

https://dspace.org/

Identifying potential biomarkers for colorectal cancer diagnosis using an RNA-seq analysis workflow

Kader, Zubayr

2021

https://hdl.handle.net/10539/32292

Kader, Zubayr (2020) Identifying potential biomarkers for colorectal cancer diagnosis using an RNA-Seq analysis workflow, University of the Witwatersrand, Johannesburg, <http://hdl.handle.net/10539/32292>

Downloaded from a DSpace repository. https://dspace.org/

Identifying Potential Biomarkers For Colorectal Cancer Diagnosis Using An RNA-Seq Analysis Workflow

by

Zubayr Kader



UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

Dissertation for the degree of

Masters of Science

in the Faculty of Science at the University Of The Witwatersrand

Supervisor: Prof. M. Kaur

December 2020

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by University Of The Witwatersrand will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: 04 February 2021 .....



Copyright © 2020 University Of The Witwatersrand
All rights reserved.

Abstract

In South Africa, colorectal cancer (CRC) is the second most common cancer in men and the third most common cancer in women, located at the lower end of the digestive system, in the colon and rectum, and is most accurately diagnosed through colonoscopy procedures. The prevalence of CRC is on the rise globally as well as locally, yet participation rates for screening tools remain low. In the present study, bioinformatics which is a multidisciplinary field that is focused on interpreting biology through the analysis of gene sequences and protein expression, was employed to explore the molecular biology of patient CRC data as well as to identify dysregulated genes as potential biomarkers to be used as an alternative screening tool. This was performed by creating an RNA-Seq analysis workflow, that identified differentially expressed genes that were further used in functional analysis to identify biological processes and pathways relating to CRC onset and progression in patients. The genes were tested as biomarkers *in silico* using statistical tests and blood expression data and the genes identified included *COL11A1, INHBA, CLDN1, ETV4* and *FOXQ1* as potential tissue biomarkers, and *MMP1, CTHRC1, KRT17* and *IGFBP1* as potential blood biomarkers. The identified biomarkers require future wet lab validation and illustrates the potential for a novel CRC screening test to reduce the dependency on traditional tools that are ineffective due to poor patient participation and associated cost.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof Mandeep Kaur, for her mentorship and guidance throughout the year and for this dissertation. I wish to thank the School of Molecular and Cell Biology for granting me this opportunity. Further thanks to all members of the Integrated Cancer Biology Research lab for their support, as well as to the NRF for financial support. Most importantly, I wish to acknowledge and thank my family, for their support and love during the year, and to the role of faith in Allah (SWT) that ultimately gave me the strength to successfully complete this degree and thesis.

Contents

Declaration i

Abstract ii

Acknowledgements iii

Contents iv

List of Figures vii

List of Tables xii

Nomenclature xvi

1 Introduction 1

1.1 Epidemiology and Research 1

1.2 The Biology of CRC 2

1.2.1 Pathways of CRC Development 2

1.2.2 Drug Resistance and Dietary Influence 4

1.3 Transcriptomics and CRC Research 6

1.3.1 Transcriptomics and CRC Subtype Classification 7

1.3.2 Transcriptomics and CRC Diagnosis 8

1.3.2.1 Current Screening Tests 10

1.3.2.2 The Need for Alternative Screening Methods . . 10

1.3.3 Bioinformatics for Biomarker Identification in CRC . . . 12

1.3.4 Introduction to Present Study 14

1.4 Aims 14

1.4.1 Objectives 14

2 Materials and Methods 16

2.1 Acquisition of RNA-Seq Data 16

2.2 Pre-Processing of Data . . . 18
2.2.1 Alignment and Quantification . . . 19
2.3 Differential Expression . . . 20
2.3.1 Background . . . 20
2.3.2 Differential Expression in R . . . 21
2.4 Functional Analysis . . . 23
2.4.1 GO Analysis . . . 23
2.4.2 KEGG Analysis . . . 24
2.5 WGCNA . . . 25
2.6 Validation . . . 27
2.6.1 Validation using CoReCG . . . 27
2.6.2 Validation using Oncomine . . . 27
2.7 Survival Analysis . . . 28
2.8 Drug Gene Interaction . . . 29
2.9 Biomarker Testing . . . 29
2.10 Galaxy . . . 31
2.11 Programs and Code Information . . . 32

3 Results 33
3.1 Quality Control . . . 33
3.2 Differential Expression . . . 35
3.3 Gene Processes . . . 46
3.4 Co-Expression . . . 53
3.5 Validation . . . 57
3.6 Survival Analysis and Stage Expression . . . 58
3.7 Drug Gene Interactions . . . 60
3.8 Biomarker Testing . . . 60

4 Discussion 63
4.1 Identified Gene Signatures . . . 64
4.1.1 Identified Primary CRC Gene Signature . . . 64
4.1.2 Identified Liver Metastases Gene Signature . . . 66
4.2 Identified Biological Processes and Pathways in CRC . . . 67
4.2.1 Gene Ontology Based Processes . . . 67
4.2.2 Identified Pathways Perturbed in CRC . . . 68
4.3 *In silico* Validation of the Identified Gene Signatures . . . 71
4.4 Identified Potential CRC Biomarkers . . . 72
4.4.1 Possible Blood Biomarkers for CRC . . . 73

4.4.2 Possible Tissue Biomarkers for CRC 74
4.5 Conclusion 74
4.6 Limitations of the Study and Future Recommendations 77

Appendices **79**

A Additional DE and Co-Expression Results **80**

B Individual Survival Analysis Plots **83**

C Individual Stage Expression Plots **90**

D Additional Validation Results **103**
D.1 CoReCG and Oncomine 103
D.2 Blood Biomarker Results 108

Bibliography **112**

List of Figures

1.1 A diagram summarising the three pathways of CRC development. A: Represents the classical pathway. B: Represents the MSI alternative pathway. C: Represents the serrated alternative pathway. Pathways B and C can present together. Taken from Mundade *et al.* (2014). 4

1.2 Figure showing the downstream pathways of EGFR, which represent possible targets for CRC therapy. PI3K may play a role in anti-EGFR drug resistance. Figure taken from Harbison *et al.* (2011). 5

1.3 A typical RNA-Seq workflow to prepare for downstream analysis. Taken from Mackenzie (2018) 8

1.4 The correlation between epithelial mesenchimal transition (EMT) and drug resistance, taken from Shibue and Weinberg (2017). The reverse process of EMT is mesenchymal epithelial transition (MET). Drug resistance and invasiveness have a higher correlation for the EMT process. 9

1.5 Figure displaying the broad overview of the characterisation of CRC subtypes through transcriptomic study. EMT = Epithelial mesenchymal transition. CIMP = CpG island methylation. MSI = Microsatellite instability. MSS = Microsatellite stability. CMS = Consensus molecular subtype. Taken from (Nakanishi *et al.*, 2019). 9

1.6 Figure showing the categories of biomarkers that can be measured from a primary CRC tumour, including proteins, RNAs and genetic mutations. 12

2.1 A flowchart of the workflow followed in this NGS analysis of the existing CRC RNA-Seq data in GSE50760. Alignment was performed using two traditional alignment tools and a pseudo-alignment tool. Downstream analysis was performed in R and using web tools. . . . 16

2.2 Figure showing the matrix layout for DESeq2 input. Genes were listed in the first column, with the samples listed as row headers. The count values populated the table according to gene and sample. 22

3.1 Figure showing the mean quality score for all the reads for each base. This quality score is included in the FASTQ file after sequencing, and represents the predicted accuracy of a correct base call. 33

3.2 Figure showing the average quality score per sequence (x-axis) according to the Phred score and the respective sequence counts (y-axis). This quality score is included in each FASTQ file and represents the number of sequences that have a high accuracy of having correct base calls. Red indicates a poor quality score of a Phred score less than 20, yellow indicates a Phred score that should be investigated and green indicates a good Phred score. The majority of sequences fall in the green block. 34

3.3 Figure showing the overall percentage GC content in the FASTQ files. The GC content can then be compared to the organisms expected GC percentage. *Homo sapiens* expected GC percentage is 46.1 ± 9%. Lines in red and yellow represent outlier samples. . . 35

3.4 A PCA plot using rlog normalised data from the Salmon method along PC1 and PC2. The blue dots indicate the Normal samples, the red dots indicate the primary CRC tumour samples, with the green dots representing the liver metastases samples. The percentage variance indicates the variance of the samples across principal components 1 and 2, the most variable of the components, which then represent variability of the samples from each other. 37

3.5 An MDS plot using rlog normalised data from the Salmon method along two dimensions of Euclidean distance differences, 1 (x-axis) and 2 (y-axis). The Euclidean distances were calculated from the rlog data, and show the relationship between samples and how they differ. The blue dots indicate the Normal samples, the red dots indicate the primary CRC tumour samples, with the green dots indicating the liver metastases samples. 38

3.6 Figures showing the volcano plots for the different contrasts, (a) CRC vs Normal, (b) Metastasis vs CRC, and (c) Metastasis vs Normal. Each dot represents a gene and their relative log_2FC. Grey dots indicate genes that had no significant change, green dots indicate genes that met the log_2FC threshold of more than 1.5, blue dots indicate genes that were significantly DE ($p<0.05$) whereas red dots indicate genes that met both the thresholds. 43
3.7 Figures showing the normalised count spread of the top 20 most DE genes according to p-value in each contrast, (a) CRC vs Normal, (b) Metastasis vs CRC, and (c) Metastasis vs Normal. The genes are labelled with their known gene symbols. 45
3.8 Figures showing bar plots for the Top 30 most significant GO biological processes for the different contrasts, (a) CRC vs Normal, (b) Metastasis vs CRC, and (c) Metastasis vs Normal. The GO analysis maps the DE genes to their known biological processes. Processes are given a colour according to adjusted p-value significance, with red being the most significant and blue being the least significant. The x-axis shows how many genes mapped to each GO term. Processes in red blocks indicate processes of interest in the present study. 49
3.9 Figures showing bar plots for the top 15 KEGG enriched pathways for all the DE genes for each contrast, (a) CRC vs Normal, (b) Metastasis vs CRC, and (c) Metastasis vs Normal. Pathways are given a darker blue according to their increasing significance. The x-axis show the number of genes in the gene set that were mapped to the pathway on the y-axis. Pathways in red blocks indicate pathways of interest in the present study. 51
3.10 Figures showing the GO analysis bar plot for each contrast, (a) CRC vs Normal, (b) Metastasis vs CRC, and (c) Metastasis vs Normal as output from WebGestalt. The input gene signatures were Tables 3.3, 3.4 & 3.5. The x-axis represent the GO terms for each category. The y-axis represent how many genes of the signatures mapped to each term, with the number displayed at the top of each bar. 52
3.11 Figure showing the samples in a dendrogram with the outlier removed and a trait heatmap added for visibility showing how the expression profiles of each run (SRR) relate to each other. 54

3.12 Figure showing two plots with which the chosen soft threshold values were used in a scale free topology plot and a mean connectivity plot. A soft threshold of 4 was chosen. A solid line is plotted at 0.825. The left panel represents the co-expression similarity as a function of soft threshold. The right panel represents the mean connectivity as a function of soft threshold. The lowest soft threshold value where R^2 stabilises above 0.8 is 4, which was selected for the analysis. 55

3.13 Figure showing the gene dendrogram along with a group heatmap in order to visualise modules that correlate with the sample group. Red indicates high correlation. 56

3.14 Figures showing the Kaplan-Meier overall survival graphs for the common genes for each contrast (a) CRC vs Normal, (b) Metastasis vs CRC, and (c) Metastasis vs Normal, using GEPIA2 and TCGA COAD, READ and LIHC studies as reference. Solid lines indicate the survival analysis results, with the dotted lines indicating the 95% confidence interval. HR represents the hazard ratio using the Cox proportional hazards model. The n number represents the number of samples used in the analysis for high and low expression. 59

4.1 Figure showing how activation of the complement system within the TME enhances cancer cell metastasis and proliferation. The release of complement proteins and activation of their receptors lead to reduced cytotoxic T-cell ability through myeloid derived suppressor cells (MDSCs), as well as inhibiting interleukin-10 (IL-10) release from tumour infiltrating lymphocytes. Figure taken from Afshar-Kharghan and Others (2017). 70

4.2 Diagram showing the analysis workflow used in this study. 76

A.1 Figure showing the PCA plot using sample names instead of dots, from original Figure 3.4. 81

A.2 Figure showing the cluster dendrogram with the red line indicating the cut-off point for outliers, leading to Figure 3.11. Here, sample SRR75594 was removed from analysis. 82

B.1 Individual survival plots for Normal vs CRC Significant genes in Table 3.3. Red line indicates high expression. Green line indicates low expression. 85

B.2 Individual survival plots for CRC vs Metastasis Significant genes in Table 3.4. Red line indicates high expression. Green line indicates low expression. 87
B.3 Individual survival plots for Normal vs Metastasis Significant genes in Table 3.5. Red line indicates high expression. Green line indicates low expression. 89

C.1 Individual stage expression plots for Normal vs CRC Significant genes in Table 3.3 using GEPIA2. The genes were mapped against TCGA COAD and READ. 92
C.2 Individual stage expression plots for Normal vs CRC Significant genes in Table 3.3 using GEPIA2. The genes were mapped against TCGA LIHC to indivate their presence in liver metastases. 94
C.3 Individual stage expression plots for CRC vs Metastasis Significant genes in Table 3.4 using GEPIA2. The genes were mapped against COAD and READ. 96
C.4 Individual stage expression plots for CRC vs Metastasis Significant genes in Table 3.4 using GEPIA2. The genes were mapped against LIHC. 98
C.5 Individual stage expression plots for Normal vs Metastasis Significant genes in Table 3.5 using GEPIA2. Genes were mapped against TCGA COAD and READ. 100
C.6 Individual stage expression plots for Normal vs Metastasis Significant genes in Table 3.5 using GEPIA2. Genes were mapped against TCGA LIHC. 102

List of Tables

1.1 Table showing the multiple existing biomarker examples and their uses in cancer. Table reproduced from (Henry and Hayes, 2012) . . 11

2.1 Table showing the dataset accession numbers that were used to acquire the SRA files from NCBI GEO. There were 54 runs in total and each sample group had 18 runs. 18
2.2 Table showing the programmes that were used in the RNA-Seq analysis workflow for quality control, alignment and quantification. 18

3.1 Table showing the gene counts for each quantification method before and after manual filtering of genes with <1 count across all samples. 36
3.2 Table showing the significant genes that were differentially expressed in the different contrasts and methods. These genes met two criteria in order to be labelled as significant, $p < 0.05$ and $log_2FC > 1.5$. . . 39
3.3 Table showing the common top DE genes amongst the three methods for CRC vs Normal and their relative descriptions, log_2FC and p-value. This list represents signature Sig1. 40
3.4 Table showing the commonly DE genes amongst the three methods for Metastasis vs CRC and their relative descriptions, log_2FC and p-value. This list represents signature Sig2. 41
3.5 Table showing the commonly DE genes amongst the three methods for Metastasis vs Normal and their relative descriptions, log_2FC and p-value. This list represents signature Sig3. 42
3.6 Table showing validation of the genes in the "CRC vs Normal" contrast and whether they are present in the CoReCG and Oncomine databases for CRC. 57
3.7 Table showing validation of the genes in the "Metastasis vs CRC" contrast and whether they are present in the CoReCG and Oncomine databases for CRC and liver cancer respectively. 58

3.8 Table showing validation of the genes in the "Metastasis vs Normal" contrast and whether they are present in the CoReCG and Oncomine databases for CRC and liver cancer respectively. 58
3.9 Table showing the known drug gene interactions using the common significant genes. Source list shows from which signature the gene is from, C = Cancer, N = Normal and M = Metastasis. C v N represents Sig1, M v C represents Sig2. 60
3.10 Table showing the biomarker testing results for the genes using Oncomine data. Avg = Average, Sens = Sensitivity, Spec = Specificity, Prec = Precision. 61
3.11 Table showing the genes that had positive meta score values for both DE and expression abundance, which represent whether a gene is both distinguishable and detectable, and from which blood sample (peripheral or EVs) they were measured as well as which gene signature they initially belonged to: C v N represents Sig1, M v N represents Sig3. 62

D.1 Table showing the primary CRC genes (Table 3.3) validated on CoReCG along with the mined summaries of the genes interaction with CRC and the sources from where the descriptions were mined. 104
D.2 Table showing the detailed Oncomine validation results for Normal vs CRC. 105
D.3 Table showing the liver metastases (Table 3.4) that were validated on CoReCG along with the mined summaries of the gene's interaction with CRC and the sources from where the descriptions were mined. 105
D.4 Table showing the detailed Oncomine validation results for CRC vs Metastasis. 106
D.5 Table showing the liver metastases genes (Table 3.5) validated on CoReCG along with the mined summaries of the genes interaction with CRC and the sources from where the descriptions were mined. 106
D.6 Table showing the detailed Oncomine validation results for CRC vs Metastasis. 107

D.7 Table showing the meta scores for differential gene expression taken from different studies comparing the blood from CRC patients to a normal patients. The genes used were from the "CRC vs Normal" contrast. The second column indicates expression from peripheral blood, whereas the third column indicates expression for extracellular vesicles in the blood. "-" indicates genes that were found but had no expression data. "N/A" indicates genes that were not found on the database. 108

D.8 Table showing the meta scores for differential gene expression taken from different studies comparing the blood from CRC patients to a normal patients. The genes used were from the "Metastasis vs CRC" contrast. The second column indicates expression from peripheral blood, whereas the third column indicates expression for extracellular vesicles in the blood. "-" indicates genes that were found but had no expression data. 109

D.9 Table showing the meta scores for differential gene expression taken from different studies comparing the blood from CRC patients to a normal patients. The genes used were from the "Metastasis vs Normal" contrast. The second column indicates expression from peripheral blood, whereas the third column indicates expression for extracellular vesicles in the blood. "-" indicates genes that were found but had no expression data. 109

D.10 Table showing the meta scores for gene expression abundance taken from different studies comparing the blood from CRC patients to a normal patients. The genes used were from the "CRC vs Normal" contrast. The second column indicates expression from peripheral blood, whereas the third column indicates expression for extracellular vesicles in the blood. "-" indicates genes that were found but had no expression data. "N/A" indicates genes that were not found on the database. 110

D.11 Table showing the meta scores for gene expression abundance taken from different studies comparing the blood from CRC patients to a normal patients. The genes used were from the "Metastasis vs CRC" contrast. The second column indicates expression from peripheral blood, whereas the third column indicates expression for extracellular vesicles in the blood. "-" indicates genes that were found but had no expression data. 110

D.12 Table showing the meta scores for gene expression abundance taken from different studies comparing the blood from CRC patients to a normal patients. The genes used were from the “Metastasis vs Normal” contrast. The second column indicates expression from peripheral blood, whereas the third column indicates expression for extracellular vesicles in the blood. "-" indicates genes that were found but had no expression data. 111

Nomenclature

Abbreviations

AGE	Advanced glycation end products
AJCC	American Joint Committee on Cancer
APC	Adenomatous polyposis coli
BP	GO Biological Processes
C	Cytosine
CANSA	Cancer Association of South Africa
CC	GO Cellular Components
cDNA	Complementary DNA
CEA	Carcinoembryonic antigen
cfDNA	Cell free DNA
CIMP-H	CpG island methylation phenotype
CIN	Chromosomal instability
CMS	Consensus molecular subtype
COAD	Colon adenocarcinoma
CoReCG	Colon Rectal Cancer Gene Database
CRC	Colorectal cancer
CSS	Colon cancer subtype
CTC	Computerised tomography colonography
DAVID	Database for Annotation, Visualisation and Integrated Discovery
DE	Differential expression / Differentially expressed
DGIdb	Drug Gene Interaction Database
DNA	Deoxyribonucleic acid
EBI	European Bioinformatics Institute
ECM	Extracellular matrix
EGFR	Epidermal growth factor receptor

EMT	Epithelial-mesenchymal transition
ERK	Extracellular signal-regulated kinase
EV	Extracellular vesicle
FC	Fold change
FTP	File transfer protocol
G	Guanine
GEO	Gene Expression Omnibus
GO	Gene Ontology
GPL	GEO Platform
GSE	GEO Series
GSEA	Gene set enrichment analysis
GTF	Gene transfer file
GUI	Graphical user interface
HCC	Hepatocellular carcinoma
HIF-1α	Hypoxia-inducible factor 1 alpha
HTS	High throughput sequencing
IBD	Inflammatory bowel disease
KEGG	Kyoto Encyclopedia of Genes and Genomes
KM	Kaplan-Meier
LDL	Low density lipoprotein
LIHC	Liver hepatocellular carcinoma
MAPK	Mitogen activated protein kinase
MDS	Multidimensional scaling
MDSC	Myeloid derived suppressor cells
MF	GO Molecular Functions
MLH1	MutL Homolog 1
MM	Module membership
mRNA	Messenger RNA
MSI	Microsattelite instable
MSS	Microsattelite stable
NCBI	Naitonal Centre for Biotechnology Information
NGS	Next generation sequencing

ORA	Over representation analysis
PAM	Paritioning around medoids
PCA	Principle component analysis
PCR	Polymerace chain reaction
QC	Quality control
RAGE	AGE Receptor
READ	Rectal adenocarcinoma
rlog	Regularised log transformation
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
ROS	Reactive oxygen species
rRNA	Ribosomal RNA
SA	South Africa / South African
SAM	Sequence alignment map
Sig1	Represents the CRC vs Normal contrast
Sig2	Represents the Metastasis vs CRC contrast
Sig3	Represents the Metastasis vs Normal contrast
SRA	Sequence Read Archive
TCGA	The Cancer Genome Atlas
TME	Tumour microenvironment
TNF	Tumour necrosis factor
tRFRNA	Transfer RNA fragments
vst	Variance stabilising transformation
WGCNA	Weighted gene correlation network analysis
Wnt	Wingless-related integration site

Chapter 1

Introduction

1.1 Epidemiology and Research

Cancer continues to place a significant burden on modern society. In 2019, the American Cancer Society reported that there were more than 1,7 million new cases of cancer with a mortality rate of above 606,000 in the US alone (Siegel *et al.*, 2019). Globally, there were 1.8 million new colorectal cancer (CRC) cases with 881,000 deaths in 2018 (Bray *et al.*, 2018). In the US, CRC is the second most common cause of cancer death amongst men and women (Siegel *et al.*, 2020*a*). In South Africa (SA) it is the fourth most common cancer overall, and the sixth in cancer-related mortality, making up 6.5% of all diagnosed cancers in 2018 (Brand *et al.*, 2018). However, according to recent data from the Cancer Association of South Africa (CANSA), CRC is currently the second most common cancer in SA men, and the third most common cancer in SA women (`https://www.cansa.org.za`). CRC is thought to be less common in developing countries, as it is usually associated with older people on Western-diets, however there has been a marked increase in the incidence of the disease in SA over the past decade (Motsuku *et al.*, 2020). Furthermore, while normally less common in black populations, evidence has shown there are disproportionately large numbers of younger black patients presenting with the disease over time (Motsuku *et al.*, 2020).

