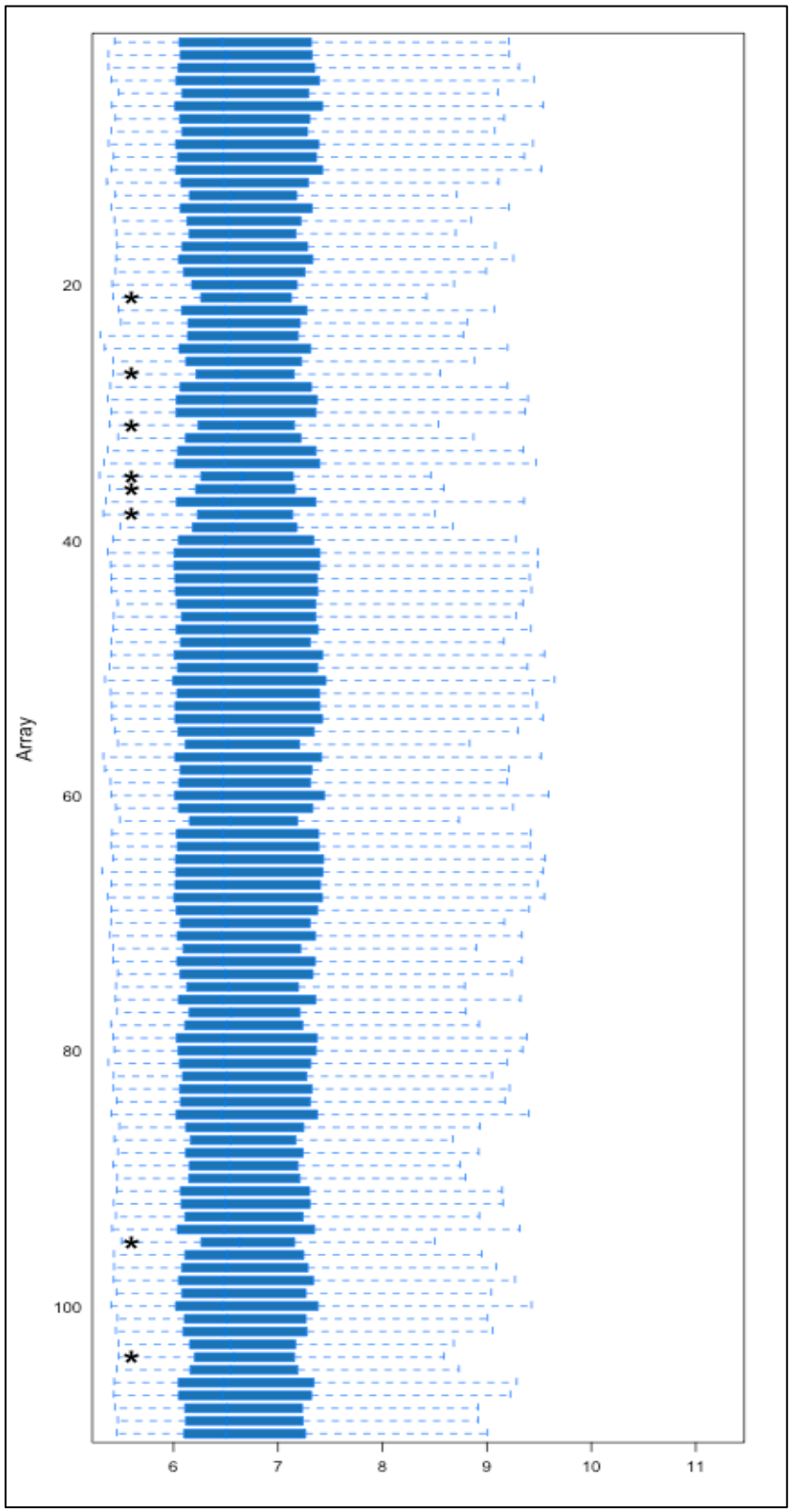


Figure D. Boxplots representing summaries of the signal intensity distributions for each array. The array identification numbers are given on the y axis. Generally, it is expected that the boxes will be in similar positions on the plot and have similar widths. Eight arrays were considered to be outliers (Figure E) and are indicated by an asterisk.