With this in mind, CRC research in SA could provide many benefits in terms of diagnosis and treatment of CRC given the increasing incidence rate. The present MSc project aims to develop an in-house ribonucleic acid sequencing (RNA-Seq) data analysis workflow that will be used in a future study to analyse SA CRC patients' tumour derived RNA-Seq data with the aim to identify potential CRC biomarkers.

1.2 The Biology of CRC

CRC is cancer of the colon or rectum, located at the lower end of the digestive system (Siegel *et al.*, 2017). CRC development has been categorised into four stages by the American Joint Committee on Cancer (www.cancerstaging.org): stage I CRC, where colon cells proliferate to form benign polyps, which transition into stage II through somatic mutations contributing to malignancy. In stage III, the tumour cells gradually increase in number, accumulating further mutations with the invasion of the surrounding muscle wall of the colon. Stage IV represents the metastasis of the cells into the surrounding blood vessels and lymphatic system (Haggar and Boushey, 2009; Siveen *et al.*, 2019).

Early diagnosis of CRC, like any cancer, is an important contributor in the overall treatment of the disease. Symptomatic diagnosis is often difficult, as the symptoms presented such as changes in bowel habit and diarrhoea, are usually non-specific with poor sensitivity for diagnosis of CRC (Vega *et al.*, 2015). Additionally, patients often delay seeking a specialist consultation, assuming the symptoms are caused by another illness. Proper diagnosis of CRC therefore only begins after a specialist has performed a colonoscopy, the gold standard diagnostic tool for CRC, and confirmed the presence of a tumour. Following diagnosis, treatment of CRC is dependent on the stage of progression with early stages resulting in colectomy (surgical removal of colon tissue), and stages II and above being treated with chemotherapy and radiation therapy (Mundade *et al.*, 2014).

The initial development of the early stage polyps can be attributed to two main pathways; the classical pathway and the alternative (serrated) pathway.

1.2.1 Pathways of CRC Development

The classical pathway of CRC development is associated with chromosomal instability (CIN) and mutations of multiple tumour suppressor genes and oncogenes, such as the adenomatous polyposis coli gene *APC* (van de Wetering *et al.*, 2015). The *APC* mutation contributes to a mutation in *KRAS*, which codes for the K-Ras protein, part of the RAS/MAPK (mitogen activated protein kinase) pathway regulating cell proliferation and differentiation. Part of the classical pathway is the adenoma-carcinoma sequence, which states that mutations in these genes contribute to the development of a pre-existing adenomatous polyp, or adenoma, which in turn acts as a precursor to CRC development (Jass *et al.*, 2002; Fearon, 2011). In the US, the prevalence of adenoma

polyps increases with age, with 25% at age 50 to 50% at age 70 (Marley and Nan, 2016; Fearon, 2011).

Studies have found that *APC* plays a central role in CRC development, and is effective in predicting overall survival, acting as the gatekeeper for the majority of CRCs (Schell *et al.*, 2016). The APC protein acts as an antagonist of the wingless-related integration site (Wnt) signalling pathway by regulating the β-catenin protein and exerting tumour-suppressing effects (Fearon, 2011). The inactivation of *APC* through mutation and subsequent activation of the Wnt/β-catenin signalling pathway is integral in adenoma initiation and CRC development (Mundade *et al.*, 2014). It is also believed that *APC* has functions in cellular processes such as cell migration, apoptosis and DNA repair, and that mutations initiate the adenoma-carcinoma sequence (Jaiswal and Narayan, 2008; Brocardo and Henderson, 2008; Fearon, 2011). The *KRAS* mutation in the classical pathway is also associated with dysregulation of the RAS signalling via the Raf–mitogen-activated protein kinase (MEK)–extracellular signal-regulated kinase (ERK) pathway, which is involved in the control of cell cycle progression (Mundade *et al.*, 2014). Mutations in *APC* and *KRAS* are associated with the poorest survival outcomes in CRC patients (Schell *et al.*, 2016).

The alternative pathway of CRC development hypothesises that tumours originate as a result of microsatellite instability (MSI), which are short tandem repeats found within the human genome, with instability and hypermutability arising from impaired DNA mismatch repair (Boland and Goel, 2010). MSI is present in 15% of all CRC diagnoses and the initial inactivation of mismatch repair genes can be attributed to aberrant methylation of the CpG promoter in repair gene MutL homolog 1 (*MLH1*), thought to occur through epigenetic silencing (Boland and Goel, 2010). This has led to CRC cases being classified as MSI, or microsatellite stable (MSS).

Furthermore, this alternative pathway of CRC development can be associated with the histological presence of serrated lesions or serrated polyps, and is associated with mutations in *BRAF* (protein kinase B-raf) leading to constitutive stimulation of the MAPK and ERK pathway, resulting in dysregulation and epigenetic silencing of genes associated with cell proliferation, differentiation, DNA repair and cell cycle control, giving rise to serrated lesions (Tuppurainen *et al.*, 2005; Mundade *et al.*, 2014; Nakanishi *et al.*, 2019). The overactive MAPK/ERK signalling reaches the nucleus DNA and induces enhanced cell proliferation, metastasis, immune system evasion, resistance to

apoptosis, and activation of hypoxia-inducible factor 1 α (HIF-1α) contributing to angiogenesis (Ascierto *et al.*, 2012).

The mutation of the lesions is associated with the aforementioned hypermethylation of the CpG island promoter regions (CpG island methylation phenotype; CIMP-H), resulting in epigenetic silencing of tumour suppressor genes such as *MLH1*. The combination of CIMP-H and MSI leads to CRC development, with the exact mechanism behind the initial *BRAF* mutation and the two events being unclear (Nakanishi *et al.*, 2019). The alternative pathway is also associated with *KRAS* mutation, similar to the classical pathway (Stefanius *et al.*, 2011). The described pathways of CRC development are summarised in Figure 1.1.

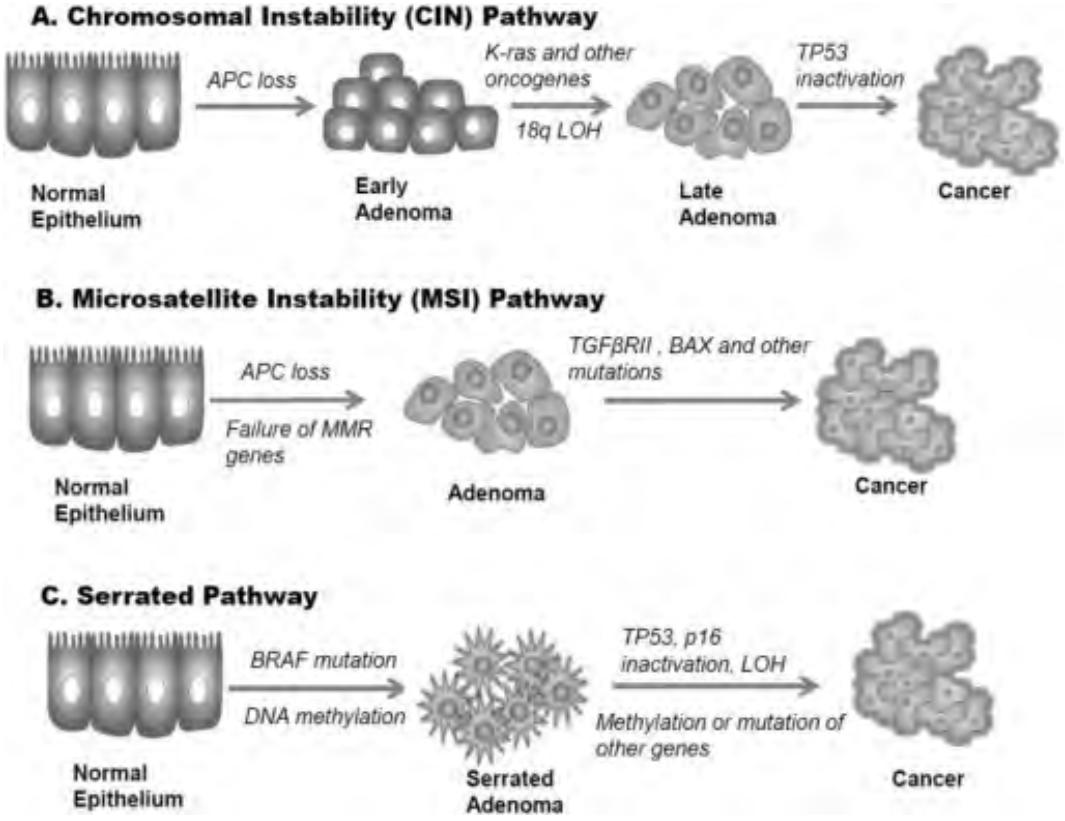


Figure 1.1: A diagram summarising the three pathways of CRC development. A: Represents the classical pathway. B: Represents the MSI alternative pathway. C: Represents the serrated alternative pathway. Pathways B and C can present together. Taken from Mundade *et al.* (2014).

1.2.2 Drug Resistance and Dietary Influence

Drug resistance is a major hurdle in cancer treatment, with almost half of metastatic CRC patients exhibiting some form of resistance (der Jeught *et al.*,

2018). Chemotherapeutic drugs used in treating CRC include 5-fluorouracil (5-FU), a topoisomerase I inhibitor irinotecan (CPT-11), oxaliplatin (often used in conjunction with 5-FU to reduce DNA repair in tumour cells), and capecitabine, a cytotoxic drug. Additionally anti-epidermal growth factor receptor (EGFR) antibodies such as Cetuximab have been used in late stage CRC and has shown to improve survival of patients with and without adjuvant chemotherapy (Jonker *et al.*, 2007; Bardelli and Siena, 2010). The EGFR signalling pathway plays a role in the MAPK/ERK signalling described above.

Retrospective analysis discovered that patients with *KRAS* mutations and those with wild type *KRAS* did not benefit from anti-EGFR therapy (Bokemeyer *et al.*, 2008; Bardelli and Siena, 2010). Believing there might be other factors in play, researchers looked at downstream signalling proteins of EGFR, such as *PI3KCA* (see Figure 1.2).

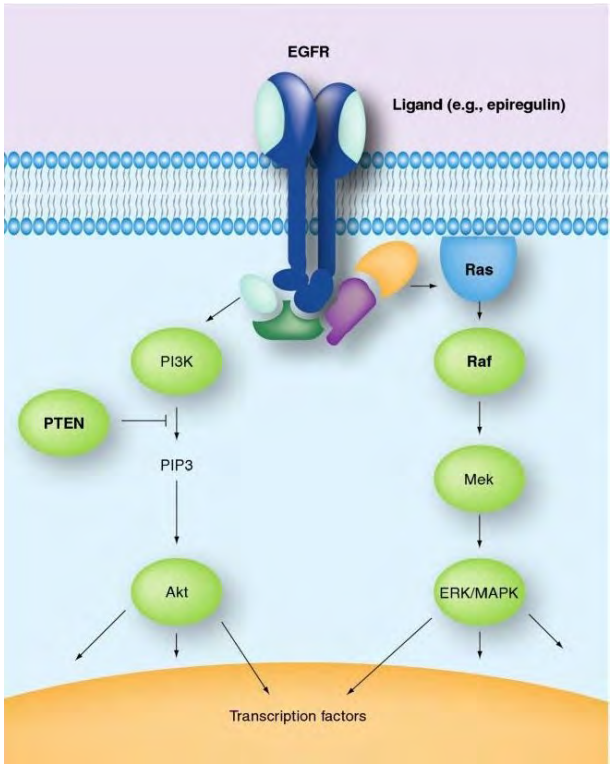


Figure 1.2: Figure showing the downstream pathways of EGFR, which represent possible targets for CRC therapy. PI3K may play a role in anti-EGFR drug resistance. Figure taken from Harbison *et al.* (2011).

The *PI3KCA* gene is mutated in about 20% of CRCs, and was also found to be present in non-responders containing wild-type *KRAS*, and could be considered as a predictor of anti-EGFR treatment resistance (Sartore-Bianchi *et al.*, 2009). The mutations in *PI3KCA* and loss of function mutations in

PTEN further contribute to CRC development through hyperactivation of *Akt*, leading to increased transcription of genes associated with apoptosis inhibition and increased cell growth and proliferation (Danielsen *et al.*, 2015). It was found that both *KRAS* and *PI3KCA* mutations have an independent role in the treatment resistance.

As CRC is also a lifestyle disease, researchers have looked into the role of diet on CRC development. Epidemiological evidence has linked CRC to dietary cholesterol intake (Jacobs *et al.*, 2012). It was found that Western diets had the highest risk for CRC development, having high saturated fat and cholesterol content. Jacobs *et al.* (2012) suggested that it is difficult to isolate individual elements in the diet and study their correlation with CRC when foods higher in cholesterol have links to other diseases such as cardiovascular disease and metabolic syndrome. On the other hand, serum cholesterol has been previously associated with CRC progression (Van Duijnhoven *et al.*, 2011), with high levels of low density lipoprotein cholesterol (LDL) also shown to be positively correlated with liver metastases in CRC patients (Wang *et al.*, 2017*a*). The LDL levels also elevated reactive oxygen species (ROS), contributing to an increase in the expression of genes associated with the MAPK pathway (Wang *et al.*, 2017*a*). The link between cholesterol and cancer continues to be researched.

1.3 Transcriptomics and CRC Research

Bioinformatics, a multidisciplinary field of study which deals with understanding and interpreting biology through the analysis of sequences, gene and protein expression as well as biological and chemical structures, has contributed extensively to cancer biology research. Transcriptomics, a branch of bioinformatics studying the transcriptome, includes methods such as DNA microarrays and next generation sequencing (NGS) in RNA-Seq.

Microarray is a widely used transcriptomic technique which makes use of a microchip with specific DNA probes attached (Kerr *et al.*, 2000). These probes correspond to cell messenger RNA (mRNA) which is reverse-transcribed to complementary DNA (cDNA) and fluorescently labelled to represent complementary sequences on the array. The fluorescence intensity is then measured which corresponds to the amount of transcripts that hybridise to a probe, illustrating the gene expression profile (Kerr *et al.*, 2000). The microarray is a relatively inexpensive transcriptomic tool that can be used for the profiling

of thousands of transcripts at once, however, it uses existing DNA probes, making them susceptible to bias as well as unable to detect novel transcripts (Mantione *et al.*, 2014). The analysis of the fluorescence intensity can also lead to differing results depending on the statistical methods used in quantifying expression in addition to interference with cross-hybridisation (Okoniewski and Miller, 2006). These limitations, along with the increased abundance of NGS technologies have made microarrays less common in research.

NGS has had a significant impact on the field of genetics research, from sequencing the entire human genome to transcriptome analysis with RNA-Seq (Van Dijk *et al.*, 2014). Whole transcriptomes can be produced by first isolating RNA from a cell or tissue of interest. As RNA is made up of different components, it is common for mRNA to be isolated and studied due to its direct role in protein synthesis, but other RNAs such as ribosomal RNA (rRNA) or micro-RNA (miRNA) can also be studied. Following RNA isolation, cDNA synthesis is performed using adaptors in order to create a cDNA library which is then sequenced on a high-throughput platform such as Illumina (Stark *et al.*, 2019). This generates millions of short sequences known as reads, which are mapped to a reference genome to produce the transcriptome and allows downstream analysis to take place. A typical RNA-Seq workflow is summarised in Figure 1.3. RNA-Seq provides many benefits over microarray, as it sequences the direct transcriptome ensuring greater accuracy and sensitivity in analysing gene expression. For example, microarrays can measure a 2-fold change in expression, whereas RNA-Seq is able to measure up to a 1.25 fold change reliably (Mantione *et al.*, 2014). Additionally, using proteomics, RNA-Seq was shown to more accurately represent absolute expression levels when compared to microarray (Fu *et al.*, 2009). This makes RNA-Seq an ideal candidate when studying gene expression.

1.3.1 Transcriptomics and CRC Subtype Classification

Transcriptomics in CRC has recently provided a more effective method at the classification of CRC, with subgroups and subtypes being discovered (Guinney *et al.*, 2015). At least two subgroups of CRC have been identified; (i) an inflammatory subgroup with a mutation in *BRAF* as well as MSI presence, and (ii) an immunosuppressive subgroup with mesenchymal characteristics (Nakanishi *et al.*, 2019).

Transcriptomics in CRC has been used previously to propose three CRC subtypes along with the abovementioned subgroups; colon cancer subtype 1

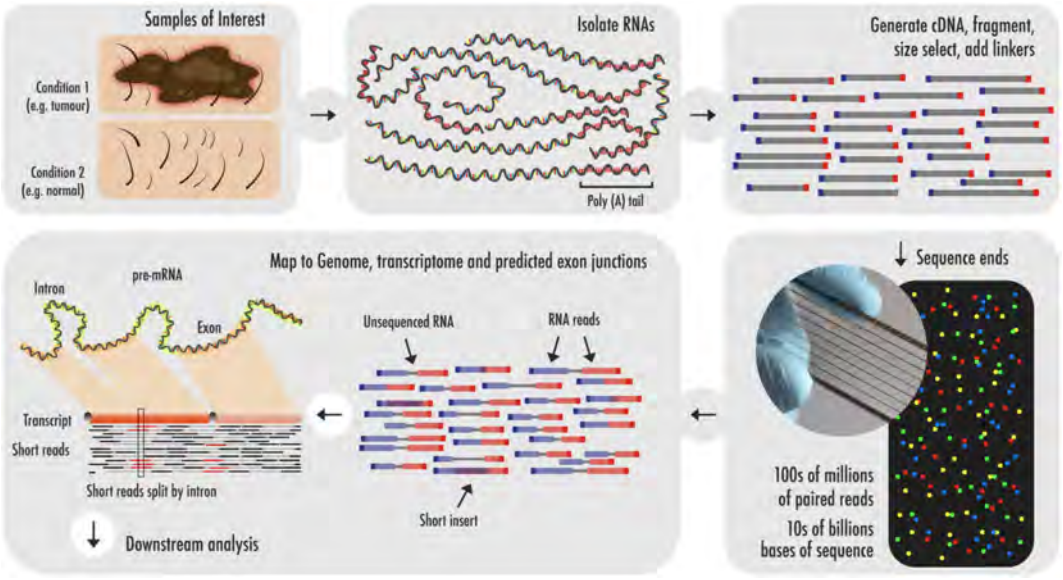


Figure 1.3: A typical RNA-Seq workflow to prepare for downstream analysis. Taken from Mackenzie (2018)

(CSS1) representing CIN, CSS2 representing MSI, and CSS3 representing MSS with mesenchymal phenotype which has the poorest prognosis as well as high drug resistance (Felipe De Sousa *et al.*, 2013). CSS3 is also associated with epithelial mesenchymal transitions (EMT), the process whereby epithelial cancer cells can transiently develop a mesenchymal phenotype by losing cell junctions, reorganising their cytoskeleton, downregulating epithelial gene expression and increasing motility which all ultimately contribute to enhanced migratory capacity and invasiveness, as well as resistance to apoptosis (Kalluri and Weinberg, 2009). EMT in cancer cells is associated with increased invasiveness, tumour-initiating ability and drug resistance (Figure 1.4) (Shibue and Weinberg, 2017).

Further transcriptome studies expanded upon the CSS work of Felipe De Sousa *et al.* (2013), and led to the consensus molecular subtypes (CMS) of CRC, containing two subgroups. The CMS1 is characterised by MSI and *BRAF* mutation, and the CMS4 is characterised by EMT genes thus having the poorest prognosis, summarised in Figure 1.5 (Guinney *et al.*, 2015).

1.3.2 Transcriptomics and CRC Diagnosis

Transcriptomics have been used extensively in the detailed research into the genetic and molecular classification of CRC, yet the avenue of diagnosis and detection has been less explored. This has become a topic in need of urgent addressing, as the increasing age of the global population is expected to translate

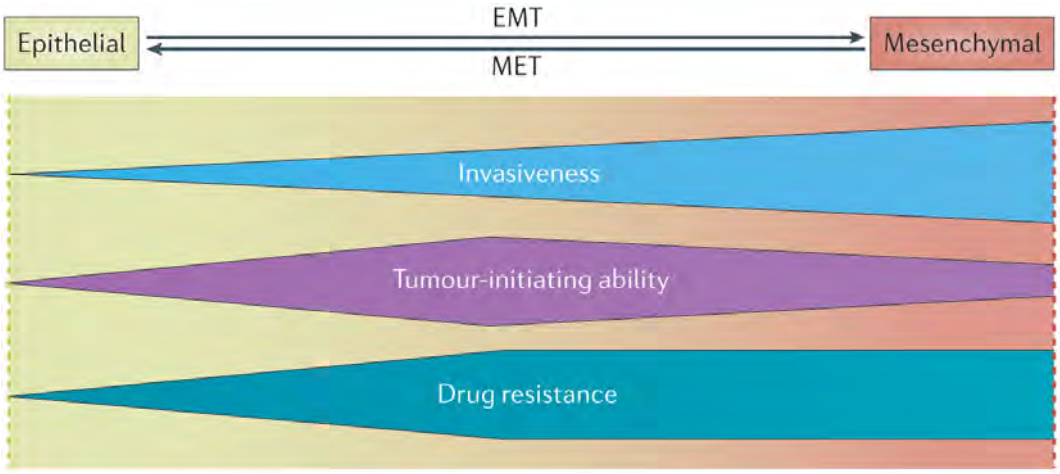


Figure 1.4: The correlation between epithelial mesenchimal transition (EMT) and drug resistance, taken from Shibue and Weinberg (2017). The reverse process of EMT is mesenchymal epithelial transition (MET). Drug resistance and invasiveness have a higher correlation for the EMT process.

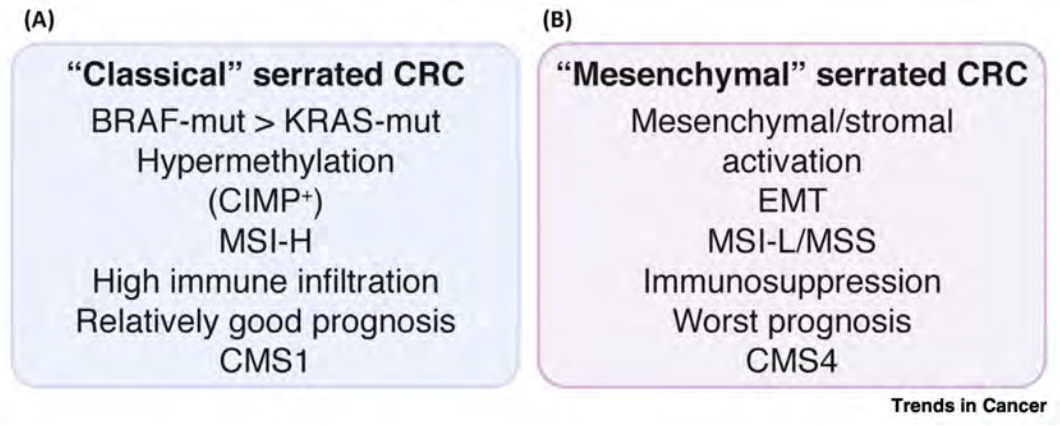


Figure 1.5: Figure displaying the broad overview of the characterisation of CRC subtypes through transcriptomic study. EMT = Epithelial mesenchymal transition. CIMP = CpG island methylation. MSI = Microsatellite instability. MSS = Microsatellite stability. CMS = Consensus molecular subtype. Taken from (Nakanishi *et al.*, 2019).

to an increasing incidence rate of cancer (Bray *et al.*, 2012), and it is well known that early detection of cancer is associated with better prognosis for patients. For example, the survival rates for CRC are as follows, 90% for cancer within the colon and bowel wall (early stage), 70% for cancer that has reached lymph nodes, and only 10% for cancer that has metastasised (late stage) (Siegel *et al.*, 2020*b*; Ganepola *et al.*, 2014). Thus, it is crucial to develop screening tools to detect cancer in the early stages to improve prognosis for patients.

1.3.2.1 Current Screening Tests

As mentioned previously, the current gold standard method for detection of CRC is colonoscopy (Agarwal *et al.*, 2016*a*), a highly invasive procedure many patients are averse to. Colonoscopy involves the use of a colonoscope, which provides a live image from the interior of the colon allowing for the visual detection of precancerous polyps or early stage cancers, which can be followed with a biopsy or surgical removal of the lesions. Colonoscopy has resulted in the prevention of two-thirds of deaths due to CRC (Baxter *et al.*, 2009). A computerised tomography colonography (CTC) is used for patients too frail to undergo colonoscopy, and employs a CT scanner to produce images of within the colon or rectum. Another screening tool is a faecal test, a non-invasive and easy to perform procedure that measures the presence of blood in the stool, however with a high false positive rate faecal tests are still required to be confirmed via a colonoscopy (Ontario and Others, 2009). A procedure known as a double contrast barium enema is another tool that can detect abnormalities in the colon wall using x-rays, which also needs to be validated by colonoscopy (Cittadini, 2012).

Given the importance of colonoscopy, the American College of Gastroenterology has recommended the procedure as a critical screening test for at risk adults, however the participation rates remain between 30% and 55% (Beydoun and Beydoun, 2008). In the United Kingdom, colonoscopy and CTC participation rates were at 20.8% and 28.8% respectively (Zhu *et al.*, 2020). The colonoscopy procedure itself although highly effective, remains costly and uncomfortable for the patient. Furthermore, a study has shown that colonoscopies may miss initial polyps in 6% to 12% of cases, thus the risk of CRC developing after a colonoscopy still exists (Levin and Corley, 2013; Shaukat *et al.*, 2013).

1.3.2.2 The Need for Alternative Screening Methods

The goal of a screening test is to detect cancer early in its progression, and improve patient prognosis, however the colonoscopy participation rates, cost and invasive nature continue to hamper its effectiveness as a screening tool. An alternative less invasive screening tool is thus needed, and there is a growing field within cancer research developing biomarkers as first line screening tools.

A biomarker is defined by the National Cancer Institute as a biological molecule found in bodily fluids, blood or tissue that acts as a sign of a normal or abnormal biological process (http://www.cancer.gov/dictionary?

Use	Example	Reference
Estimate risk of developing cancer	BRCA1 germline mutation (breast and ovarian cancer)	(Easton *et al.*, 1995)
Screening	Prostate specific antigen (prostate cancer)	(Lin *et al.*, 2008)
Differential diagnosis	Immunohistochemistry to determine tissue of origin	N/A
Determine prognosis of disease	21 gene recurrence score (breast cancer)	(Paik *et al.*, 2004)
Predict response to therapy	KRAS mutation and anti-EGFR antibody (colorectal cancer)	(Allegra *et al.*, 2009)
Monitor for disease recurrence	CEA (colorectal cancer)	(Locker *et al.*, 2006)
Monitor for response or progression in metastatic disease	CA15-3 and CEA (breast cancer)	(Harris *et al.*, 2007)

Table 1.1: Table showing the multiple existing biomarker examples and their uses in cancer. Table reproduced from (Henry and Hayes, 2012)

CdrID=45618). A biomarker can thus differentiate a healthy person from a person with a disease, and can be proteins or nucleic acids such as RNAs representing genetic expression that result from molecular alterations occurring in the disease state (Henry and Hayes, 2012). Biomarkers can be used as a screening test for various cases, and examples of biomarker usage in cancer are demonstrated in Table 1.1, with a *KRAS* mutation used as a predictive biomarker in CRC.

Biomarkers can also be used after the diagnosis of cancer in order to determine the prognosis of a patient (Henry and Hayes, 2012). In terms of CRC, Table 1.1 shows the use of carcinoembryonic antigen (CEA) used as a marker for detecting liver metastases in order to monitor disease recurrence (Locker *et al.*, 2006). The measurement of biomarkers can occur in different ways, such as measuring expression from the tissue (invasive) or the blood (non-invasive). A figure illustrating the possible biomarker categories in CRC is displayed in Figure 1.6.

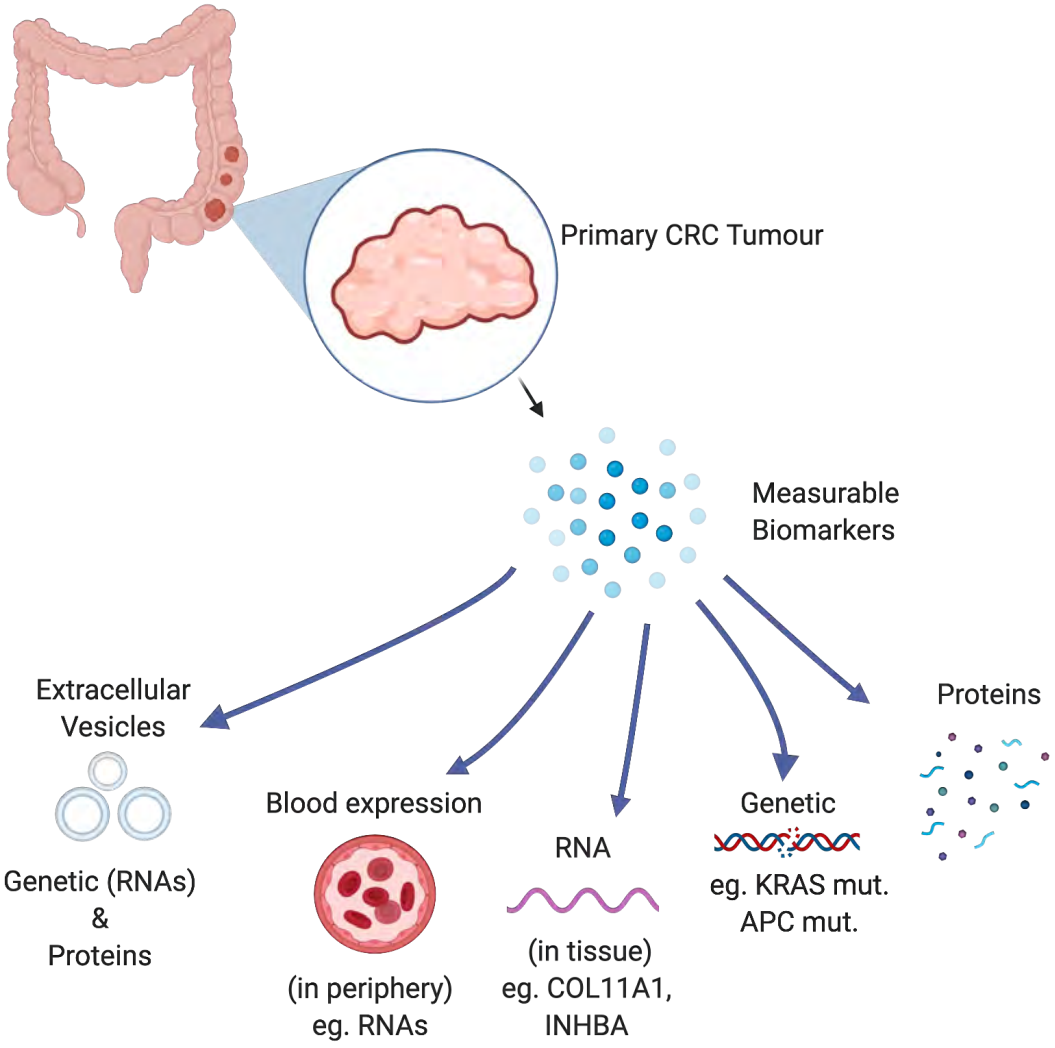


Figure 1.6: Figure showing the categories of biomarkers that can be measured from a primary CRC tumour, including proteins, RNAs and genetic mutations.

1.3.3 Bioinformatics for Biomarker Identification in CRC

Potential biomarkers can be identified through different approaches, however an effective emerging approach involves bioinformatics and the use of NGS and high throughput sequencing (HTS) in an *in silico* approach (Henry and Hayes, 2012). Using an *in silico* approach is a cost-effective way to predict possible biomarkers before testing them clinically or in a wet lab environment where there is a high cost for lab resources.

Gene expression arrays in the form of microarrays have been used previously with *in silico* approaches to identifying biomarkers, however with its rapid development, RNA-Seq has become more popular in use due to having greater efficiency and higher resolution (Xu *et al.*, 2013). Cancers are associ-

ated with differential gene expression and the identification of biomarkers using RNA-Seq data can be accomplished by performing differential expression (DE) analysis on genes between normal cells and cancer cells, with genes found to be commonly DE in the cancer cells investigated as possible biomarkers. RNA itself is normally highly regulated in a healthy person, with aberrant expression becoming more common in pathological states such as cancer (Coskun *et al.*, 2012; Shen *et al.*, 2013), and has been shown to be stable outside of the tumour cell environment , measurable in blood plasma, serum or extracellular vesicles (EVs) as alternatives to measuring in the cancer tissue (Tsui *et al.*, 2002; Sourvinou *et al.*, 2013). RNA-Seq has previously been used to analyse RNA biomarkers in the tissues of breast and prostate cancers (Berger *et al.*, 2010; Sinicropi *et al.*, 2012; Ganepola *et al.*, 2014), and analysis in CRC can not only identify potential biomarkers, but also aid in the research into the molecular basis of CRC initiation and progression through the functional analysis of DE genes directly from the transcriptome (Wang *et al.*, 2012*b*).

Many bioinformatics tools have been created for this type of analysis, leading to the establishment of Bioconductor, an open-source software project for the creation and distribution of bioinformatics tools (Gentleman *et al.*, 2004; Huber *et al.*, 2015). These software tools work in the R programming language, meaning the researcher will often be required to have basic programming knowledge. Using tools provided in Bioconductor, a researcher can perform analyses such as DE analysis, to identify DE genes, and functional annotation of the DE genes through Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) database analysis, allowing the researcher to identify the biological processes and pathways that the genes of interest are involved in. Furthermore, co-expression analysis can be performed to identify genes that are commonly expressed together which can then be further analysed to determine which processes and pathways they contribute to, as well as survival analysis to determine the impact the genes have on patient survival.

Bioinformatics has previously been used to identify potential biomarkers *in silico*, such as the discovery of potential diagnostic biomarkers in ovarian cancer (Kaur *et al.*, 2011). In CRC, microRNAs have been found as biomarkers *in silico*, with the aim for molecular validation to further investigating the findings (Fadaka *et al.*, 2019).

1.3.4 Introduction to Present Study

When using the *in silico* approach in identifying biomarkers, it is imperative that an appropriate dataset be used if an in-house sequence procedure is not performed. Fortunately, with the exponential increase in NGS data, various online databases have been set up hosting sequencing files from existing RNA-Seq studies for research purposes. Using an existing CRC dataset allows for the creation of an RNA-Seq bioinformatics workflow for identifying potential biomarkers in CRC. The present study aims to accomplish this through DE analysis, and further functional and survival analysis on the DE genes, identifying processes and pathways of interest in the development of CRC, and their impact on patient survival. A selection of notable genes will be evaluated as biomarkers by testing their prognostic and predictive values *in silico*. These genes would be found with the intention of using them in a novel screening tool for CRC to improve patient participation, early diagnosis and overall survival rates.

1.4 Aims

This research project aimed to use existing CRC patient RNA-Seq data and to perform downstream bioinformatics analysis for the identification of DE genes as well as identifying potential biomarkers through subsequent functional and *in silico* analysis. Specifically, the study aimed to:

1. Identify DE genes and altered processes and pathways with regards to CRC diagnosis and development for the *in silico* discovery of potential biomarkers.

2. Develop an in-house RNA-Seq workflow to be used on future patient sample data for CRC research in order to identify DE genes.

To fulfil the above-mentioned aims, the following objectives were undertaken.

1.4.1 Objectives

1. Acquire and pre-process a CRC RNA-Seq dataset.

2. Identify DE genes between normal and cancer cells.

3. Use DE genes to identify gene signatures responsible for CRC and metastasis.

4. Perform functional and pathway analysis on the gene signatures.

5. Perform survival analysis on the gene signatures.

6. Validate gene signatures in existing databases.

7. Identify and evaluate potential CRC biomarkers from the gene signatures using an *in silico* approach.

Chapter 2

Materials and Methods

The NGS workflow used in this study is summarised in Figure 2.1.

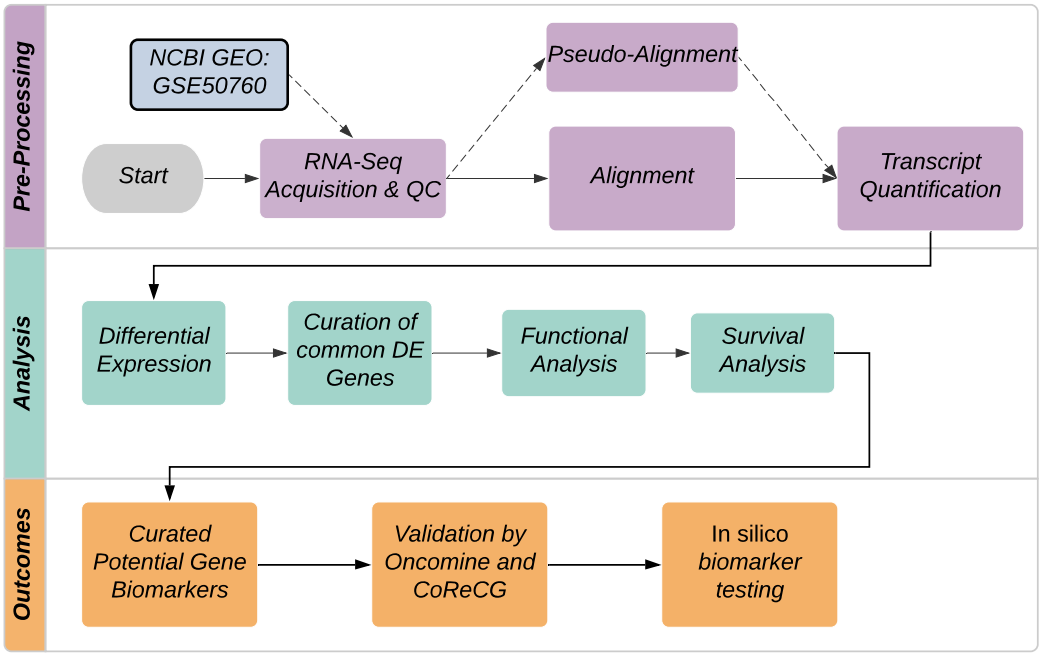


Figure 2.1: A flowchart of the workflow followed in this NGS analysis of the existing CRC RNA-Seq data in GSE50760. Alignment was performed using two traditional alignment tools and a pseudo-alignment tool. Downstream analysis was performed in R and using web tools.

2.1 Acquisition of RNA-Seq Data

The Gene Expression Omnibus (GEO) was established in 2002 and is a large international public repository for NGS and microarray data developed by the National Centre for Biotechnology Information (NCBI) (Barrett *et al.*,

2012). The NCBI GEO web interface is accessible at http://www.ncbi.nlm.nih.gov/geo/. This interface allows for the submission as well as retrieval of NGS data that can either be raw or pre-processed. Studies containing NGS data are listed as series, or GSE, which contain all the samples of the experiment. The GSE holds all relevant information for a user, such as the title and citation of the study, the organism, the technology used for sequencing, as well as study design and contact information. Individual files within a GSE series can be downloaded through the Sequence Read Archive (SRA), which is the primary repository for NGS data. Files on SRA can be accessed through NCBI or through the European Bioinformatics Institute (EBI) at http://www.ncbi.nlm.nih.gov/Traces/sra and http://www.ebi.ac.uk/ena respectively (Leinonen *et al.*, 2010). Files are loaded on the NCBI SRA in the '.sra' format, which requires the NCBI SRA Toolkit for conversion into FASTQ files. One advantage of downloading through EBI is that the raw FASTQ files can be downloaded without needing conversion tools.

In this project, the series GSE50760 was downloaded from the NCBI GEO database (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE50760) and was created in a study by Kim *et al.* (2014*b*). This series consisted of 54 CRC RNA-Seq data samples, of which 18 were primary tumour samples, 18 were normal colon samples and 18 were liver metastases samples. As per the authors, the RNA was isolated using the RNeasy Mini Kit (Qiagen, CA, USA) from the colon tissue. Agarose gel electrophoresis, ethidium bromide staining and ultraviolet light examination was used in the quality and integrity control of the isolated RNA. Following this, a sequencing library was created using TruSeq RNA Sample Preparation kit v2 (Illumina, CA, USA) by purifying mRNA from total RNA using poly-T oligo-attached magnetic beads. The mRNA was then converted into cDNA, and then polymerase chain reaction (PCR) was performed. Sequencing on the library was performed in paired-end reads of 100bp, using Illumina Hiseq-2000 (GPL11154) (Kim *et al.*, 2014*b*). The accession list for the individual files is listed in Table 2.1. Each of the 54 SRA files were downloaded from the NCBI SRA accession links. Each SRA file contained two FASTQ files, as they were paired end sequences.

Sample	Accession Number	SRA Accession
Primary CRC	GSM1228184 - 201	SRR975551 - 568
Normal Colon	GSM1228202 - 219	SRR975569 - 586
Liver Metastases	GSM1228220 - 237	SRR975587 - 604

Table 2.1: Table showing the dataset accession numbers that were used to acquire the SRA files from NCBI GEO. There were 54 runs in total and each sample group had 18 runs.

2.2 Pre-Processing of Data

Pre-processing allows for the data to be assessed for quality as well as for the removal of artefacts that could affect downstream results. RNA-Seq protocols have intrinsic biases and limitations, some of which include nucleotide composition bias and GC bias (Wang *et al.*, 2012*a*). Proper quality control (QC) of RNA-Seq data is able to pick up these biases, and make them known to the researcher. Pre-processing after QC would then involve the removal of poor-quality reads or adapters that were used in the RNA-Seq process, however many raw sequence files that are uploaded to online databases such as GEO have already been pre-processed by the submitting authors.

The table of programmes used for pre-processing are listed in Table 2.2.