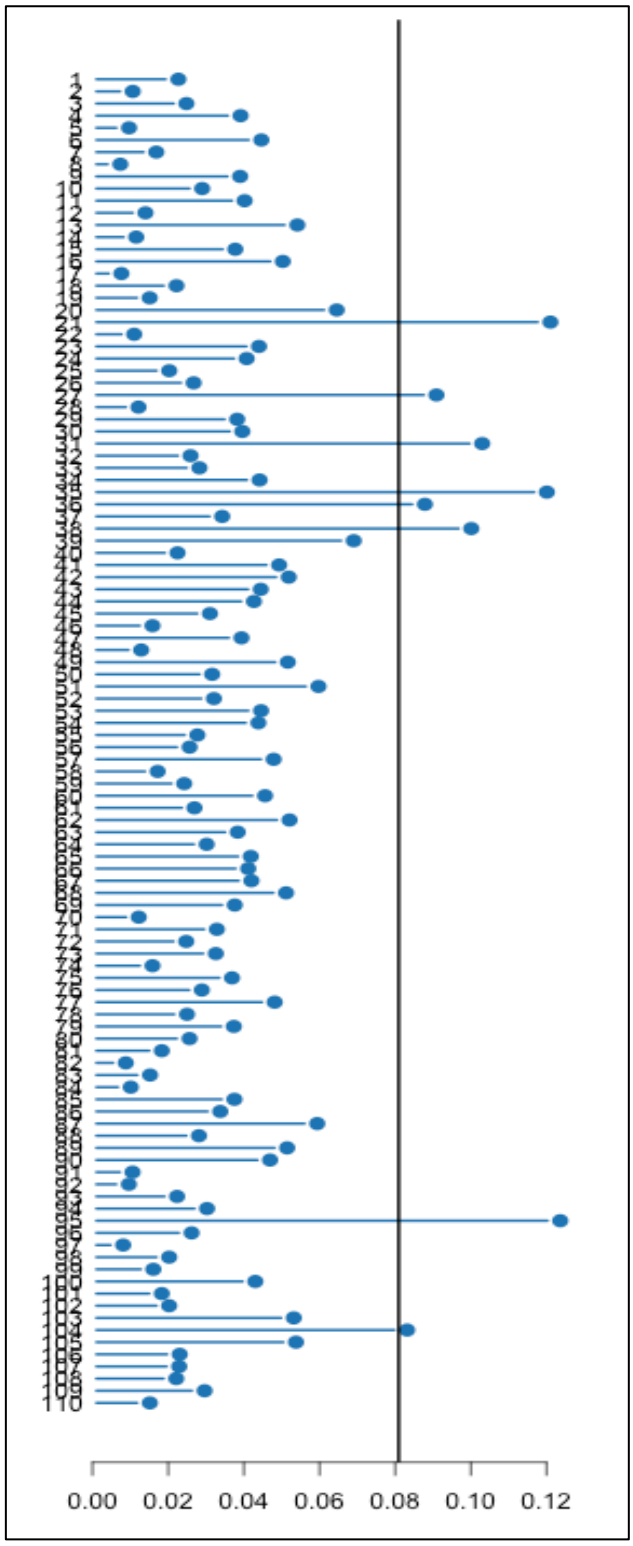


Figure E. Bar chart illustrating outlier detection for signal intensity distributions for each array. The array identification numbers are given on the y axis. Detection of outliers is achieved by calculating the Kolmogorov-Smirnov (KS) statistic between each array's distribution and the overall distribution of the combined data (Chakravarti and Laha 1967). A threshold is determined based on the distribution of the values across all arrays and represented as a vertical line. Outliers are found outside (to the right) of the threshold line. A threshold of 0.081 was determined. Eight arrays exceeded the threshold and were considered outliers.

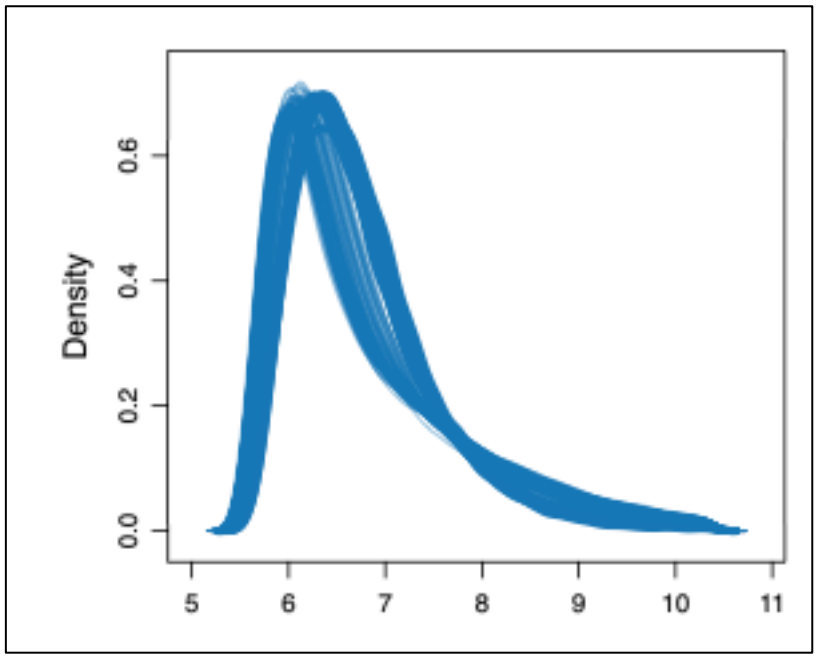


Figure F. Density histograms depicting the density estimates of the array intensity data. The plots are superimposed on a single graph enabling the observation of unusual distributions more readily. Generally, the histograms are expected to have similar shapes and ranges.