Package Name	Version	Purpose	Platform
NCBI SRA-Toolkit	2.10.4	Pre-Processing	Command Line
FastQC	0.11.9	Quality Control	Command Line
MultiQC	1.8	Quality Control	Python
HISAT2	2.2.0	Alignment	Command Line
HTSeq	0.11.4	Transcript Quantification	Command Line
featureCounts	3.11	Transcript Quantification	R
Salmon	1.2.1	Psuedo-alignment w/ Quant	Command Line

Table 2.2: Table showing the programmes that were used in the RNA-Seq analysis workflow for quality control, alignment and quantification.

The SRA files downloaded from NCBI SRA were first converted to FASTQ files using the NCBI SRA Toolkit. FastQC, developed by Barbraham Bioinformatics, is perhaps the most commonly used QC program for HTS data, is able to analyse files in BAM, SAM or FASTQ formats (Andrews *et al.*, 2012). The program is simple to use, and is available as a command line tool for researchers with coding knowledge, as well as having a graphical user interface (GUI) for easier use.

For the present study, MultiQC (Ewels *et al.*, 2016) was used in conjunction with FastQC. MultiQC allows for the individual FastQC reports to be summarised in a single report. This was performed in the Z Shell environment, an extension of the Bash shell environment commonly found on Linux machines, on macOS 10.15. Once quality was confirmed, the FASTQ files were processed further.

2.2.1 Alignment and Quantification

Alignment was performed using HISAT2 (Kim *et al.*, 2015) against the human genome. The complete human reference genome assembly, GRCh38, and the corresponding annotation file in gene transfer format (GTF) was downloaded using the file transfer protocol (FTP) from Ensembl and represents a robust and high quality reference assembly (Schneider *et al.*, 2017). The reference assembly was indexed for alignment using the 'hisat2-build' command which produced eight genome '.ht2' files for alignment.

Alignment of the FASTQ files to the '.ht2' files was performed using the 'hisat2' command to produce sequence alignment map (SAM) files. Unaligned reads were excluded from further processing.

Quantification of the SAM files was done using HTSeq, with the 'htseq-count' command (Anders *et al.*, 2014). This was done against the GRCh38 GTF annotation file, where the reads were annotated to genes and counted. HTSeq gives output as a count matrix for each sample consisting of the genes and their respective counts, representing their expression within the sample. An additional quantification method was used with featureCounts which is included in the R package 'Rsubread' (Liao *et al.*, 2013). The SAM files were quantified against the same GRCh38 GTF annotation file and a count matrix was produced within R.

In addition to the traditional alignment method offered by HISAT2 and featureCounts, a pseudo-alignment was performed using Salmon, a lightweight pseudo-alignment program capable of indexing and quantifying transcripts with high accuracy and at faster speeds (Patro *et al.*, 2017). Salmon does not use the genome, but rather uses the human transcriptome. The FASTA file of the human transcriptome (release 34) was downloaded from Gencode (`www.gencodegenes.com`), run by the EBI (`www.ebi.ac.uk`). The transcriptome was indexed by Salmon using the 'salmon index' command with the '–gencode' flag. The FASTQ files were pseudo-aligned and quantified against the index built by Salmon using the 'salmon quant' command. This created

a 'quant' file for each sample, which consisted of the transcript IDs and their respective counts.

2.3 Differential Expression

2.3.1 Background

DE is a statistical parameter used to measure genes that are either upregulated or downregulated amongst samples by comparing the normalised gene expression levels across the samples. In the present study, DE was performed across the normal colon, primary CRC tumour and liver metastases samples in order to identify genes that had altered expression as the disease progresses. The genes identified as DE were analysed further and their potential as biomarkers for CRC was assessed.

DESeq2 is a popular DE tool designed for RNA-Seq DE analysis and is available in the Bioconductor project in R (Love *et al.*, 2014). DESeq2 takes a count matrix (K) with genes as rows (i) and samples as columns (j). The count value of the genes, or the expression, as recorded from the quantification methods above, is captured in K_{ij}. The reads are modelled according to a negative binomial distribution with the mean (μ_{ij}) and dispersion (α_i) as follows:

$$K_{ij} \sim NB(\mu_{ij}, \alpha_i)$$

The mean (μ_{ij}) is calculated using the quantity (q_{ij}) multiplied by a normalisation factor for all genes (s_j) in the equation:

$$\mu_{ij} = s_j q_{ij}$$

The normalisation or size factor (s_j) is estimated using a median of ratios method (Anders and Huber, 2010). A design matrix (x) with elements j and r is created, which indicates which group the samples come from (i.e. normal or primary tumour or metastases). DESeq2 uses a generalised linear model (GLM) with a logarithmic link with coefficients (β_{ir}) in the following equation:

$$log_2(q_{ij}) = x_{jr}\beta_{ir}$$

The coefficients β_{ir} give the log_2 fold changes for gene i in each column of the model matrix x. The GLM model allows flexibility for more complex designs in genomic analysis (Smyth, 2004; Love *et al.*, 2014). The variance of the observed count and the calculated mean value is dependent on the size

factor s_j for the quantity q_{ij} and is further defined by the dispersion parameter α_i. The variance of the counts is described in the following equation:

$$Var(K_{ij}) = E[(K_{ij} - \mu_{ij})^2] = \mu_{ij} + \alpha_i \mu_{ij}^2$$

Because of its direct implication in the variance of the counts, the dispersion factor needs to be calculated accurately. DESeq2 is able to ensure this accuracy regardless of the sample size by first calculating a gene-wise dispersion factor using a maximum likelihood model then an Empirical Bayes model is used to shrink the factor as needed (Love *et al.*, 2014). This allows DESeq2 to account for gene-specific variation.

Once GLM coefficients are fit for each gene, DESeq2 is able to test whether the coefficient differs significantly from zero, indicating a DE gene. DESeq2 uses a Wald test for testing this significance (Love *et al.*, 2014).

Often times DE is performed on two sample groups, for example a control and a treated group. In the present study, there were three groups: normal colon, primary CRC tumour, and liver metastases samples, which will be referred to as normal, CRC and metastasis in the text respectively. DESeq2 can account for this by providing contrast functionality which allows two groups to be isolated and compared to each other once GLM coefficients have been fit to the genes of the samples of interest. The following contrasts were picked, with the second group in each contrast acting as the baseline for comparison (i.e. CRC compared to Normal (as the baseline expression)):

1. CRC vs Normal
2. Metastasis vs CRC
3. Metastasis vs Normal

2.3.2 Differential Expression in R

The three different count matrices created from Section 2.2.1 were used as input for three different instances of DESeq2 in the R statistical environment using RStudio (R Core Team, 2019; RStudio Team, 2015). The matrix from HTSeq was loaded into R to create a DESeqDataSet using 'DESeqDataSetFromMatrix'. The matrix is structured as shown in Figure 2.2. The 'quant' files created by Salmon were imported using the 'tximeta' package in R as a 'SummarisedExperiment'. From there, the DESeqDataSet was created. Because featureCounts was run within R, the DESeqDataSet was created from

the matrix already produced. The following analyses steps were performed identically for each DESeqDataSet created.

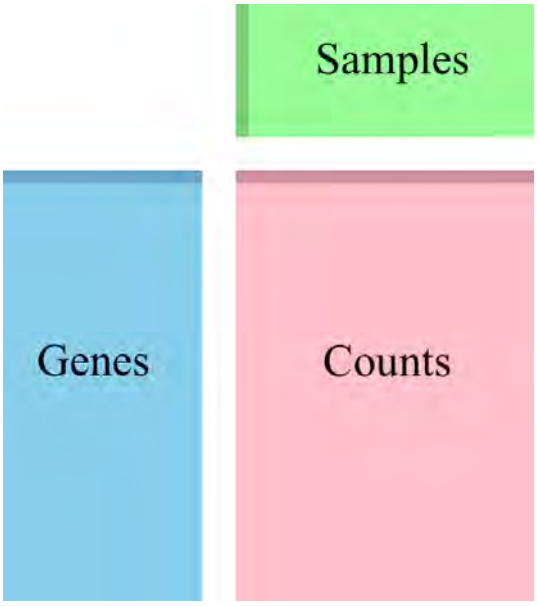


Figure 2.2: Figure showing the matrix layout for DESeq2 input. Genes were listed in the first column, with the samples listed as row headers. The count values populated the table according to gene and sample.

The dataset was subjected to minimal pre-filtering to keep rows that have more than one read total, thus excluding rows (genes) with zero and single counts across samples. Multidimensional statistics, such as principle component analysis (PCA) works best with data that has the same range of variance or is said to be *homoskedastic* (Love *et al.*, 2014). Variance can be stabilised across the mean in DESeq2 using transformations such as the variance stabilising transformation (vst) or the regularised logarithm transformation (rlog), and the latter was employed in the present study. In order to test the consistency between samples within groups, PCA and multidimensional scaling (MDS) plots were created using the rlog stabilised values. PCA uses the genes and counts as data, whereas MDS uses calculated Euclidean distances between the data points (Love *et al.*, 2014).

The DESeq2 DE pipeline was run in R on the previously created DESeqDataSets. This pipeline runs the DE method described above, briefly: the estimation of size factors (s_j), the estimation of dispersion values for each gene (α_i), and fitting the GLM to the data to produce the coefficients (β_{ir}). A results table was created, extracting key values from the DE analysed data namely the log_2 fold change (log_2FC) and p-values of the DE genes. Results

were created for each of the contrasts. The p-values were calculated from a null hypothesis defined as a log_2FC difference of zero between the sample groups in the separate contrasts. In order to account for the multiple comparisons and false positives that may be detected, DESeq2 adjusts the p-value using the Benjamin and Hochberg (BH) method (Benjamini and Hochberg, 1995). Significance for the adjusted values was set at $p < 0.05$ as well as manually setting the log_2FC threshold to 1.5.

A volcano plot was created for each contrast, which visualises the distribution of the coefficients calculated from DESeq2 by using the log_2FC values (Dudoit *et al.*, 2002; Love *et al.*, 2014). Because there might be genes with low counts or high dispersion parameters, the log_2FC values can be shrunk using estimators and does not change the total number of genes that were identified as significantly DE. The default shrinkage estimator provided by DESeq2 was used to shrink the log_2FC values before plotting the volcano plots (Love *et al.*, 2014).

Genes that were identified as significantly DE according to the p-value and log_2FC thresholds were extracted into a separate dataset for each contrast. The top 20 genes according to their p-value were further extracted into separate lists.

The top 20 DE genes for each contrast of each quantification method (HTSeq, Salmon, featureCounts) were compared and the common genes that were present in all three were extracted into separate lists per contrast. This resulted in three smaller gene lists encompassing the most significant DE genes as commonly found by the different alignment and quantification methods. These gene lists, or signatures, will be referred to as Sig1, Sig2 and Sig3 in line with the order of contrasts described above.

2.4 Functional Analysis

In order to get biological insights from the list of significantly DE genes functional analysis was performed using the GO and KEGG databases.

2.4.1 GO Analysis

The GO database is regularly maintained by the Gene Ontology Consortium and includes annotations for a wide variety of organisms and ontologies for biological processes (BP), molecular functions (MF) and cellular components (CC) (Consortium, 2019). Enrichment, or over-representation, is a type of

analysis that determines the probability that functional categories within a certain list of genes are present at a higher proportion than a background set of genes. GO ontologies are classified as terms and have names as well as accession numbers. A single gene is able to be associated with more than a single GO term. These GO terms are loosely hierarchical with differing levels of specificity according to how much is known about the gene product.

GO over-representation analysis can be performed in R using various packages. Here, the package 'clusterProfiler' was used which is available through Bioconductor (Yu *et al.*, 2012). The package uses a GO database package 'GO.db' which has up-to-date annotation maps for the entire ontology (Carlson *et al.*, 2017). clusterProfiler is able to perform enrichment tests using the 'enrichGO' function for the specified category. The datasets containing the significant DE genes from Section 2.3.2 were used as the input data of interest, with all genes that were DE, significant or not, used as the background dataset. Bar and dot plots were created in order to visualise the enrichment results.

In addition to using R for GO analysis, web based tools can also be used. Web-based tools offer ease of use, only requiring a list of significant DE genes. A recently released web tool is WebGestalt (Liao *et al.*, 2019). The DE genes from the lists Sig1, 2 and Sig3 were uploaded as separate inputs, with the human genome set (GRCh38) as the background set. WebGestalt advanced parameters were set as follows: minimum genes for a category was set at 5, maximum genes for a category was set at 2000, significance test adjustment was set to use the BH method, significance level was set to top 10, number of categories from set cover was set at 10, number of categories visualised in report were set at 40, colour in DAG was set to continuous. WebGestalt performed GO analysis across the different categories, and output was an 'html' file containing all the relevant results and plots.

2.4.2 KEGG Analysis

KEGG is a database for the high-level analysis of biological functions from the genomic level. KEGG describes itself as the computer representation of the biological system, with 18 databases ranging from molecular pathways to genes to diseases and is currently maintained by the Kanehisa Laboratory (Kanehisa and Goto, 2000). The KEGG pathway database consists of manually drawn pathway maps representing the molecular interaction of genes and proteins. This allows one to map genes of interest to a pathway. Each pathway map is identified by the combination of 2-4 letter prefix code and a 5 digit number

(Kanehisa and Goto, 2000).

Gene set enrichment analysis (GSEA) was performed using the KEGG database in R using clusterProfiler. GSEA uses the log_2FC from the DE analysis. The changes over specific gene sets are measured instead of the changes in individual genes. These gene sets were derived from KEGG pathways, allowing for the identification of changes in specific pathways. The GSEA analysis output a '.csv' file of the KEGG pathways in order of their adjusted p-value significance which was used to create plots in RStudio.

2.5 WGCNA

Co-expression is a method used to identify genes that are commonly dysregulated together. Co-expression analysis was performed in R using the 'WGCNA' package, which performs weighted correlation network analysis (Langfelder and Horvath, 2008). Co-expression is performed in three general steps (van Dam *et al.*, 2018). First, the correlation between pairs of genes is assessed (their expression similarity). Next, the associations in correlations are used to create a network where genes act as nodes. Finally, groups of similarly co-expressed genes are clustered into modules. These modules can be used for further analysis. Co-expression can further be classified into signed and unsigned, as well as weighted and unweighted. In this study, a signed and weighted network was produced. Signed networks offer more biologically meaningful results, and weighted networks have been shown to produce more robust results (van Dam *et al.*, 2018).

WGCNA analysis is performed on a matrix x_{ij} of interconnecting nodes, or genes. A network is formed by calculating the adjacency matrix α_{ij}, which represents the connectedness between the node pairs based on the expression data generated by the DE analysis. This is done by first calculating the similarity between nodes, known as the co-expression similarity (s_{ij}), by taking the correlation coefficient of the nodes i and j. In order to avoid information loss, a weighted network approach is used by using a soft threshold instead of a hard threshold, which allows the network adjacency to assume a value continuous between 0 and 1. The adjacency value is calculated by taking the similarity value to the power of the soft threshold β ($\beta >= 1$), described below.

$$\alpha_{ij} = s_{ij}^{\beta}$$

In order to determine the threshold value β, a plot is created using $log(p(k))$ against $log(k)$, where k is the connectivity and $p(k)$ is the frequency distribu-

tion. A square regression correlation, or R^2 is determined for this plot and is used to measure the correlation of genes with high connectivity to genes with low connectivity (Zhang and Horvath, 2005). The R^2 values were plotted against different soft thresholding powers (β). A β value is then picked where the R^2 value is at least above 0.8 and seems to be levelling off (Zhang and Horvath, 2005). In this instance, a β value of 4 was chosen, representing the lowest value where R^2 stabilises above 0.8. The adjacency matrix created using the determined β value is then transformed into a Topological Overlap Matrix (TOM). The TOM is created to reduce the effects of noise and possible false associations. This is done using the following formula

$$w_{ij} = \frac{l_{ij} + \alpha_{ij}}{min(k_i, k_j) + 1 - \alpha_{ij}}$$

w_{ij} represents the TOM similarity value, calculated using l_{ij}, the number of nodes connected to i and j nodes, k_i and k_j, the interconnectivity for the i and j nodes, and the adjacency value being between or equal to 0 and 1. A TOM based dissimilarity value allows for more distinct gene modules (Zhang and Horvath, 2005). Thus, in order to create a dissimilarity value (d_{ij}), the original value is subtracted from 1.

$$d_{ij} = 1 - w_{ij}$$

The dissimilarity value is used to create a dendrogram of the genes. The vertical lines (or leafs) in the dendrogram represent genes, and the branches group together for highly interconnected and co-expressed genes. Clustering of these genes into modules was performed using a combination of the partitioning around medoids (PAM) method, and a general hierarchical clustering method. The PAM method partitions the dissimilarity matrix into equal size k-mers which are clustered according to the minimal distance between gene correlations whereas the hierarchical clustering method uses the average distance between a node in relation to all other nodes. Modules with similar expression values with higher than 0.75 correlation with each other were merged. The modules were clustered with the following parameters of having a minimum of 30 genes and a maximum of 500 genes, and can be ranked according to significance. Genes and gene modules were assessed in terms of their significance by their correlation to the trait of the samples, being normal, CRC or metastases. The gene significance for a gene at node i (GS_i) relates to how biologically relevant the gene is. Gene significance can be plotted against module membership (MM) to identify genes of potential importance.

$$K_{cor,i} = cor(x_i, E_q)$$

In the above equation, $K_{cor,i}$ represents the MM, of a gene or node i in the module q and is defined by the correlation between the expression profile of node i and the respective module eigengene (E_q). Furthermore, hub genes that connect to many modules can be identified as genes with the highest associated module eigengene based connectivity (K_{me}).

2.6 Validation

In order to validate the identified genes in the 'Sig' lists they were compared to previous CRC literature and published studies. This was completed using two online tools, the Colon Rectal Cancer Gene Database (CoReCG), and Oncomine.

2.6.1 Validation using CoReCG

CoReCG is a curated database of validated CRC genes from previously published literature (Agarwal *et al.*, 2016*b*). A web-tool, accessible at `http://lms.snu.edu.in/corecg/` allows the user to input a gene ID and search the database. If the gene is found in the database, a summary of the findings and references to the previous literature is provided. Additionally, links to external resources such as NCBI, Uniprot and Ensembl for the gene of interest are provided. The CoReCG database has 2056 genes referenced from 2486 papers from 1980 to 2015, with the majority of papers being published in 2015 (Agarwal *et al.*, 2016*b*). The genes from the 'Sig' lists were uploaded individually to query the database. Their presence in the database and respective summaries of their role in CRC was recorded.

2.6.2 Validation using Oncomine

Oncomine is a bioinformatics initiative aimed at analysing and collecting data for over 18 cancer types and is accessible at `https://www.oncomine.org` (Rhodes *et al.*, 2004, 2007). Oncomine was created in response to the exponential increase in microarray data and has since incorporated HTS data such as RNA-Seq data for many cancer types, and uses the expression data to perform a specialised DE workflow. This allows the data to be analysed in combinations such as 'normal vs cancer' of the same tissue type and 'cancer vs cancer' of different types or subtypes (Rhodes *et al.*, 2007). There are currently 9 large studies for colon and rectal cancer with 468 samples in total. The genes in

these samples are ranked according to their DE significance in the analysis combinations. Genes in the 'Sig' gene lists were uploaded to Oncomine to determine how significant they are in terms of disease progression as found in other studies.

On the Oncomine web interface, the following filters were added for the 'Sig1' list of genes: Cancer Type = 'Colorectal Cancer', Analysis Type = 'Normal vs Cancer', Dataset Size = '150+'. Because the 'Sig2' and 'Sig3' list of genes were DE genes in liver metastases samples, many were not found in Oncomine CRC studies. The filters were adjusted as follows: Cancer Type = 'Liver Cancer', Analysis Type = 'Cancer vs Cancer'. Thresholds for results were set to $p<0.05$ and fold change of 1.5. Each gene in the lists was uploaded individually and added to the filters. If the gene matched to a study, the DE results and study name were recorded into a separate table.

2.7 Survival Analysis

Survival analysis is a statistical technique used in data analysis to determine the time until an event occurs (Kleinbaum and Klein, 2010). In cancer disease instances, the event can be regarded as death, relapse, metastasis or overall survival. Survival data is recorded in studies found in databases online, such as the Cancer Genome Atlas (TCGA), whereby tables consisting of the individuals and their time to an event is recorded. Survival data is often plotted using Kaplan-Meier (KM) curves, which allow researchers to visualise the probability an individual will survive an event of interest over a period of time. One benefit to using a KM plot is that censored data, such as data before an event is reached, is taken into account, allowing for empirical distribution of data (Kaplan and Meier, 1958). In a bioinformatics example, the input can be a gene or gene signature, with the reference survival information taken from TCGA. The resulting KM plot will show the survival probability of individuals with cancer who express the gene or genes of interest. This plot can be used to show the change in survival probability of survival in patients according to changes in gene expression, i.e. when there is high or low expression of the genes. When creating KM plots, a log-rank test can be performed in order to determine whether the plots from high and low expression are statistically significant or not (Kleinbaum and Klein, 2010). The log-rank test is a large chi-square test that compares two KM curves with the null hypothesis being that the curves show no statistical difference.

In the present study, the web tools PROGgeneV2 and GEPIA were used (Goswami and Nakshatri, 2014; Tang *et al.*, 2019). PROGgeneV2 is a tool created to study the prognostic data of various cancers and can be used to create KM plots for individual genes. The tool is a web interface for a backend R script using the 'survival' package (Therneau, 2020). The genes in each list were uploaded individually to the PROGgene web interface. The 'colorectal' cancer type was selected, with 'death' as the survival mode and gene expression bifurcated at the 'median' meaning that the gene expression was split into high and low according to the median expression of that gene in the patient data. PROGgene has survival information from 15 large CRC studies and the present analysis the TCGA-COAD (colon adenocarcinoma) study (Peng *et al.*, 2015) was used. The analysis produced separate KM plots for each gene in the list. In order to conduct batch analysis of the gene signatures, the Gene Expression Profiling Interactive Analysis (GEPIA) tool was used (Tang *et al.*, 2017, 2019). GEPIA2 has data sources from TCGA and Genotype Tissue Expression (GTEx) databases. Each 'Sig' gene list was input as separate gene signatures. The 'overall survival' measure was selected, with the 'median' cutoff selected for high and low gene expression. The other parameters were left as default. The datasets used were the TCGA COAD, READ (rectal adenocarcinoma) and LIHC (liver hepatocellular carcinoma) datasets. Resulting KM plots for the gene signatures were generated. Additionally using GEPIA, stage expression plots were created for each individual gene.

2.8 Drug Gene Interaction

Drug gene interactions allow researchers to determine the therapeutic significance of genes. The Drug Gene Interaction database (DGIdb) is an online web-tool hosting all known drug gene interactions for FDA approved and unapproved drugs (Cotto *et al.*, 2018). The genes from the 'Sig' lists were uploaded to DGIdb with the FDA approved and antineoplastic drug categories filtered. The genes that matched to drug interactions were recorded.

2.9 Biomarker Testing

In order to assess the prognostic predictive potential of the selected genes, their expression in the Oncomine database was assessed in terms of sensitivity, specificity and precision (Parikh *et al.*, 2008). The assessments include true

positives (TP) of patients that have CRC and express the gene, true negatives (TN) of patients that do not have CRC and do not express the gene, false positives (FP) of patients who do not have CRC but do express the gene, and false negatives (FN) of patients who do have CRC but do not express the gene. The sensitivity is the ability of the gene to correctly detect CRC presence and is defined by:

$$SE = \frac{TP}{TP + FN}$$

Specificity is the ability to correctly identify patients not having CRC presence and is defined by:

$$SP = \frac{TN}{TN + FP}$$

The positive predictive value (PPV) or precision value shows the percentage of positives that have CRC presence and is defined by:

$$PPV = \frac{TP}{TP + FP}$$

The genes were uploaded individually to Oncomine and the expression data from the CRC studies was used to calculate the results using the above equations.

An additional parameter was measured by assessing the blood expression levels of the gene. This was done using BBCancer, a recently made available blood based biomarker atlas for the early diagnosis of different cancers (Zuo *et al.*, 2020). BBCancer represents the largest blood sample resource for cancer biomarker testing. BBCancer contains the DE analysis of different RNAs between normal and tumour samples for different cancers including CRC as taken from previous studies up to July 2018. These studies were used in a meta analysis of the fold changes. These fold changes were then integrated with a robust rank algorithm to create a meta score for each gene. The meta scores are of interest as they show a single value indicating whether the gene has high or low expression in the blood of a cancer or normal patient. BBCancer also provides a measurement in addition to the DE analysis called expression abundance. Together, the DE analysis and the expression abundance analysis show whether a genes expression can be distinguished from normal blood and detectable as a biomarker respectively.

Each gene of the 'Sig1' list was entered into BBCancer and analysed using DE as well as expression abundance. The resulting meta scores were recorded in a table format. The expression levels of the related proteins were matched against the patient blood samples, as well as extracellular vesicle (EV) samples.

2.10 Galaxy

Galaxy was developed in 2010 to address the increasing computational requirements to analyse HTS data (Blankenberg *et al.*, 2010). Galaxy hosts a framework that is able to encapsulate high-end bioinformatics tools and reduces the need for researchers to have coding knowledge and powerful computer systems as all that is required is a modern web browser. Over the past 10 years Galaxy has grown in popularity and usage, and provides a simple yet comprehensive way to perform downstream analysis on NGS data. Galaxy has dedicated tutorials on RNA-Seq data analysis (`https://galaxyproject.org/tutorials/rb_rnaseq/`), and allows researchers to upload their data and either use pre-existing workflows developed by other researchers, or create their own workflow. Galaxy calls on other programs to perform the steps of analyses without the researcher needing to know how to operate the programs. Galaxy is free to use, only limited by the amount of data one can store (capped at 250GB on the public server).

Galaxy was used in the present study to evaluate an RNA-Seq analysis workflow that is completely separate to the methods described above. A workflow for alignment, quantification and DE analysis was created on Galaxy as follows.

A Galaxy account was set up in the free to use Galaxy Main server. As the FASTQ files take up more than the 250GB storage quota, an additional 250GB was acquired from the Galaxy support staff (Jennifer Hillman-Jackson, USA). Using Galaxy, the FASTQ files were downloaded directly to the Galaxy server, without having to download them to a personal computer or device first. This was done by using the 'Download and extract data' tool and importing the individual SRR accession numbers. Galaxy downloaded the SRA files, and using the NCBI-Toolkit extracted the FASTQ files for each accession. Another option that was assessed was using the 'Fetch Data' option by pasting the direct links to the gzipped FASTQ files from EBI.

The FASTQ files were grouped in their pairs. Galaxy hosts many tools for alignment such as HISAT2, Salmon, Kalisto and RNA-STAR. STAR was unavailable at the time due to maintenance. Here, the HISAT2 tool was used for alignment, using the built in human reference genome (hg38) and the FASTQ files as input. This created a list of BAM files. The BAM files were used as input with the featureCounts tool and quantification was performed using the built in human genome annotation file (hg38).This created a list of tabular count matrices on the Galaxy server. These count matrices were then used as

input for DE using both DESeq2 and edgeR. This led to an output of basic plots showing initial DE results.

2.11 Programs and Code Information

Detailed 'sessionInfo' from the R session, including packages and their version numbers, can be found in the Appendix. All codes used are available at GitHub (https://github.com).

Chapter 3

Results

3.1 Quality Control

QC was performed using FastQC and merged into a batched summary file using MultiQC as described above. Figure 3.1 represents the mean quality score displaying the quality of each read in each FASTQ file, whereas Figure 3.2 shows the quality for each FASTQ sequence file, and the the per sequence GC content showing the percentage of guanine and cytosine nucleotide bases in the FASTQ sequence files is shown in Figure 3.3.



Figure 3.1: Figure showing the mean quality score for all the reads for each base. This quality score is included in the FASTQ file after sequencing, and represents the predicted accuracy of a correct base call.

Figure 3.1 shows the distribution of quality scores across the reads. This is an important analysis in QC, as it can show and alert the researcher if there were any problems in the sequencing. The y-axis shows the Phred quality score, a measure of the quality of the identified base. The Phred score is

calculated during sequencing, and indicates the probability of the sequencer misidentifying the base. The Phred score is calculated as Q in the following equation:

$$Q = -10 * log_{(10)}(P)$$

Where P is the estimated error probability of the base identification or call (Ewing and Green, 1998). In FastQC, a Phred (Q) score of 30 is highlighted in green and is equivalent to a 99.9% base call accuracy (or a 1 in 1000 probability of incorrect base call). The x-axis represents the base pair position in the read. Figure 3.1 shows the reads in this study are of high quality and start to drop in quality as the read position increases which is commonly associated with Illumina sequencing phenomena: signal decay and phasing (Ledergerber and Dessimoz, 2011). Signal decay occurs as sequencing progresses each cycle, leading to the degradation of fluorophores on the 3' end of the read. Thus, the signal degrades each cycle leading to a drop in the quality score. Phasing occurs during a cycle when a strand fails to incorporate a base call, thus in the next cycle it will lag behind. This is also known as a loss in the synchronicity of the sequencing (Ledergerber and Dessimoz, 2011). The reads in this study showed no signs of instrument damage or signs that warrant concern, and therefore the quality of the reads was deemed good.



Figure 3.2: Figure showing the average quality score per sequence (x-axis) according to the Phred score and the respective sequence counts (y-axis). This quality score is included in each FASTQ file and represents the number of sequences that have a high accuracy of having correct base calls. Red indicates a poor quality score of a Phred score less than 20, yellow indicates a Phred score that should be investigated and green indicates a good Phred score. The majority of sequences fall in the green block.

Figure 3.2 shows the per sequence quality score for the FASTQ sequences.

The Phred scores are plotted on the x-axis with the number of sequences plotted on the y-axis. The majority of sequences had a Phred score of above 30, demonstrating that the quality of the reads was good with no signs that warrant concern.

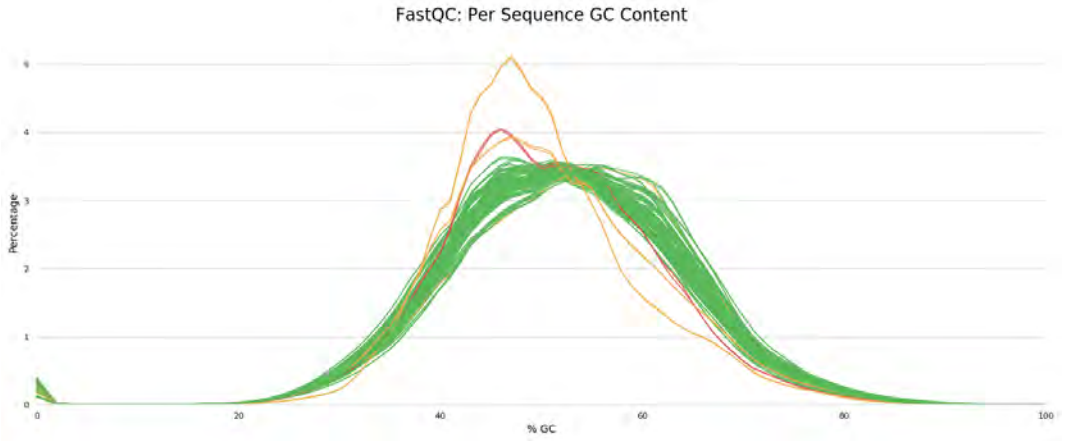


Figure 3.3: Figure showing the overall percentage GC content in the FASTQ files. The GC content can then be compared to the organisms expected GC percentage. *Homo sapiens* expected GC percentage is 46.1 ± 9%. Lines in red and yellow represent outlier samples.

Figure 3.3 shows the per sequence GC content. Of the 108 FASTQ files, 7 of them were outliers. The %GC peaks at around 50, showing a normal distribution. A biased distribution would indicate a contaminated sequencing library. FastQC calculates the distribution using the observed data, as it does not know the expected distribution. In homo sapiens, the expected %GC is $46.1 \pm 9\%$ (Romiguier *et al.*, 2010), meaning the sequences in this study were deemed good for further analysis.

3.2 Differential Expression

DE using DESeq2 in RStudio was performed on the different count matrices as aligned by the different methods described above. DE analysis identified genes that were either upregulated or downregulated amongst the different sample relative to a baseline group, i.e. "CRC vs Normal" denotes the DE analysis on genes in the "CRC" group as compared to the baseline "Normal" group, thus highlighting DE genes present in CRC. The DE dataset was pre-filtered of genes that had total counts across all samples of <1 and the result of this filtering is displayed in Table 3.1.

Method	Pre-Filtering	Post-Filtering
HTSeq	60676	45798
Salmon	60240	43636
featureCounts	60676	44818

Table 3.1: Table showing the gene counts for each quantification method before and after manual filtering of genes with <1 count across all samples.

Due to the following DE analyses being done separately for the three different methods as described above, and in the interest of legibility, only the Salmon DE results are displayed in the following text. Salmon has been previously shown to be as effective if not more than traditional alignment methods (Patro *et al.*, 2017). The DE plots using the other two alignment methods showed similar results, and can be obtained upon request.

The initial results from the DE dataset were transformed and normalised using rlog as described above, and were plotted in two plots, a PCA and an MDS plot. These plots were chosen in order to visualise the samples and groups in terms of their relation to each other. The PCA plot in Figure 3.4 shows distinct groupings between the three sample groups, Normal (blue dots), CRC (red dots) and Metastasis (green dots). The Normal samples show close clustering with one another, indicating consistency in the gene expression in the sample group. The CRC samples appear to cluster closer to the Normal samples, with intra-group variation in expression, with Metastasis having the most distinct separation and variance amongst individuals, showing and possibly predicting how the samples and gene expression will differ in further analysis. It also shows a pattern with regards to cancer progression, with the metastatic cancer having the highest variation from the primary tumour possibly as genetic differences continue to arise (Huang *et al.*, 2018*a*). Interestingly, some dots cluster with the other sample groups, for example there are a few red CRC dots clustering with the blue Normal dots, and a few green Metastasis dots clustering with both the Normal and CRC dots. The exact sample (SRR) names are shown in Appendix Figure A.1. This is hypothesised to be due to patient sample variability, as well as the highly heterogeneous nature of CRC and metastasis (Grady and Carethers, 2008; Vogelstein *et al.*, 2013). It could also represent samples at different stages in the disease, with the green dots clustering with red indicating samples in the early stages of metastasis that still show similar gene expression profile to primary CRC, and the red dots clustering with the blue dots indicating early stages of primary CRC tumour

progression that still shows gene expression profile similar to healthy colon cells.

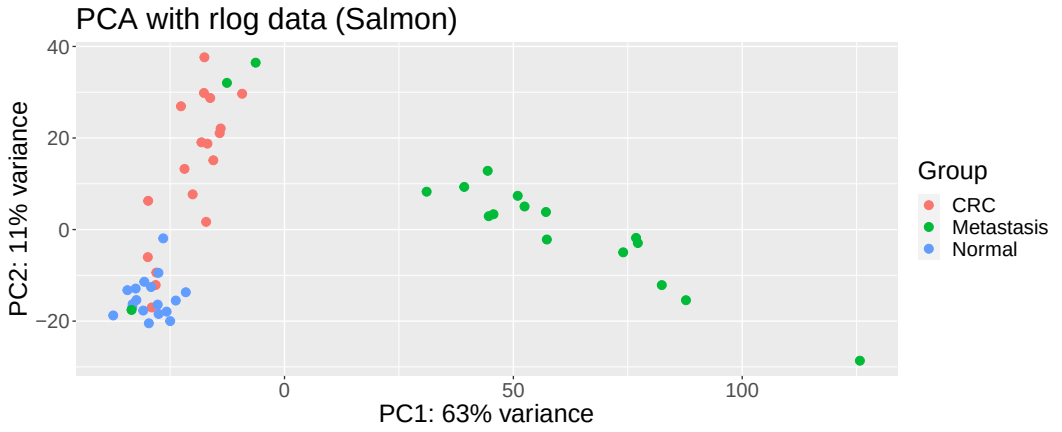


Figure 3.4: A PCA plot using rlog normalised data from the Salmon method along PC1 and PC2. The blue dots indicate the Normal samples, the red dots indicate the primary CRC tumour samples, with the green dots representing the liver metastases samples. The percentage variance indicates the variance of the samples across principal components 1 and 2, the most variable of the components, which then represent variability of the samples from each other.

The MDS plot shows the consistency of the gene expression amongst individual samples within their groups by measuring the Euclidean distances between samples. Samples with similar gene expression values cluster together. In Figure 3.5, similar to the PCA results, there is distinct separation of the sample groups, with clustering amongst the Normal samples. In this instance, the CRC and Metastasis sample groups show separation from each other, but a high degree of variability amongst individual samples as in the PCA plot. This is not uncommon from patient RNA-Seq data as there are many variables in deriving these sequences.

DE analysis was performed as described previously, identifying genes in each contrast that were either upregulated or downregulated. In order to visualise these results, volcano plots (Figure 3.6) were created for each contrast. An volcano plot is useful for showing the distribution of genes across the contrasts of interest (Dudoit *et al.*, 2002). The volcano plot illustrates the log_2FC of genes across the groups. Each dot in the volcano plot represents a gene that is upregulated or downregulated, whereas the red dots indicate genes that are significantly ($p<0.05$ and $log_2FC > 1.5$) DE. The plots show significantly upregulated genes in Figure 3.6a represented by the red dots beyond the log_2FC

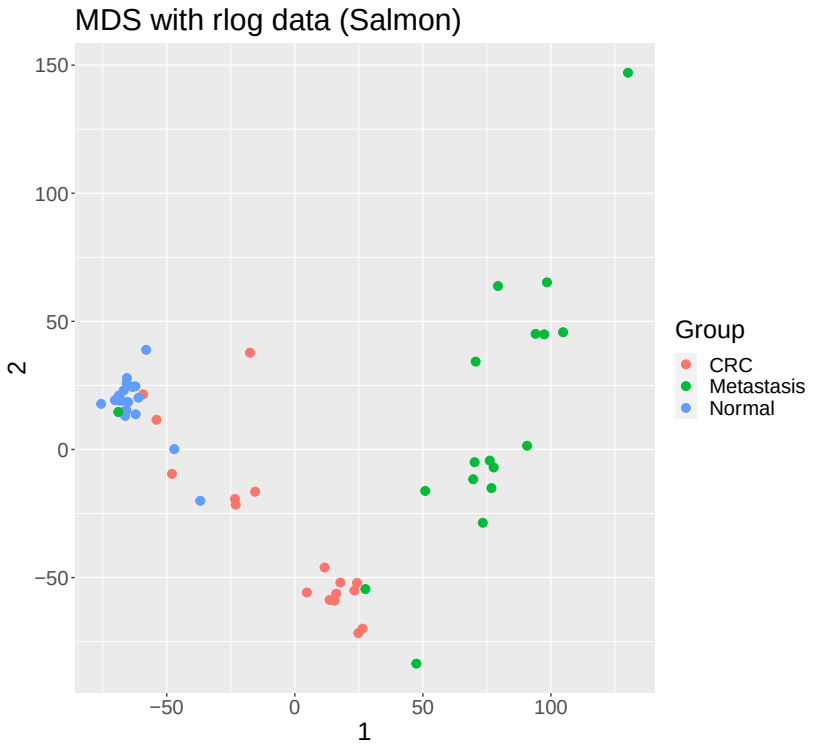


Figure 3.5: An MDS plot using rlog normalised data from the Salmon method along two dimensions of Euclidean distance differences, 1 (x-axis) and 2 (y-axis). The Euclidean distances were calculated from the rlog data, and show the relationship between samples and how they differ. The blue dots indicate the Normal samples, the red dots indicate the primary CRC tumour samples, with the green dots indicating the liver metastases samples.

midline of 0, therefore illustrating how genes in the CRC group are upregulated as compared to the Normal group. In Figures 3.6b & 3.6c the increase in amount of red dots indicate an increase in significantly upregulated as well as downregulated genes in the Metastasis group as compared to the CRC and Normal groups respectively. Notably, the Metastasis upregulated genes in Figures 3.6b & 3.6c were also above a log_2FC of 5.

In order to quantify the volcano plots, the number of significantly DE genes ($p < 0.05$ and $log_2FC > 1.5$, i.e the red dots in each contrast) were extracted and are displayed in Table 3.2 according to method and contrast. The 'CRC vs Normal' group had the least number of DE genes, with 'Metastasis vs CRC' having slightly more and 'Metastasis vs Normal' having the most.

Method	CRC vs Normal	Metastasis vs CRC	Metastasis vs Normal
HTSeq	468	862	3344
Salmon	481	866	3194
featureCounts	543	930	3549

Table 3.2: Table showing the significant genes that were differentially expressed in the different contrasts and methods. These genes met two criteria in order to be labelled as significant, $p < 0.05$ and $log_2FC > 1.5$.

The top 20 DE genes according to p-value for each contrast were selected from the extracted significant genes, and their respective counts for each contrast were plotted in Figure 3.7 to illustrate the gene spread across groups.

These 20 genes were different amongst the three different alignment methods used. Taking this into account the genes that were common in all three quantification methods were extracted and made into a separate list for each contrast using Microsoft Excel (2020). Tables 3.3, 3.4 & 3.5 show the common genes as well as brief descriptions for "CRC vs Normal"(Sig1) and "Metastasis vs CRC" (Sig2) and "Metastasis vs Normal" (Sig3) respectively. The descriptions for the genes were taken from GeneCards (https://www.genecards.org) which hosts user-friendly summaries for all annotated human genes. These genes formed the foundation for further analysis, as they represented the most significant DE genes found using the three different methods and it is from these genes that a potential biomarker could be found.

CRC vs Normal	Details	log_2FC	p-value
COL11A1	Encodes the alpha unit for type XI collagen	4.755618	5.105867e-20
ETV4	Functions in the regulation of cell growth, angiogenesis, migration proliferation, and differentiation	3.993919	2.956365e-20
INHBA	Encodes a member of the TGF-β family	4.450682	1.088434e-19
ADAM12	Encodes disintegrin and metalloprotease, involved with cell-cell interactions	4.101672	4.000987e-20
CLDN1	Encodes claudin-1, an integral membrane protein	3.708042	7.314958e-16
COL10A1	Encodes the alpha unit for type X collagen	5.972869	7.246757e-22
MMP1	Encodes matrix metalloproteinase-1, a collagenase	4.506133	2.094360e-15
FAP	Encodes fibroblast activation protein, a transmembrane glycoprotein	3.653027	4.044356e-15
CTHRC1	Encodes collagen triple helix repeating protein, involved in vascular remodelling	3.650354	2.018502e-15
CASC19	Cancer associated susceptibility gene 19	3.117882	2.524442e-11
MMP3	Encodes matrix metalloproteinase-3, a collagenase	4.370396	9.031456e-13
KRT17	Encodes type I intermediate filament chain keratin 17, involved in follicle development	4.276951	1.799842e-11
FOXQ1	Encodes forkhead box Q1, a protein involved in cell cycle regulation	3.955872	9.914317e-12

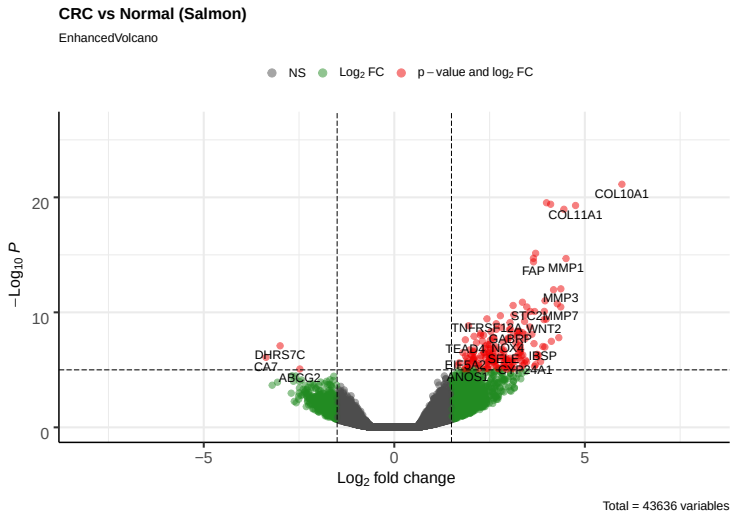
Table 3.3: Table showing the common top DE genes amongst the three methods for CRC vs Normal and their relative descriptions, log_2FC and p-value. This list represents signature Sig1.