Section 3. Variance mean dependence

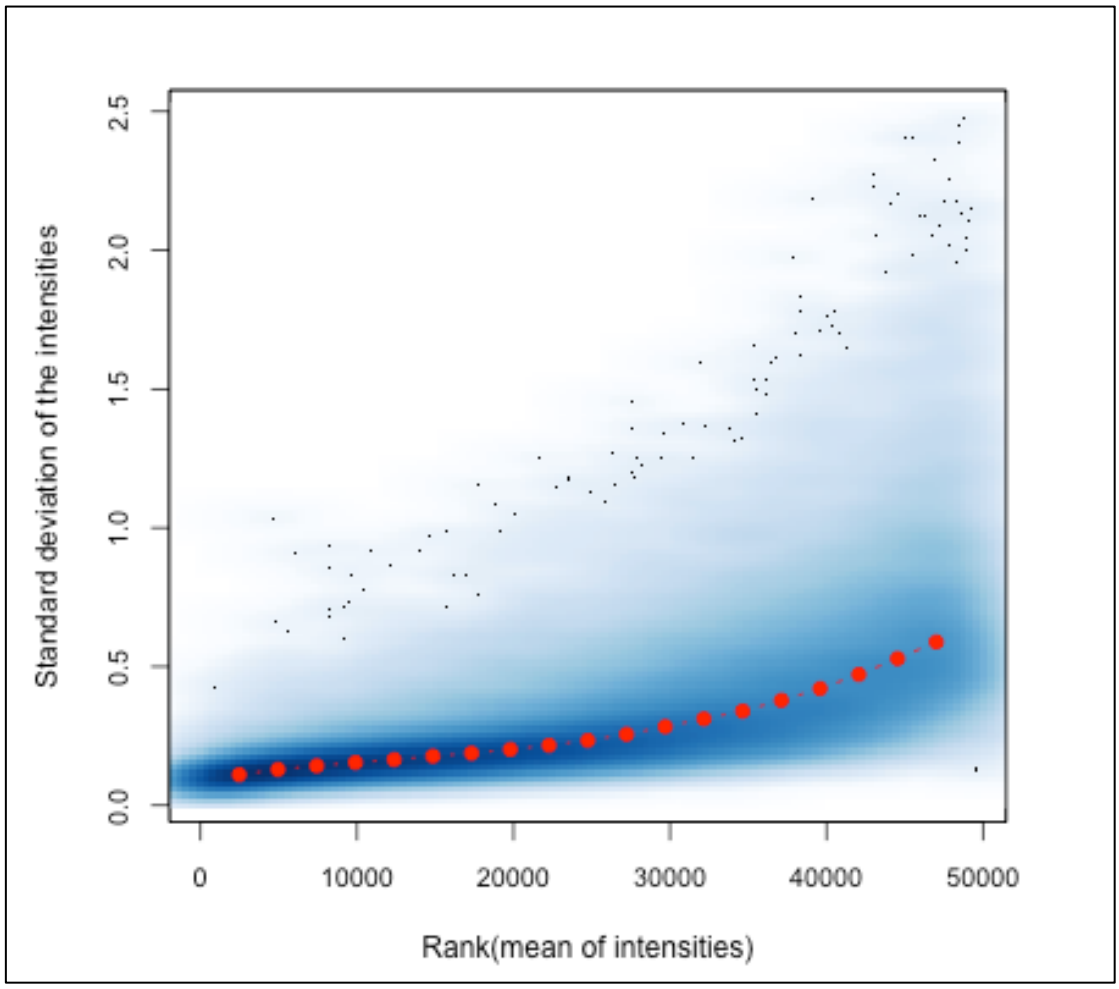


Figure G. Density plot representing the standard deviation of the signal intensities for the arrays versus the rank of their mean. The running median of the standard deviation is shown using red dots connected by lines. After the intensity data has been normalised and transformed to a logarithm-like scale, it is generally expected that the red line will be approximately horizontal, which would signify no significant trend.

Section 4. Individual array quality

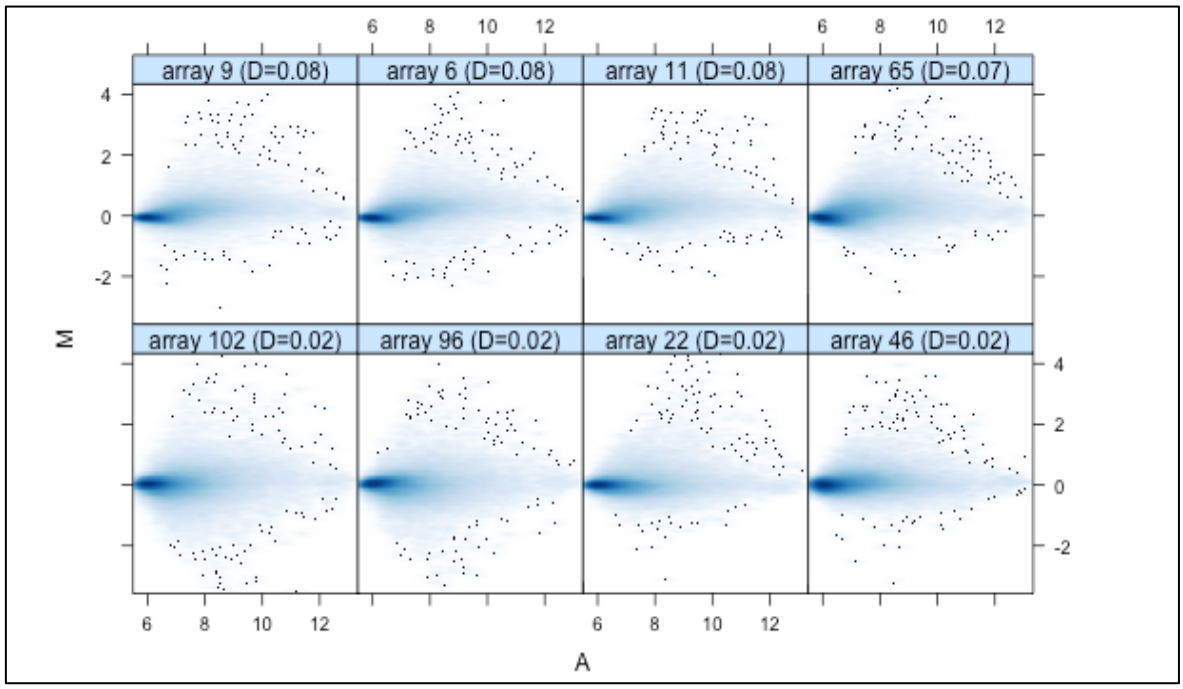


Figure H. MA plots to visualise the distribution of the difference between log intensities (M) for an array versus the average log intensities (A) for the array. The array of interest is compared to an 'array' consisting of the median log intensities across the set of experiment arrays. Generally, it is expected that the majority of the distribution would be focussed along the M=0 axis, signifying no trend. Arrays with the highest and lowest Hoeffding's D statistics are presented, along with a bar chart of the Hoeffding's D statistics (Figure I). The D statistic is shown in the panel heading.