Metastasis vs CRC	Details	log_2FC	p-value
CYP2E1	Cytochrome liver membrane protein and key activator of cytochrome P450 enzyme activity	8.222054	8.330575e-71
HPX	Encodes hemopoxin, a heme transport protein	8.053424	9.275393e-59
APOH	Encodes apolipoprotein H, a functional plasma protein	8.426904	3.693762e-65
APOA1	Encodes apolipoprotein A1, involved in lipid metabolism	8.442701	1.958622e-61
APOB	Encodes apolipoprotein B, involved in lipid and cholesterol metabolism	8.174899	7.146724e-60
FGL1	Encodes fibrinogen like protein, involved in T-cell activation	8.441066	2.258128e-65
FGB	Encodes beta subunit of fibrinogen	8.257455	2.555077e-55
ITIH2	Encodes inter alpha typist inhibitor heavy chain 2, involved in ECM stabilisation	8.407494	5.303918e-62
HP	Encodes haptoglobin, binds free plasma haemoglobin	8.129479	2.112212e-54
FGA	Encodes alpha subunit of fibrinogen	8.345762	1.959888e-58
ITIH1	Encodes inter alpha typist inhibitor heavy chain 1, involved in cell adhesion	8.407494	5.303918e-62
AMBP	Encodes alpha-1-microglobulin precursor a complex glycoprotein in plasma	8.081744	1.826885e-57
ORM1	Encodes ososomccoid 1, an acute inflammation plasma protein	8.343218	1.024156e-57

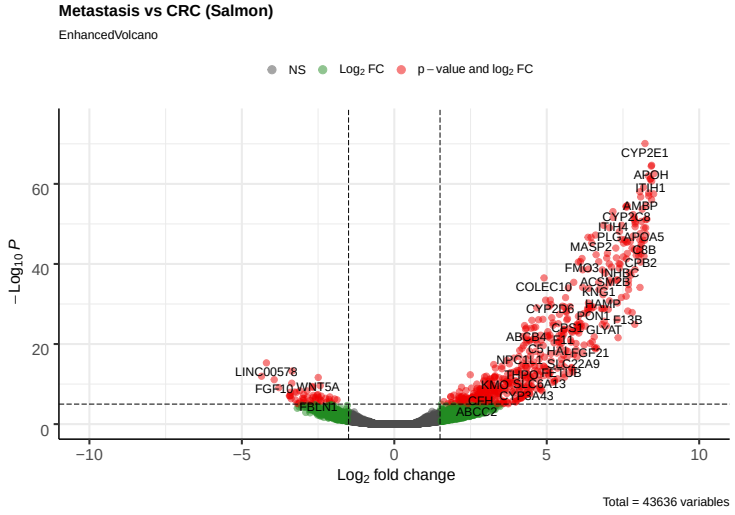
Table 3.4: Table showing the commonly DE genes amongst the three methods for Metastasis vs CRC and their relative descriptions, log_2FC and p-value. This list represents signature Sig2.

Metastasis vs Normal	Details	log_2FC	p-value
CYP2E1	Cytochrome liver membrane protein involved in metastasis	7.929285	1.220889e-65
APOH	Encodes apolipoprotein H, a functional plasma protein	8.041904	3.964893e-59
APOA1	Encodes apolipoprotein A1, involved in lipid metabolism	8.331261	9.578143e-60
HPX	Encodes hemopoxin, a heme transport protein	7.808650	4.194697e-55
FGL1	Encodes fibrinogen like protein, involved in T-cell activation	8.156394	6.833551e-61
ITIH1	Encodes inter alpha typist inhibitor heavy chain 1, involved in cell adhesion	8.222296	4.667700e-59
IGFBP1	Encodes insulin like growth factor binding protein 1, involved in cell migration and metabolism	8.098457	5.798662e-62
ITIH2	Encodes inter alpha typist inhibitor heavy chain 2, involved in ECM stabilisation	8.230591	5.899587e-61
AMBP	Encodes alpha-1-microglobulin precursor a complex glycoprotein in plasma	8.277759	2.728997e-60
ITIH3	Encodes inter alpha typist inhibitor heavy chain 3, involved in ECM stabilisation	7.530781	1.401363e-59
APOA2	Encodes apoliprotein A2, involved in lipid metabolism	8.226643	8.332104e-57
PLG	Encodes plasminogen	7.903639	8.616285e-63

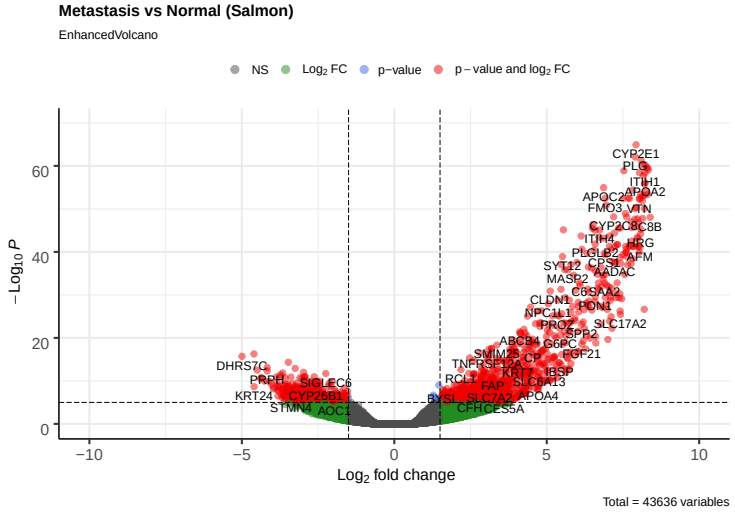
Table 3.5: Table showing the commonly DE genes amongst the three methods for Metastasis vs Normal and their relative descriptions, log_2FC and p-value. This list represents signature Sig3.



(a) CRC vs Normal

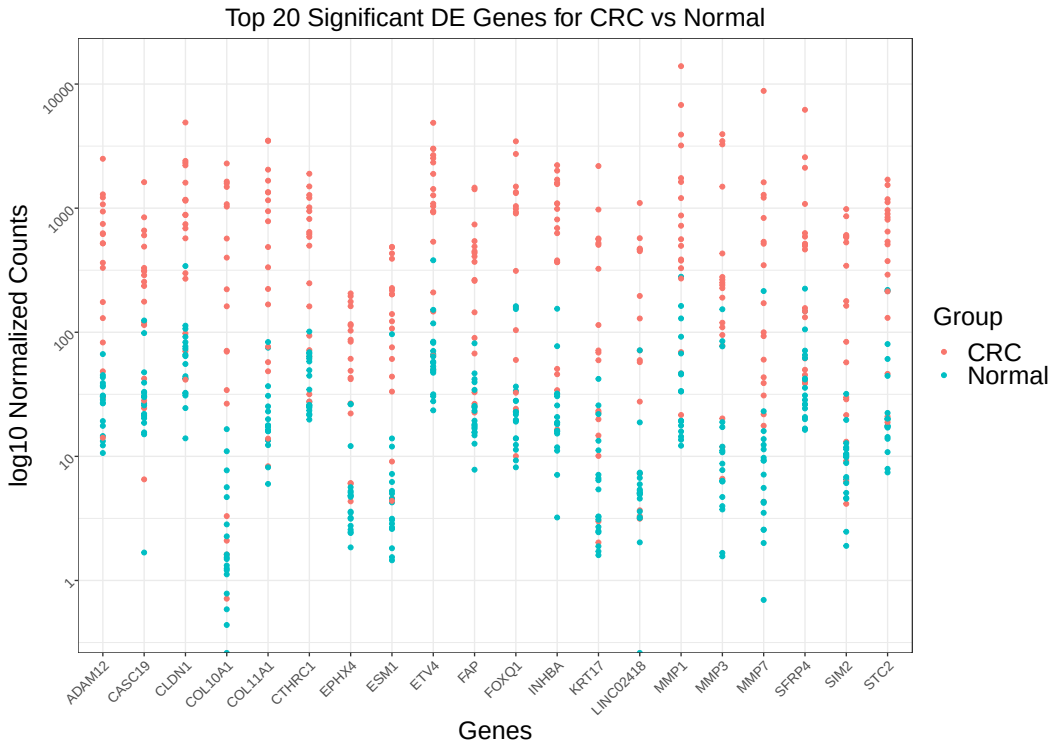


(b) Metastasis vs CRC

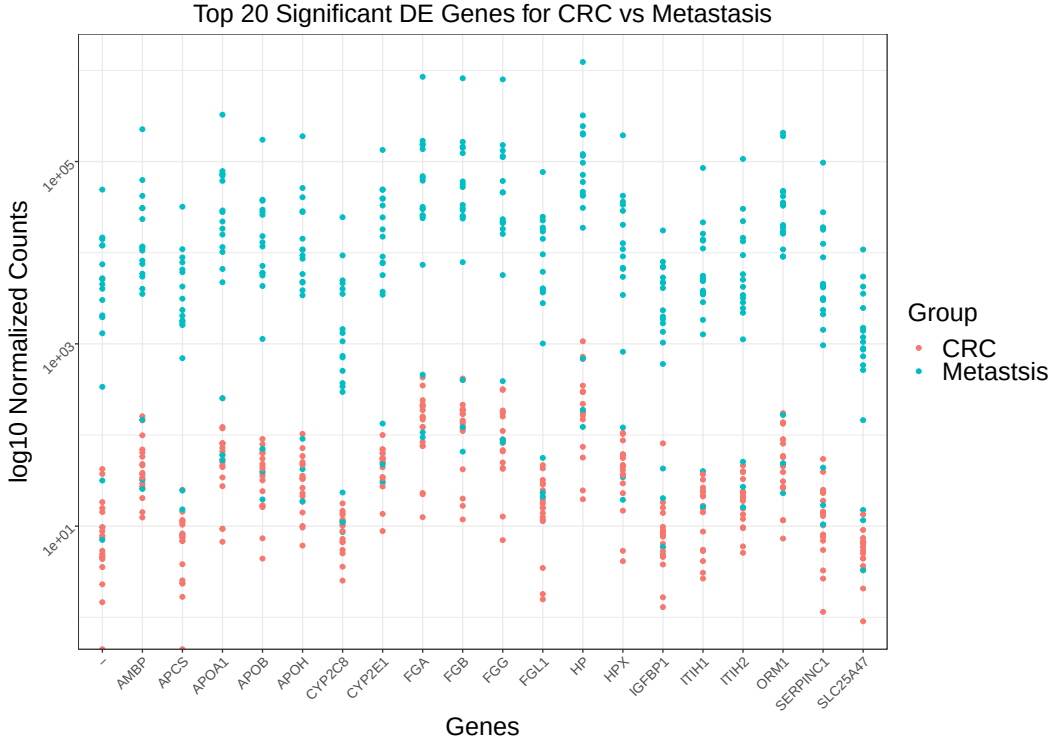


(c) Metastasis vs Normal

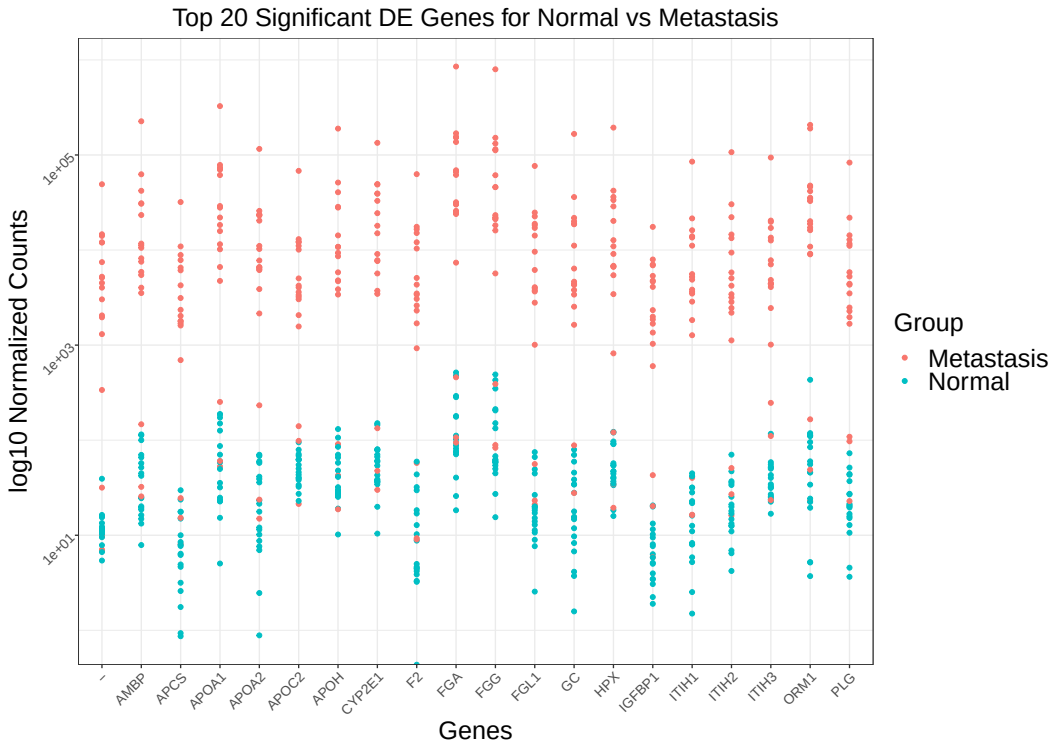
Figure 3.6: Figures showing the volcano plots for the different contrasts, (a) CRC vs Normal, (b) Metastasis vs CRC, and (c) Metastasis vs Normal. Each dot represents a gene and their relative log_2FC. Grey dots indicate genes that had no significant change, green dots indicate genes that met the log_2FC threshold of more than 1.5, blue dots indicate genes that were significantly DE (p<0.05) whereas red dots indicate genes that met both the thresholds.



(a) CRC vs Normal



(b) Metastasis vs CRC



(c) Metastasis vs Normal

Figure 3.7: Figures showing the normalised count spread of the top 20 most DE genes according to p-value in each contrast, (a) CRC vs Normal, (b) Metastasis vs CRC, and (c) Metastasis vs Normal. The genes are labelled with their known gene symbols.

3.3 Gene Processes

Functional analysis using GO (Consortium, 2019) and KEGG pathway analysis (Kanehisa and Goto, 2000) was performed.

GO analysis used the list of significant genes as an input and mapped them to biological processes within an organism, in this case *homo sapiens*. The top 30 enriched biological processes of all the significant DE genes were plotted into a bar plot in Figure 3.8. The enriched GO terms were ordered according to their significance (p-value).

In Figure 3.8a, the top 30 most significant GO terms were 'extracellular matrix organisation', 'extracellular structure organisation' and 'collagen fibril organisation', meaning these biological processes are over-expressed in the CRC samples when compared to the Normal samples. Notably, there exist processes which carry high biological significance in cancer such as 'angiogenesis' and 'immune regulation.' Furthermore, processes related to the colorectal development were also enriched, such as 'epithelial tube morphogenesis' and 'endoderm formation.'

Figures 3.8b & 3.8c represent the GO processes upregulated in the Metastasis samples, against CRC and Normal samples respectively. 'Complement activation' was the top process in Figure 3.8b, as well as various metabolic processes such as 'steroid metabolic process and 'triglyceride metabolic process. A similar result was seen in Figure 3.8c. The metabolic processes could be attributed to the fact that these samples were extracted form liver metastases, where metabolism takes place and the lipid transport and metabolism genes may show an association between lipid metabolism and cancer metastasis. Interestingly, there were a few immune response processes such as 'regulation to humoral immune response' that were also enriched in the metastases samples.

The KEGG enrichment analysis on all the significant DE genes is displayed in Figure 3.9.

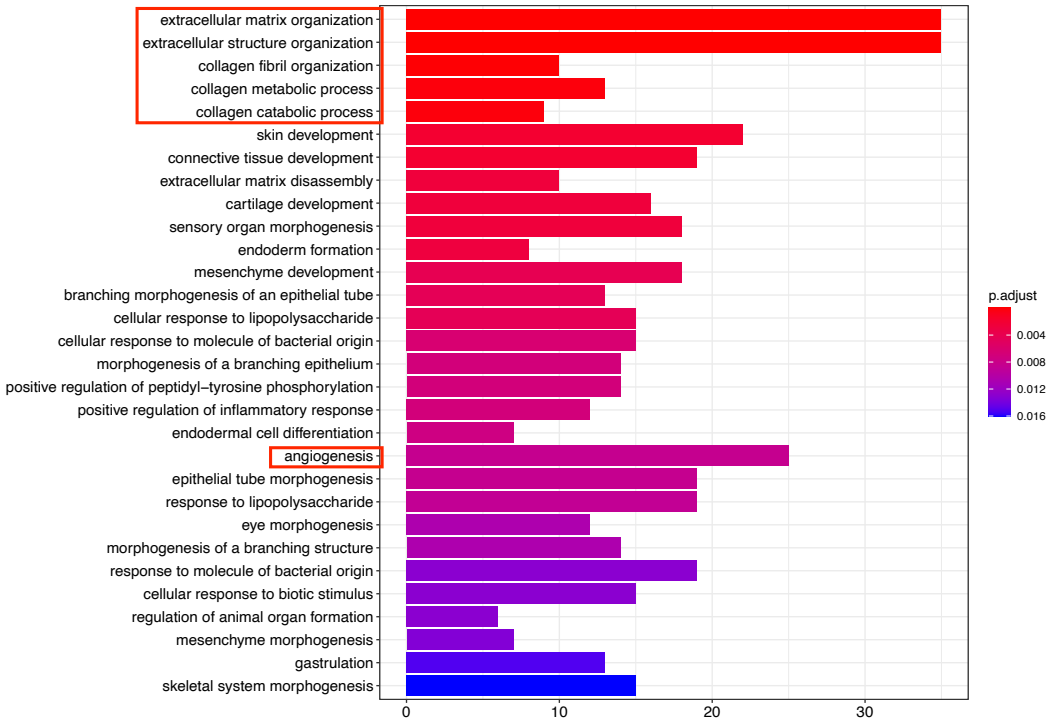
The pathways enriched in primary CRC (Figure 3.9a) include those related to cancer in the 'Wnt signalling pathway', 'Cell cycle', and 'ECM receptor interaction', the latter of which had related GO processes in Figure 3.8a. Pathways 'RNA transport', 'Ribosome' and 'Spliceosome' indicate control of gene expression. Additionally, inflammatory related pathways such as 'Cytokine-cytokine receptor interaction' and 'IL-17 signalling pathway' as well as inflammatory disease pathways 'Systemic lupus erythematosus' and 'Rheumatoid arthritis' were also enriched.

In the liver metastases (Figures 3.9b & 3.9c) the pathway 'Complement

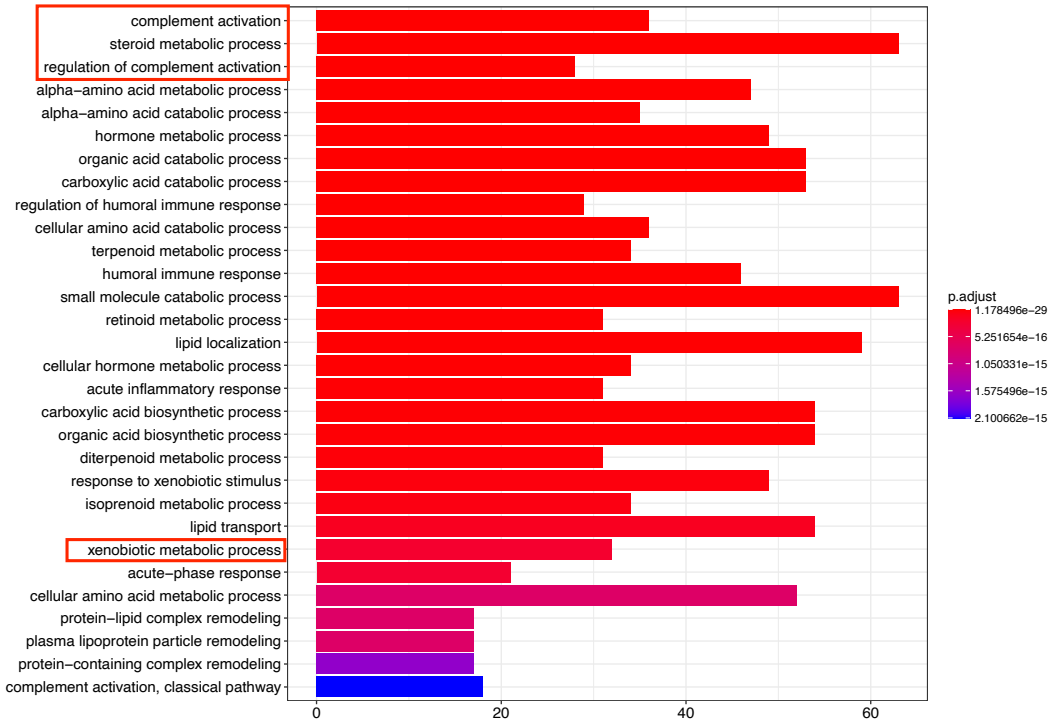
and coagulation cascades' was enriched, similar to the GO analysis of these genes (Figures 3.8b & 3.8c). 'Peroxisomes' and the 'PPAR signalling pathway' indicate lipid metabolism pathways, alongside the drug and other metabolism pathways that were also enriched. Key liver related pathways enriched included 'Biosynthesis of amino acids', 'Bile secretion' and 'Glycolysis and gluconeogenesis'.

The WebGestalt GO summaries on the gene signatures are displayed in Figure 3.10.

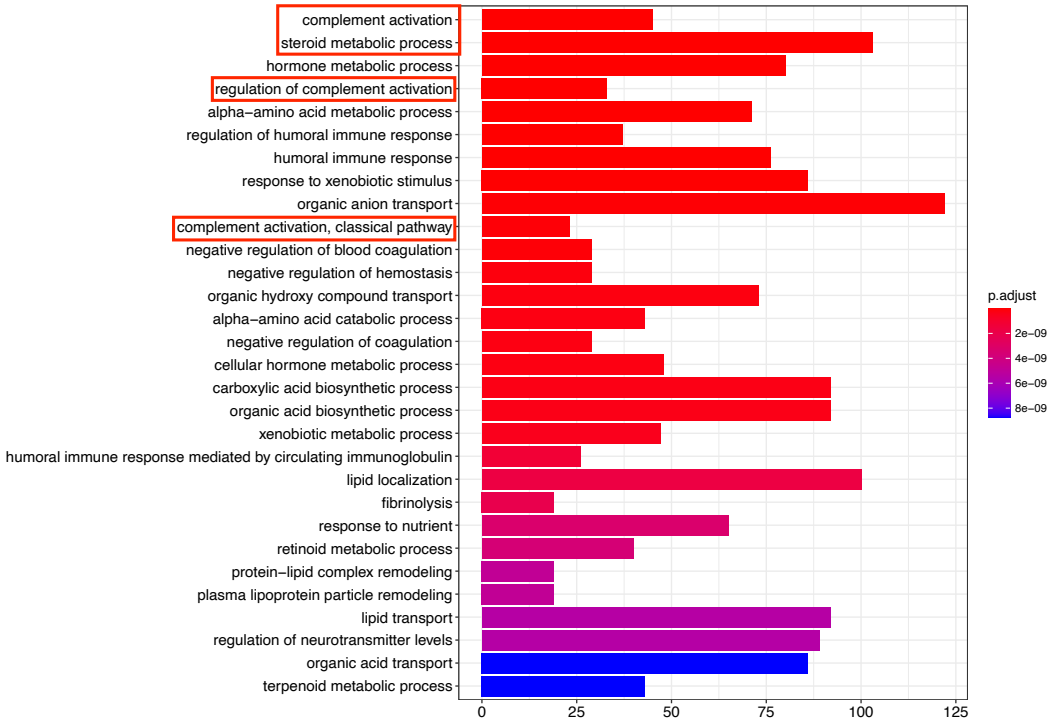
As seen in Figure 3.10a, the top biological processes involved in the CRC gene signature, with 10 of the 13 genes being involved, are 'cellular component organization', 'developmental processes' and 'biological regulation, all processes that are associated with cancer and proliferation. Similar to the previous GO results (see Figure 3.8) the signature of the metastases samples showed that 'metabolic processes' and 'biological regulation' were the top processes upregulated (Figure 3.10b).



(a) CRC vs Normal

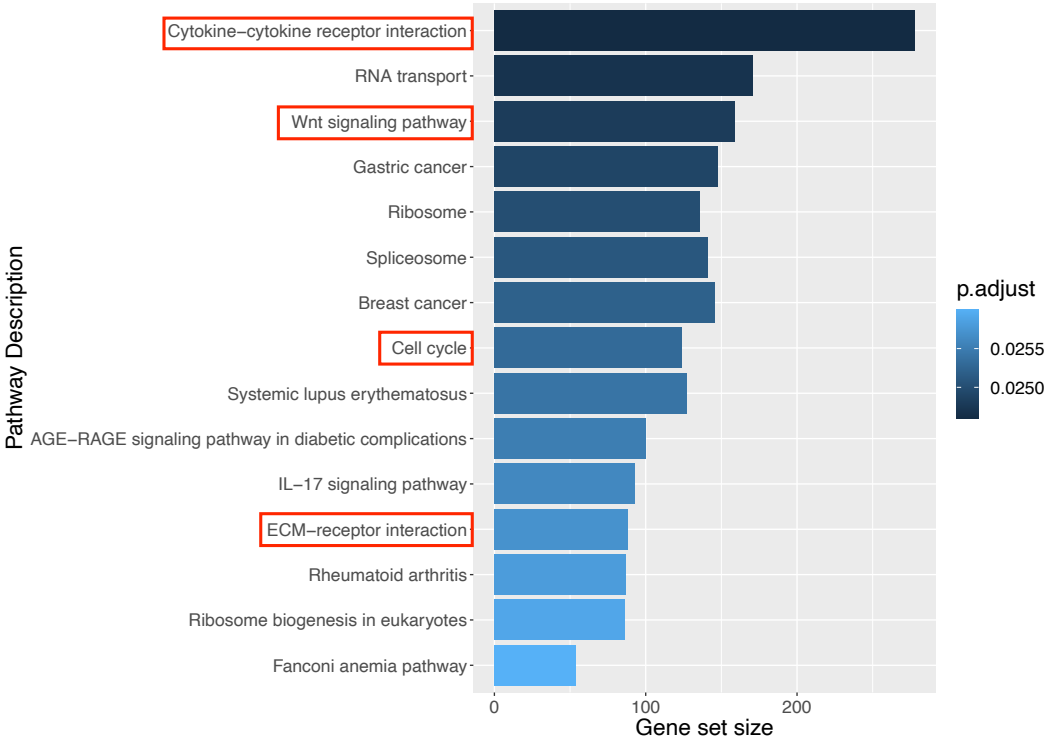


(b) Metastasis vs CRC

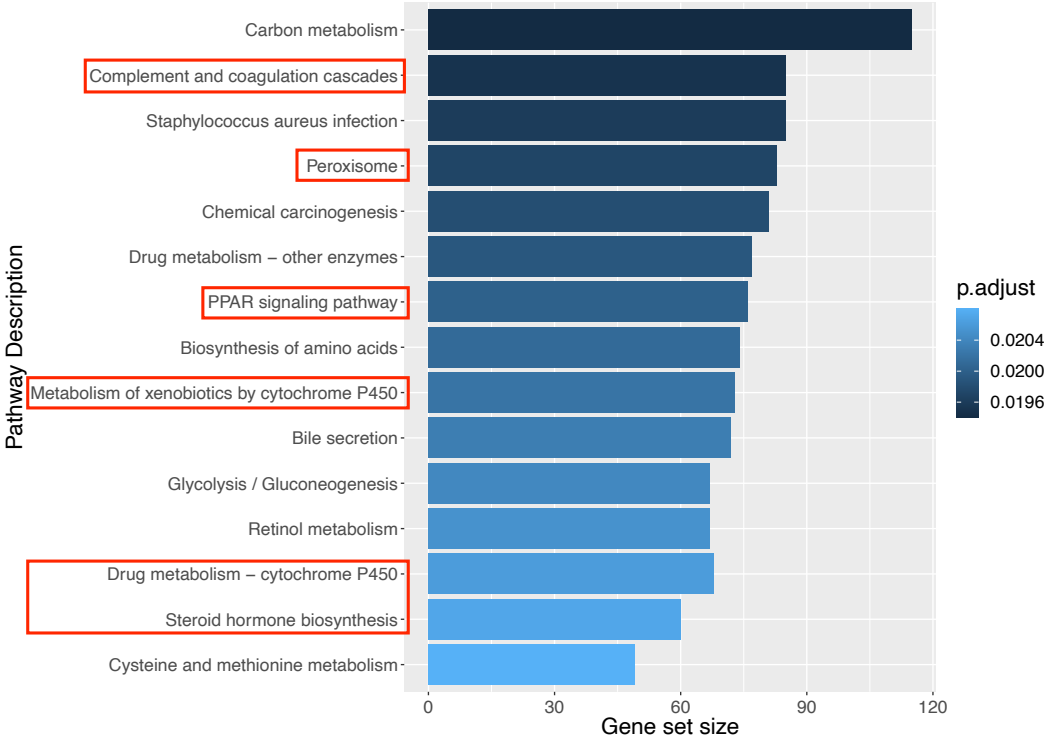


(c) Metastasis vs Normal

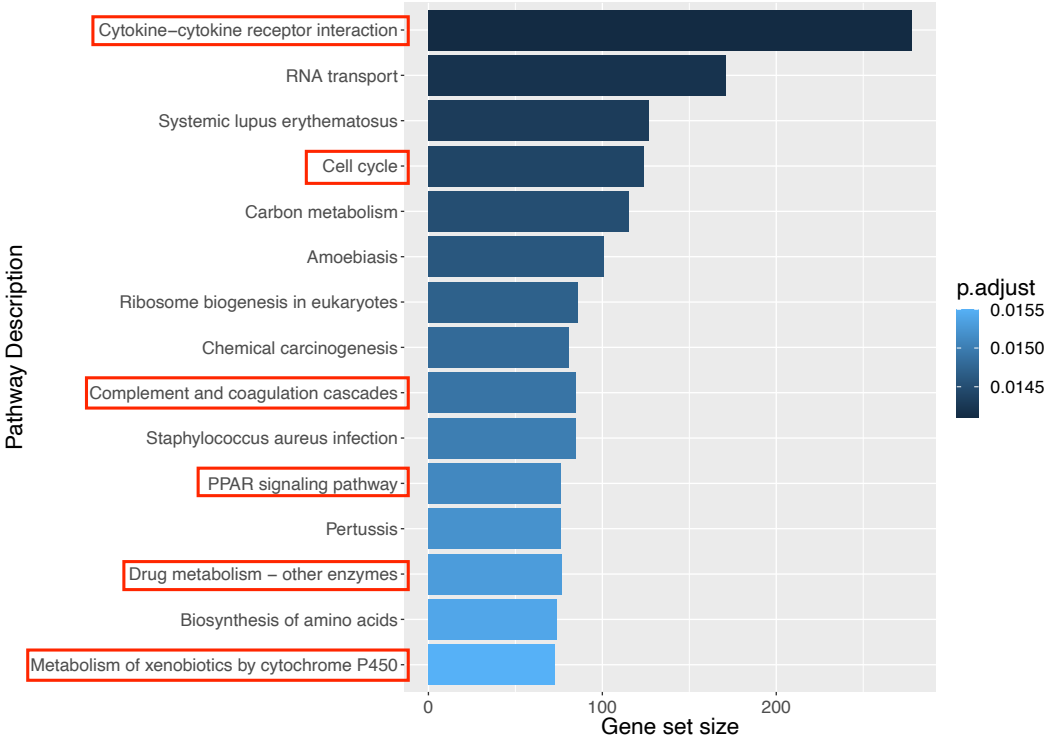
Figure 3.8: Figures showing bar plots for the Top 30 most significant GO biological processes for the different contrasts, (a) CRC vs Normal, (b) Metastasis vs CRC, and (c) Metastasis vs Normal. The GO analysis maps the DE genes to their known biological processes. Processes are given a colour according to adjusted p-value significance, with red being the most significant and blue being the least significant. The x-axis shows how many genes mapped to each GO term. Processes in red blocks indicate processes of interest in the present study.



(a) CRC vs Normal

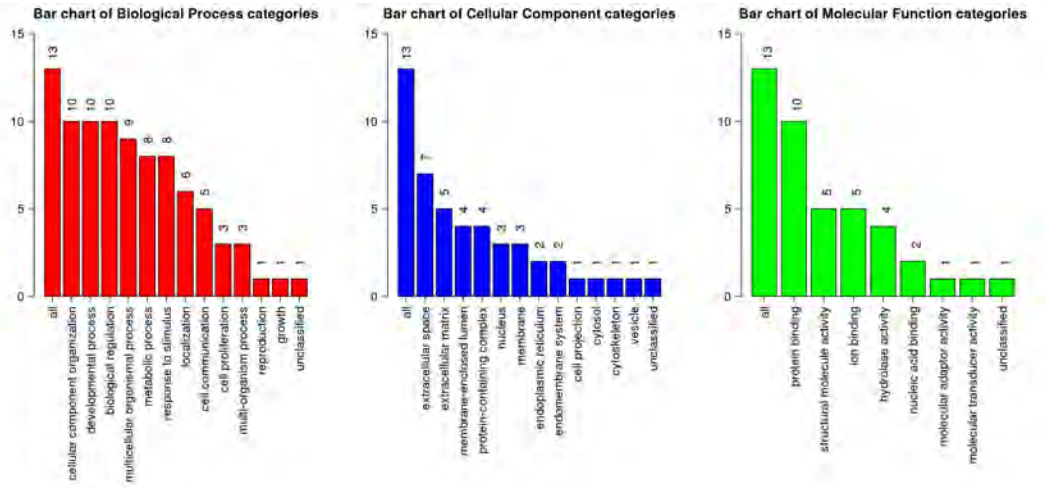


(b) Metastasis vs CRC

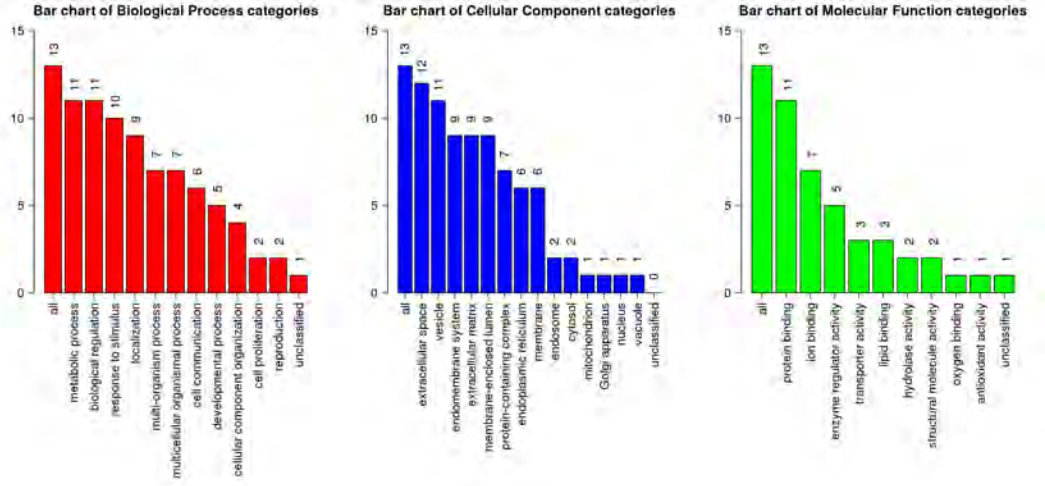


(c) Metastasis vs Normal

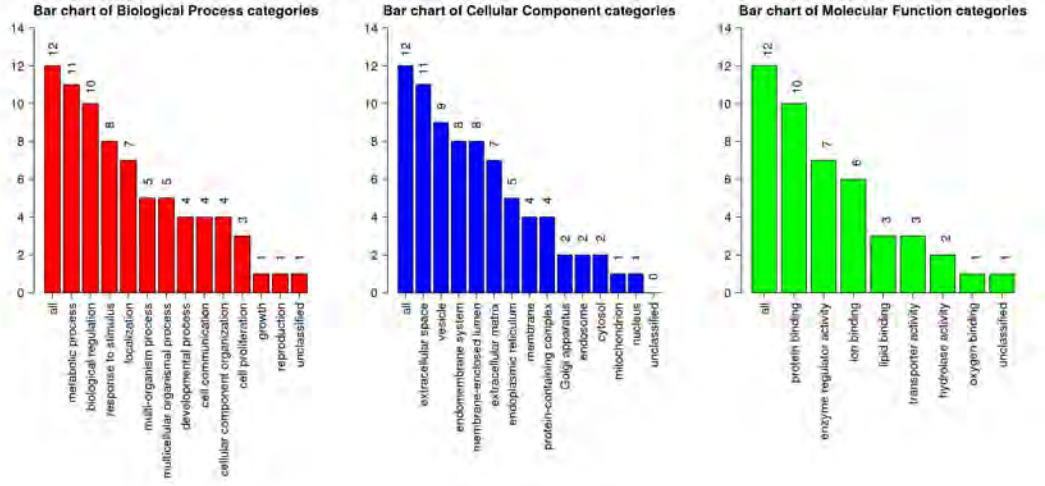
Figure 3.9: Figures showing bar plots for the top 15 KEGG enriched pathways for all the DE genes for each contrast, (a) CRC vs Normal, (b) Metastasis vs CRC, and (c) Metastasis vs Normal. Pathways are given a darker blue according to their increasing significance. The x-axis show the number of genes in the gene set that were mapped to the pathway on the y-axis. Pathways in red blocks indicate pathways of interest in the present study.



(a) CRC vs Normal



(b) Metastasis vs CRC



(c) Metastasis vs Normal

Figure 3.10: Figures showing the GO analysis bar plot for each contrast, (a) CRC vs Normal, (b) Metastasis vs CRC, and (c) Metastasis vs Normal as output from WebGestalt. The input gene signatures were Tables 3.3, 3.4 & 3.5. The x-axis represent the GO terms for each category. The y-axis represent how many genes of the signatures mapped to each term, with the number displayed at the top of each bar.

3.4 Co-Expression

A co-expression workflow was developed using WGCNA as described above. Co-expression analysis can identify co-expressed genes that are associated with specific biological functions. First, the variance stabilised expression data across all samples was used in the co-expression analysis. Initially, all samples were clustered using a simple hierarchical clustering method in R. The outlier was detected and removed by setting a cut-height that removes individual samples not stemming from the main cluster (see Figure A.2). From there, the sample dendrogram was plotted, with a group heatmap attached in Figure 3.11. While the Normal samples show clustering to one half of the dendrogram, the CRC and Metastasis samples show diversity in their cluster pattern.

In order to perform co-expression analysis, an optimal soft threshold power of β was required in order to create the initial TOM as described above. A set of powers from 1 to 20 were plotted on a scale free topology plot and a mean connectivity plot. A good soft threshold power is the lowest power seen where the curve flattens out at a high value ($R^2 > 0.8$), meaning there are higher numbers of low correlated genes helping to distinguish gene clusters. Figure 3.12 visualises these plots and a soft threshold of $\beta = 4$ was selected, as described previously.

Once the threshold power was chosen, the TOM was created and used to create a TOM dissimilarity matrix. This matrix was used to create a gene dendrogram of the expressed genes that were clustered using a typical hierarchical technique and a PAM technique. The cluster modules were merged with a distance threshold of 0.075, resulting in 32 merged module colours representing genes commonly co-expressed in each module. These merged module colours were used along with a group heatmap for each gene in Figure 3.13. The clearest association is the correlation between the metastasis genes and the yellow module. Although performing WGCNA analysis was not an objective to this study, the method was developed in line with developing an RNA-Seq analysis workflow that covers multiple bioinformatics analyses types, and the results represent a point of departure into future studies.

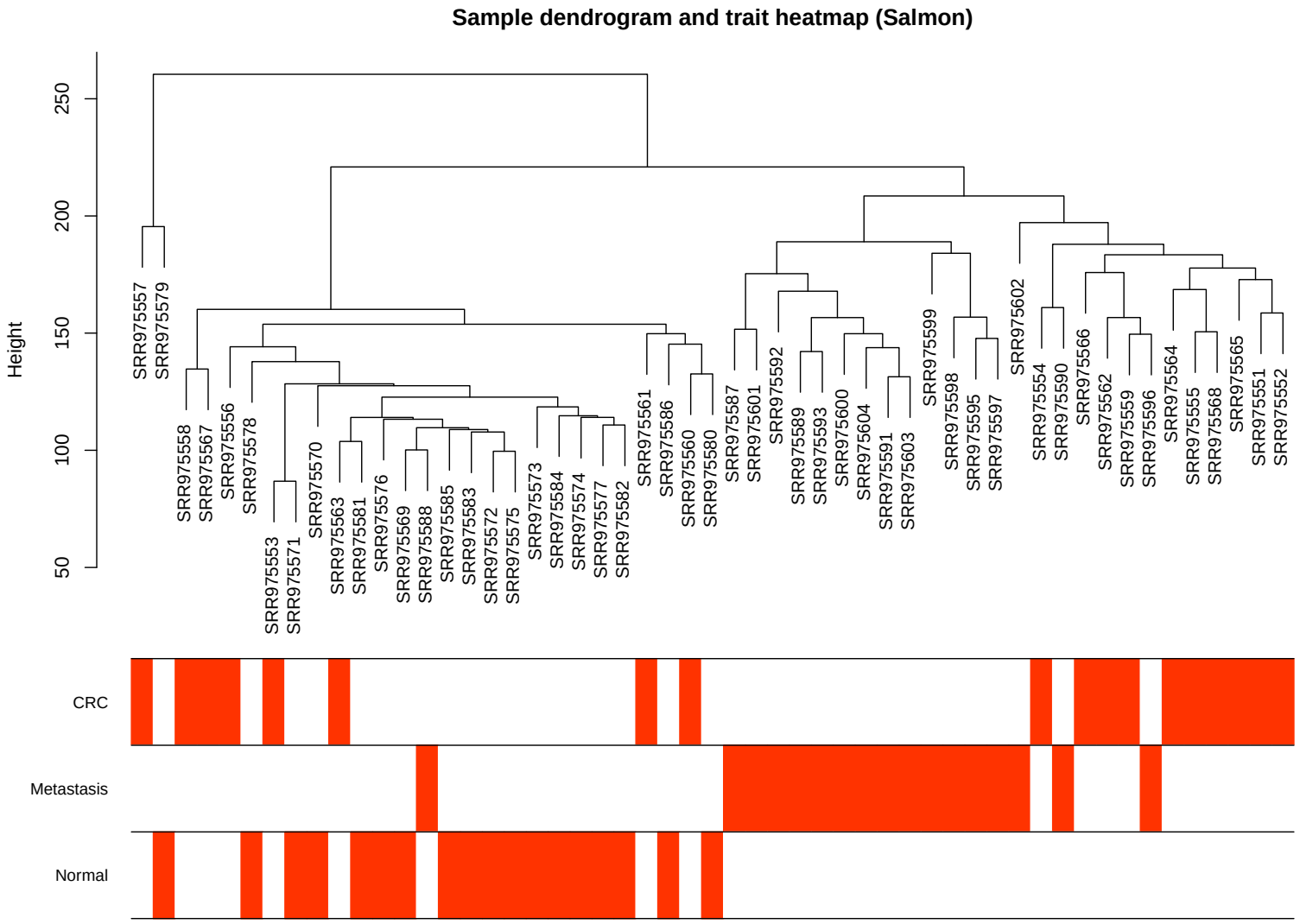


Figure 3.11: Figure showing the samples in a dendrogram with the outlier removed and a trait heatmap added for visibility showing how the expression profiles of each run (SRR) relate to each other.