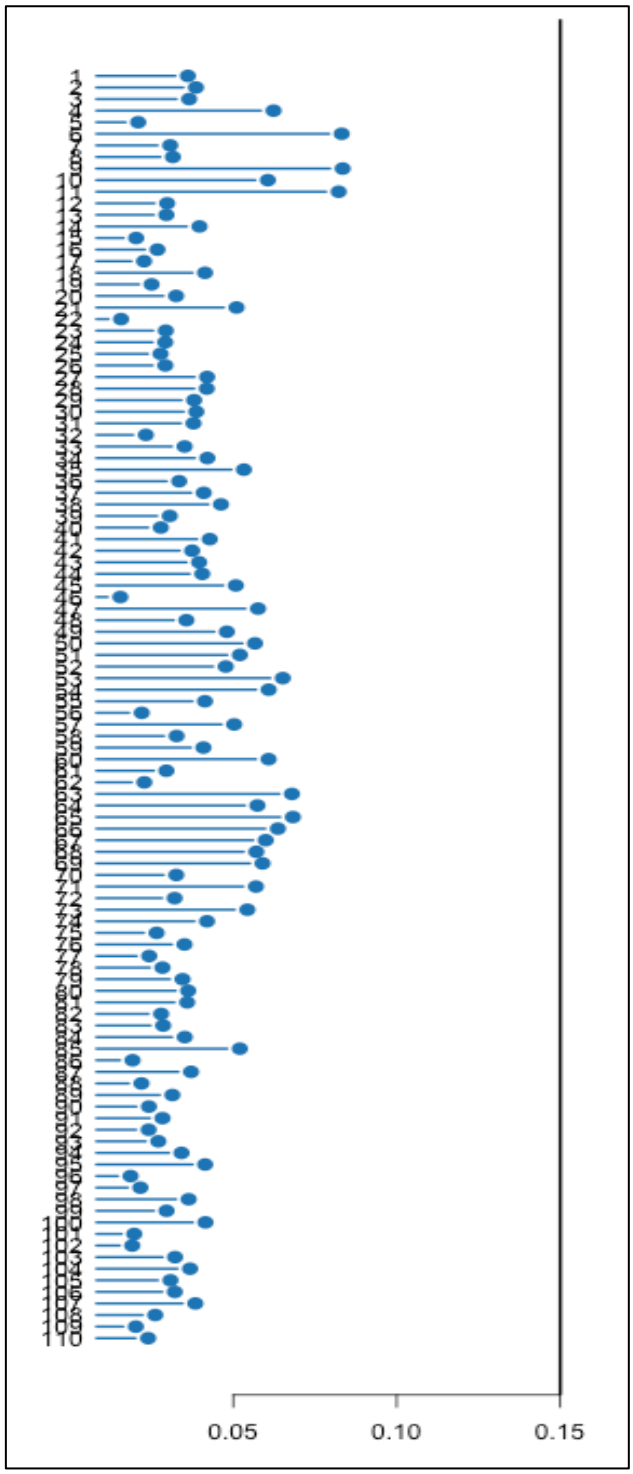


Figure I. Bar chart illustrating outlier detection for the distribution of the difference between log intensities (M) for an array versus the average log intensities (A) for the array. The array identification numbers are given on the y axis. Detection of outliers is achieved by calculating Hoeffding's D statistic on the paired distribution of M and A for each array (Hoeffding 1948). Hoeffding's D statistic is a non-parametric test of independence, which depends on the rank order of the observations, in this case the rank order of M for an array versus the rank order of A for the same array. The range for D in this context is -0.5 to 1.0. The higher the D statistic, the more dependent the variables being tested. A threshold of 0.15 is used in this context. Outliers are found outside (to the right) of the threshold line. None of the arrays exceeded the threshold.

APPENDIX 5a

RNA quality- purity

260/280 ratios

The ratio of absorbance at 260nm and 280nm was used to assess RNA purity. For RNA, a ratio of approximately 2.0 is typically accepted as being 'pure'. If the ratio is appreciably lower than this, the presence of protein or other contaminants that absorb strongly at 280nm may be indicated. RNA purity was assessed using the Implen NanoPhotometer® (Implen GmbH, Munchen, Germany). Following are the 260/280 ratios for the RNA used in this study. Abbreviations: N= normal, T= tumour.

CEL File	Study Number	260/280	Comment
VD21	VDS01 N	0.734	
VD22	VDS01 T	0.751	
VD111	VDS02 N	1.77	
VD112	VDS02 T	1.39	
VD23	VDS03 N	0.552	
VD24	VDS03 T	0.672	
VD25	VDS04 N	2.04	
VD26	VDS04 T	1.96	
VD113	VDS05 N	1.41	
VD114	VDS05 T	1.17	
VD27	VDS06 N	1.99	
VD28	VDS06 T	2.05	
VD29	VDS07 N	2.04	

VD30	VDS07 T	2.2	
VD31	VDS08 N	2.08	
VD32	VDS08 T	2.03	
VD9	VDS09 N	0.789	
VD10	VDS09 T	0.919	
VD33	VDS10 N	2.11	
VD34	VDS10 T	2	
VD35	VDS11 N	1.99	
VD36	VDS11 T	2	
VD37	VDS12 N	2.09	
VD38	VDS12 T	2.05	
VD39	VDS13 N	2.02	
VD40	VDS13 T	2.01	
VD41	VDS14 N	2.05	
VD42	VDS14 T	2.12	
VD43	VDS15 N	2.02	
	VDS15 T		
VD45	VDS16 N	2.07	
VD46	VDS16 T	1.6	
VD47	VDS17 N	2.2	
VD48	VDS17 T	2.14	

	VDS18 N		No tissue collected
	VDS18 T		
VD49	VDS19 N	2.12	
VD50	VDS19 T	2.11	
VD51	VDS20 N	0.39	
VD52	VDS20 T	0.79	
VD44	VDS21 N	2.02	
	VDS21 T		
VD57	VDS22 N	2.12	
VD58	VDS22 T	2.07	
VD17	VDS23 N	1.23	
VD18	VDS23 T	1.19	
VD11	VDS24 N	1.51	
VD12	VDS24 T	1.14	
VD59	VDS25 N	2.12	
VD60	VDS25 T	2.13	
VD61	VDS26 N	2.12	
VD62	VDS26 T	2.04	
VD19	VDS27 N	1.23	
VD20	VDS27 T	0.711	

VD63	VDS28 N	2.1	
VD64	VDS28 T	2.06	
VD65	VDS29 N	2.1	
VD66	VDS29 T	2.04	
VD13	VDS30 N	0.674	
VD14	VDS30 T	1.27	
VD54	VDS31 N	2.09	
VD55	VDS31 T1 caecal	2.05	
VD56	VDS31 T2 transverse	2.02	
	VDS32 N		Tissue put in formalin
	VDS32 T		
VD67	VDS33 N	2.09	
VD68	VDS33 T	2.08	
VD69	VDS34 N	2.12	
VD70	VDS34 T	2.08	
VD102	VDS35 N	1.36	
VD103	VDS35 T1 transverse	1	
VD104	VDS35 T2 rectum	1.84	
VD15	VDS36 N	0.933	
VD16	VDS36 T	1.09	
VD71	VDS37 N	2.11	
VD72	VDS37 T	2.05	

VD73	VDS38 N	2.11	
VD74	VDS38 T	2.05	
VD75	VDS39 N	2.12	
VD76	VDS39 T	2	
VD77	VDS40 N	2.14	
VD78	VDS40 T	2.07	
VD83	VDS41 N	2.02	
VD84	VDS41 T1 rectosigmoid	2.07	
VD85	VDS41 T2 B transverse	1.82	
VD79	VDS42 N	2.1	
VD80	VDS42 T	2.11	
VD81	VDS43 N	0.998	
VD82	VDS43 T	2.04	
VD86	VDS44 N	0.414	
	VDS44 T		
VD87	VDS45 N	1.11	
VD88	VDS45 T	1.47	
VD89	VDS46 N	1.24	
VD90	VDS46 T	2.18	
VD91	VDS47 N	1.33	

VD92	VDS47 T	1.24	
VD115	VDS48 N	0.96	
VD116	VDS48 T	1.51	
VD53	VDS49 N	2.07	
	VDS49 T		
VD93	VDS50 N	1.3	
VD94	VDS50 T	1.09	
VD95	VDS51 N	1.4	
VD96	VDS51 T	1.74	
	VDS52 N		Patient did not have surgery
	VDS52 T		
VD117	VDS53 N	1.17	
VD118	VDS53 T	1.48	
VD97	VDS54 N	1.44	
VD98	VDS54 T	1.77	
VD99	VDS55 N	1.75	
	VDS55 T		
VD100	VDS56 N	1.86	
VD101	VDS56 T	1.15	
VD105	VDS57 N	0.816	
VD106	VDS57 T	1.54	

VD107	VDS58 N	1.27	
VD108	VDS58 T	0.59	
VD109	VDS59 N	1.42	
VD110	VDS59 T	0.97	

APPENDIX 5b

RNA quality

RNA integrity number

An algorithm is used to determine the RNA integrity numbers (RINs), which account for the entire electrophoretic trace of the RNA as well as the ratio of 28s to 18s rRNA. The RIN scale ranges from 0 to 10, with 10 denoting the maximum RNA integrity. As RNA degradation progresses, peak heights for 28s and 18s rRNA become increasingly diminished while degraded RNA peaks become more prominent. The 28s and 18s peaks will be barely visible where there is significant degradation. RNA integrity was assessed using the Agilent 2100 Bioanalyser (Agilent Technologies, Santa Clara, USA).

The RNA extracted from the patient samples were to be used for gene expression analysis regardless of integrity. There were two main reasons for this. Firstly, these specimens are difficult to access and thus we needed to use what we had and secondly, because even if the RINs are low the RNA can still generate good data (L. McNoe, personal communication, 2015). However, RINs were measured for a small number of our sample. We wanted to assess whether tissue samples treated with RNAlater® overnight fared better than those frozen down directly at -80^{o}C. Three normal and tumour pairs were examined. We had samples for the three patients that were treated and untreated. Following are the RINs measured for the samples both before and after baseline adjustment. Several of the RINs became not applicable (NA) after baseline adjustment. Treatment with RNAlater® appeared to make little difference but this was a small sample. Adjusted RINs were all low signifying low RNA integrity. Abbreviations: N= normal, T= tumour.

Study Number	Treatment	RIN pre-adj	RIN post-adj
VDS01N	RNAlater®	4.6	N/A
VDS01T	RNAlater®	2.7	2.7
VDS01N	-80°C	2.8	2.8
VDS01T	-80°C	2.7	2.7
VDS02N	RNAlater®	2.3	2.3
VDS02T	RNAlater®	2.6	2.6
VDS02N	-80°C	3.9	3.9
VDS02T	-80°C	6.5	N/A
VDS05N	RNAlater®	4.6	N/A
VDS05T	RNAlater®	3.1	N/A
VDS05N	-80°C	2.6	N/A
VDS05T	-80°C	2.7	2.7

APPENDIX 6

Heatmaps for top 100 genes in normal, tumour and paired tissue analyses

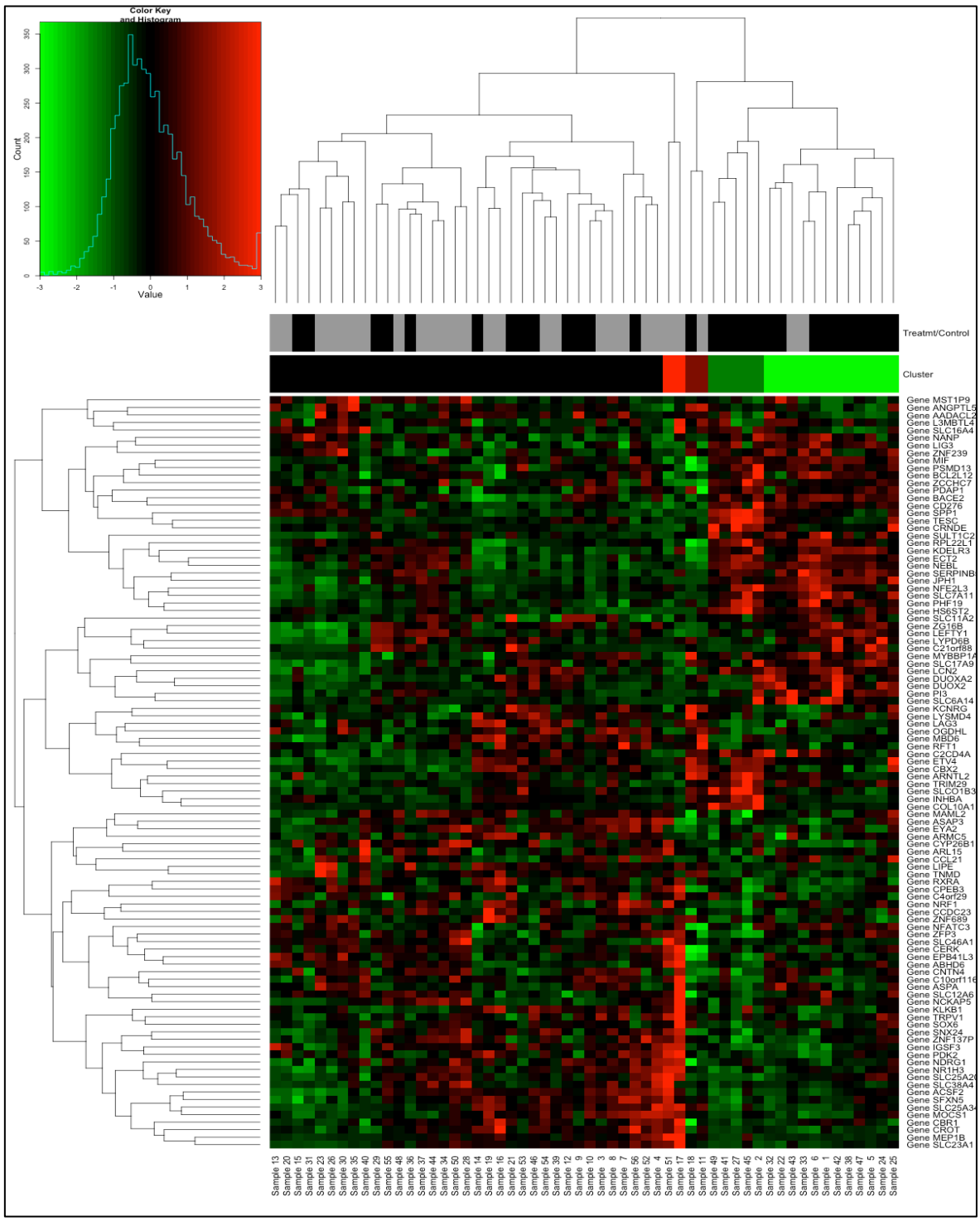


Figure 4A. Heatmap of the top 100 genes expressed in the single gene analysis in the **NORMAL** tissue. Treatmt = treatment; black indicates the treatment arm of the study. The cluster colour bar has been coded to demonstrate the 5 main clusters in the sample dendrogram. The red colouring indicates elevation in gene expression. Conversely, the green colouring indicates reduction in gene expression. The intensity of colour is directly related to the magnitude of elevation or reduction of gene expression.

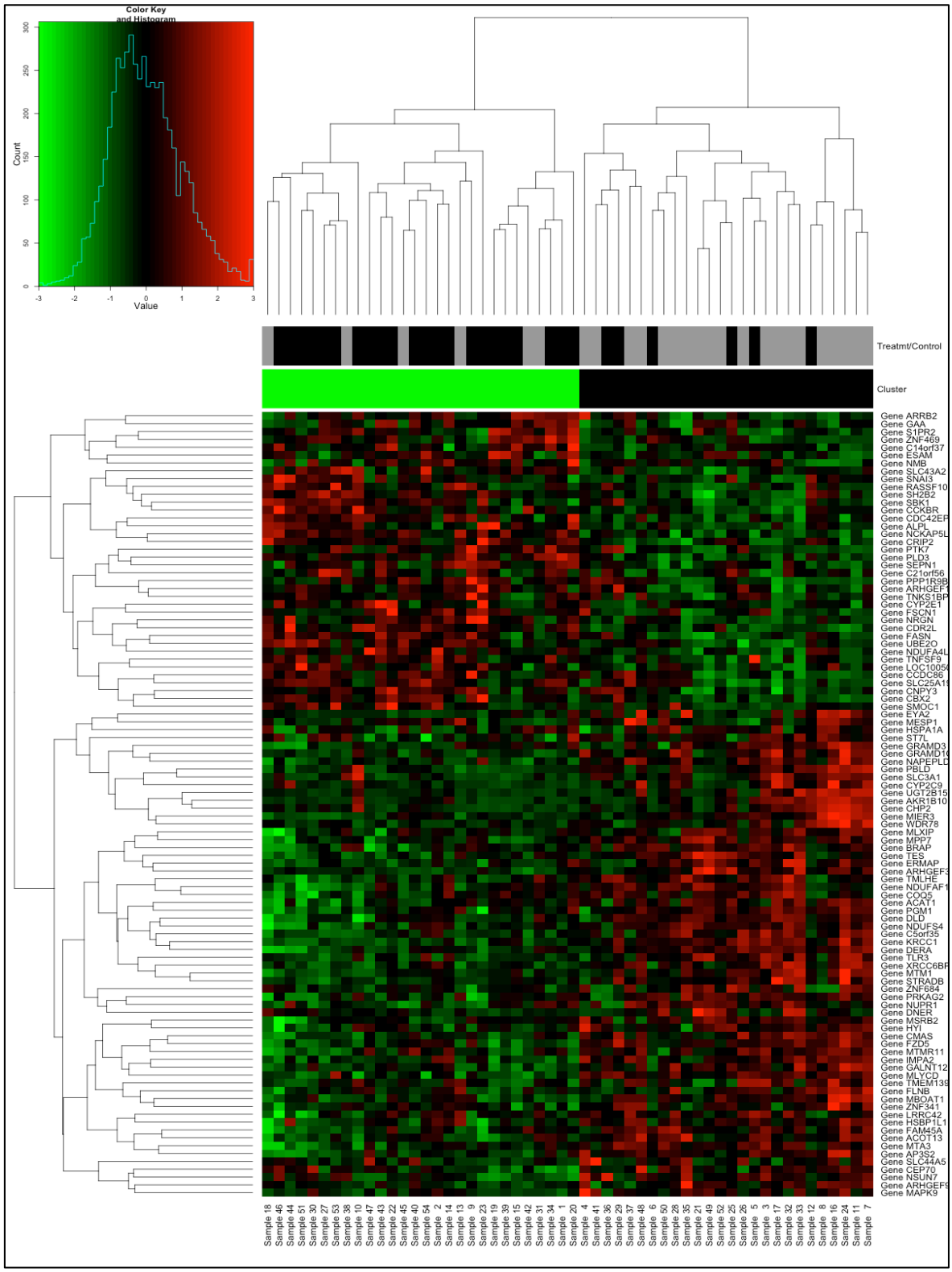


Figure 4B. Heatmap of the top 100 genes expressed in the single gene analysis in the **TUMOUR** tissue. Treatmt = treatment; black indicates the treatment arm of the study. The cluster colour bar has been coded to demonstrate the 2 main clusters in the sample dendrogram to highlight the separation of the treatment and control groups into two clusters. The red colouring indicates elevation in gene expression. Conversely, the green colouring indicates reduction in gene expression. The intensity of colour is directly related to the magnitude of elevation or reduction of gene expression.

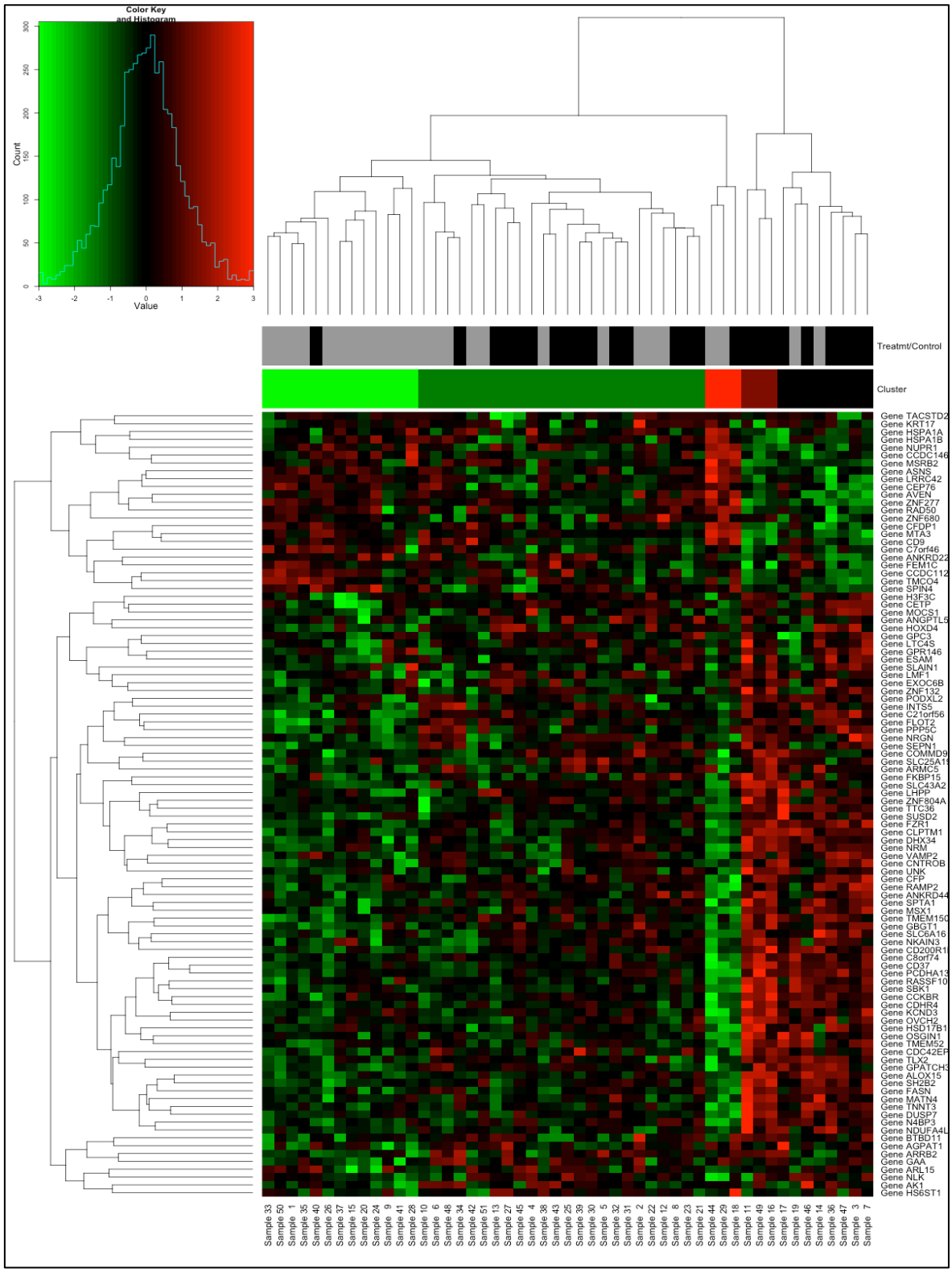
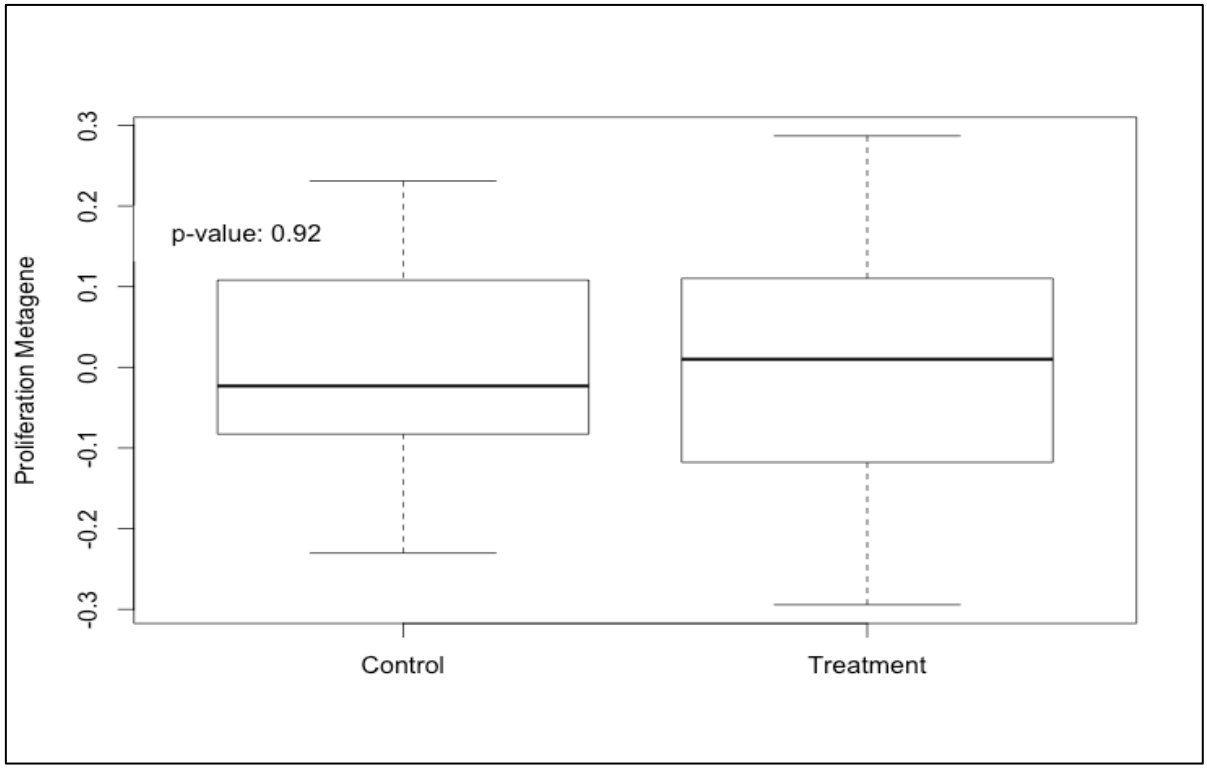


Figure 4C. Heatmap of the top 100 genes expressed in the single gene analysis in the **PAIRED** tissue. Treatmt = treatment; black indicates the treatment arm of the study. The cluster colour bar has been coded to demonstrate the 5 main clusters in the sample dendrogram. The red colouring indicates elevation in gene expression. Conversely, the green colouring indicates reduction in gene expression. The intensity of colour is directly related to the magnitude of elevation or reduction of gene expression.

APPENDIX 7



Proliferation metagene across the treatment and placebo groups. Boxplots of the proliferation metagene across the treatment and placebo arms.