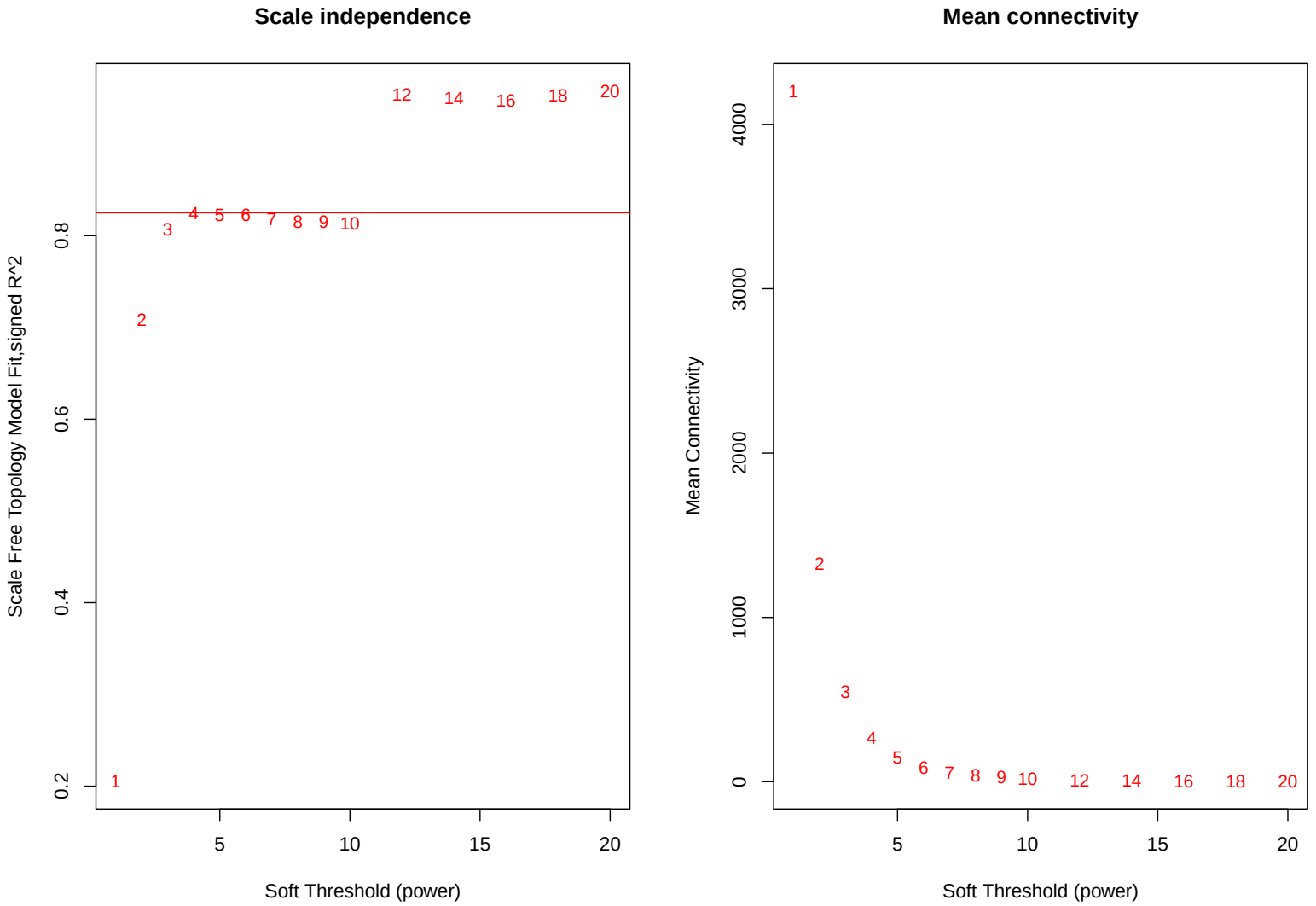


Figure 3.12: Figure showing two plots with which the chosen soft threshold values were used in a scale free topology plot and a mean connectivity plot. A soft threshold of 4 was chosen. A solid line is plotted at 0.825. The left panel represents the co-expression similarity as a function of soft threshold. The right panel represents the mean connectivity as a function of soft threshold. The lowest soft threshold value where R^2 stabilises above 0.8 is 4, which was selected for the analysis.

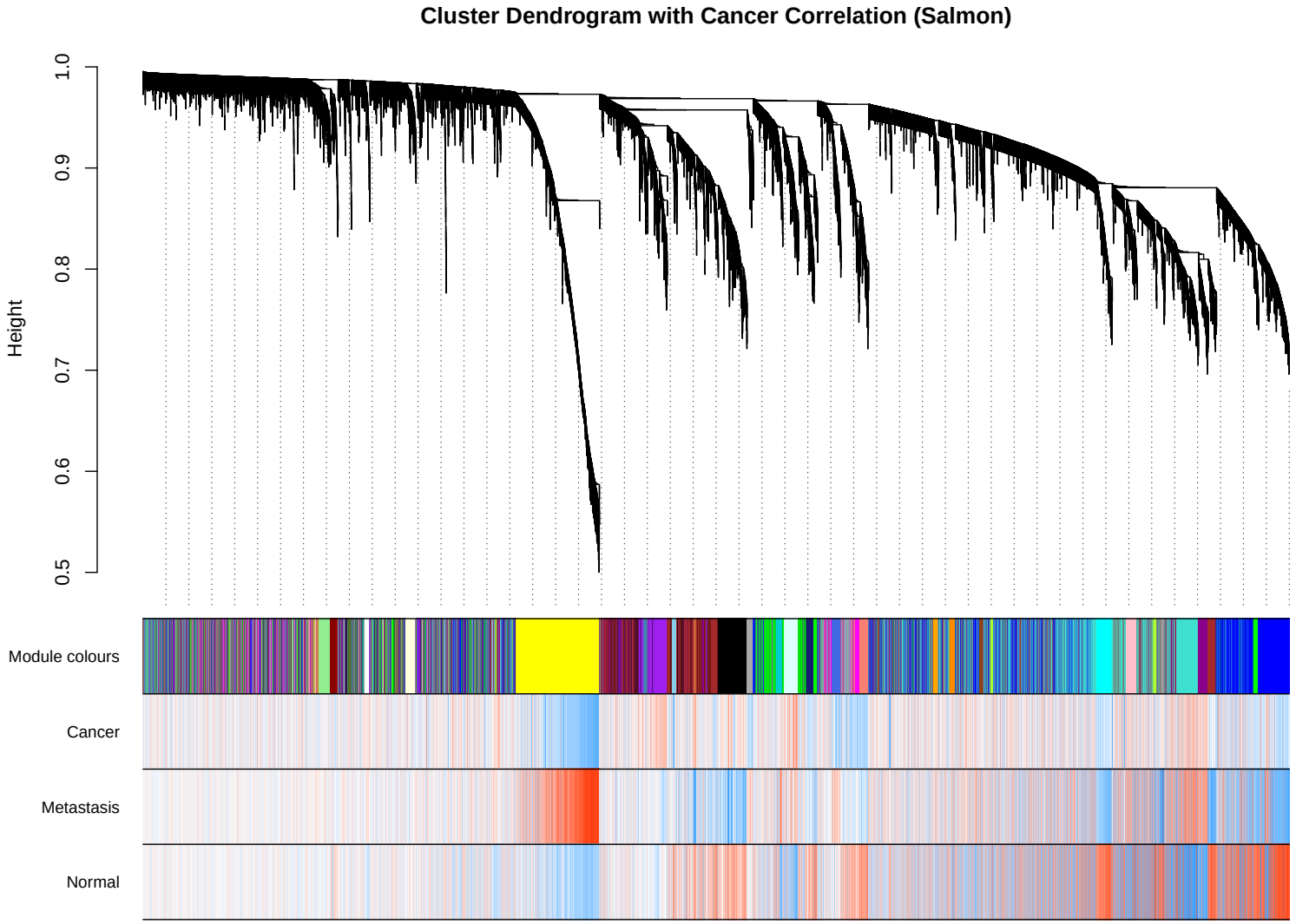


Figure 3.13: Figure showing the gene dendrogram along with a group heatmap in order to visualise modules that correlate with the sample group. Red indicates high correlation.

3.5 Validation

The gene signatures from Tables 3.3, 3.4 & 3.5 were validated for their presence in CRC patients using the CoReCG and Oncomine database (Agarwal *et al.*, 2016*b*; Rhodes *et al.*, 2004). The results of the validation are displayed in Tables 3.6, 3.7 & 3.8. The detailed Oncomine results can be found in the Appendix D.1.

Gene	Presence in CoReCG	Presence in Oncomine
COL11A1	Yes	Yes
ETV4	Yes	Yes
INHBA	Yes	Yes
ADAM12	No	Yes
CLDN1	Yes	Yes
COL10A1	No	Yes
MMP1	Yes	Yes
FAP	No	Yes
CTHRC1	Yes	Yes
CASC19	No	No
MMP3	Yes	Yes
KRT17	Yes	No
FOXQ1	Yes	Yes

Table 3.6: Table showing validation of the genes in the "CRC vs Normal" contrast and whether they are present in the CoReCG and Oncomine databases for CRC.

In Tables 3.7 & 3.8 many of the genes were not found in the CoReCG database. One reason for this is that the CoReCG database only included genes from CRC tumours, and not from the liver metastases of CRC. This reason also explains why many of the genes were found in liver cancer profiles in Oncomine, and not in CRC profiles. As the samples themselves were taken from liver metastases in patients, the results are to be expected.

Gene	Presence in CoReCG	Presence in Oncomine
CYP2E1	Yes	No
HPX	No	Yes
APOH	No	Yes
APOA1	No	Yes
APOB	Yes	Yes
FGL1	No	Yes
FGB	Yes	Yes
ITIH2	No	Yes
HP	Yes	Yes
FGA	No	Yes
ITIH1	No	No
AMBP	No	Yes
ORM1	No	Yes

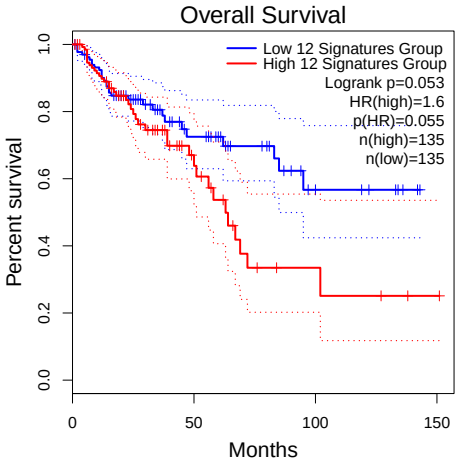
Table 3.7: Table showing validation of the genes in the "Metastasis vs CRC" contrast and whether they are present in the CoReCG and Oncomine databases for CRC and liver cancer respectively.

Gene	Presence in CoReCG	Presence in Oncomine
CYP2E1	Yes	No
HPX	No	Yes
APOH	No	Yes
APOA1	No	Yes
FGL1	No	Yes
IGFBP1	Yes	Yes
ITIH2	No	Yes
ITIH3	No	No
APOA2	No	Yes
ITIH1	No	No
AMBP	No	Yes
PLG	Yes	Yes

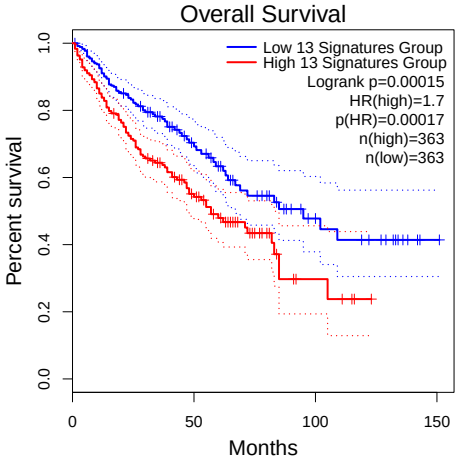
Table 3.8: Table showing validation of the genes in the "Metastasis vs Normal" contrast and whether they are present in the CoReCG and Oncomine databases for CRC and liver cancer respectively.

3.6 Survival Analysis and Stage Expression

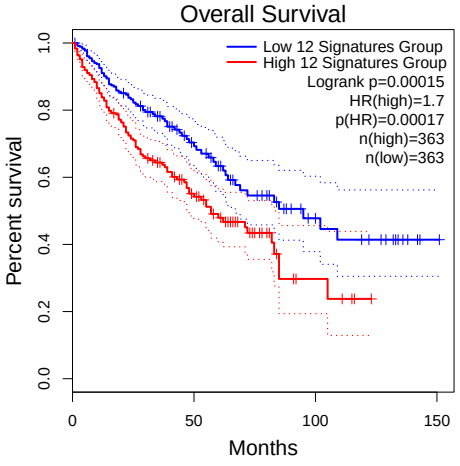
The gene signatures were then subjected to survival analysis and stage expression analysis using GEPIA2. The resulting plots in Figure 3.14 show survival analysis on the genes as a batch, and not individually. The survival analysis of individual genes was done using PROGgeneV2 as described above and these individual plots are displayed in the Appendix B.



(a) CRC vs Normal



(b) Metastasis vs CRC



(c) Metastasis vs Normal

Figure 3.14: Figures showing the Kaplan-Meier overall survival graphs for the common genes for each contrast (a) CRC vs Normal, (b) Metastasis vs CRC, and (c) Metastasis vs Normal, using GEPIA2 and TCGA COAD, READ and LIHC studies as reference. Solid lines indicate the survival analysis results, with the dotted lines indicating the 95% confidence interval. HR represents the hazard ratio using the Cox proportional hazards model. The n number represents the number of samples used in the analysis for high and low expression.

The sets of gene signatures show decreased survival with higher expression of their respective genes (Figure 3.14). Figure 3.14a was referenced against data from the TCGA COAD and READ studies, whereas Figure 3.14b & 3.14c was referenced against data from TCGA COAD, READ and LIHC to represent the liver metastases. The decreased survival with high expression of the genes show the influence these genes have on the severity of the disease.

GEPIA2 was also used in order to analyse gene expression across CRC stages I - IV, which were described above. The "CRC vs Normal" genes were mapped to TCGA COAD and READ to determine stage expression across common CRC, and were separately mapped to TCGA LIHC to determine stage expression in liver metastases. The remaining contrasts were mapped first to COAD and READ and separately to LIHC. In the interest of legibility, the individual stage expression plots for each gene in each contrast are displayed in the Appendix C.

3.7 Drug Gene Interactions

In order to analyse known drug-gene interactions, the gene signatures in Tables 3.3, 3.4 & 3.5 were uploaded to DGIdb. The interactions with drugs associated with cancer were recorded. Not all the genes were found to have an interaction.

Gene Name	Drug	Drug Description	Source List
E1AF	TRAMETINIB	Allosteric kinase inhibitor	C v N
MMP1	MARIMASTAT	MMP inhibitor	C v N
MMP3	MARIMASTAT	MMP inhibitor	C v N
APOB	LOVASTATIN	Anticholesteromic	M v C

Table 3.9: Table showing the known drug gene interactions using the common significant genes. Source list shows from which signature the gene is from, C = Cancer, N = Normal and M = Metastasis. C v N represents Sig1, M v C represents Sig2.

3.8 Biomarker Testing

The selection of genes for the prediction of CRC onset was determined by calculating sensitivity, specificity and precision for each of the genes in the gene signature in the "CRC vs Normal" contrast. This specific contrast was chosen as it represented the genes most commonly associated with the onset

of CRC, making them ideal as first line screening markers. Eight genes were selected for biomarker testing based on whether they were present in both CoReCG and Oncomine, as well as according to their stage expression profiles (seen in Appendix C). The Oncomine database was used for the calculations of predictive values. Oncomine has six major CRC studies with mRNA expression data from 100+ samples that were used in the present study. These were TCGA COAD and TCGA READ (Network and Others, 2012), Kaiser CRC (Kaiser *et al.*, 2007), Skrzypczak CRC (Skrzypczak *et al.*, 2010), Gaedcke CRC (Gaedcke *et al.*, 2010), Ki CRC (Ki *et al.*, 2007). The average sensitivity, specificity and precision was calculated for each gene using the results from the five different studies. These results are displayed in Table 3.10.

Genes	Avg Spec	Avg Sens	Avg Prec
COL11A1	92,86	99,17	99,55
INHBA	87,79	89,90	93,20
MMP1	73,55	77,78	91,26
MMP3	84,37	74,58	93,58
FOXQ1	95,82	69,79	91,05
CLDN1	98,21	81,88	93,04
COL10A1	91,64	59,09	90,42
ETV4	91,08	81,67	93,43

Table 3.10: Table showing the biomarker testing results for the genes using Oncomine data. Avg = Average, Sens = Sensitivity, Spec = Specificity, Prec = Precision.

A good biomarker is one that has around 90% for each of the parameters (Parikh *et al.*, 2008). Genes COL11A1, INHBA and CLDN1 have average of 90+% for the parameters.

The genes were also assessed using BBCancer, as previously described. The BBCancer DE and expression abundance analysis used peripheral blood expression measurements as well as EV expression measurements.

Tables for blood DE meta scores and the meta scores for the blood expression abundance of the signature genes are displayed in the Appendix D.2. A ideal biomarker is both distinguishable and detectable. The DE meta scores help show whether a biomarker could be distinguishable, with positive meta scores indicating a higher probability that the gene can be distinguished between cancer and normal patients. The expression abundance meta scores will then show whether the gene is detectable in the blood of a cancer patient. Table 3.11 shows the genes that had positive meta score values for both DE

analysis and expression abundance analysis, indicating the potential for them to be both distinguishable and detectable as biomarkers in blood.

Gene	Blood Sample	Contrast
MMP1	Blood	C v N
CTHRC1	Blood	C v N
KRT17	Blood	C v N
IGFBP1	Blood	M v N

Table 3.11: Table showing the genes that had positive meta score values for both DE and expression abundance, which represent whether a gene is both distinguishable and detectable, and from which blood sample (peripheral or EVs) they were measured as well as which gene signature they initially belonged to: C v N represents Sig1, M v N represents Sig3.

Chapter 4

Discussion

Bioinformatics continues to rise in popularity amongst cancer researchers due to biological characteristics that cancer presents through altered gene expression (Lu *et al.*, 2018). The rapidly advancing field of bioinformatics along with the rise in CRC prevalence has enabled the hybrid science to play a major role in diagnosis and treatment research by analysing DE genes, while also leading to the development of many software tools. The aim of this study was to identify DE genes in CRC with the intention to identify possible processes, pathways and biomarkers to predict the development (Normal to CRC) and progression (CRC to Metastasis) of the disease, while also developing an in house RNA-Seq analysis workflow using a variety of software and web tools for future studies.

Transcriptomics using RNA-Seq data has shown to be effective in CRC research and classification (Guinney *et al.*, 2015). Quality control of RNA-Seq data is imperative before attempting to draw out any meaningful biological conclusions. Fortunately, databases that store RNA-Seq data, such as GEO, often require or recommend researchers to perform quality control on their end and pre-process the data themselves before making it public. This has reduced the need for researchers using the existing data to perform QC themselves. However, this study had an aim to create a RNA-Seq analysis workflow that covers all the steps necessary for an in house RNA-Seq experiment and a quality control workflow was used even though the existing data had been pre-processed and trimmed by the researchers whose dataset was used for analysis in the present study Kim *et al.* (2014*b*).

FastQC was employed for quality control due to its ease of use and reliability (Wang *et al.*, 2012*a*). The reports in Figures 3.1, 3.2 indicate that the FASTQ files used from GSE70560 are of high quality and without concerning errors.

Furthermore, the GC percentage in Figure 3.3 illustrate a result close to the expected *Homo sapiens* GC percentage, again demonstrating a good quality of FASTQ files that were used.

Genes involved in the development or progression of cancers, whether it be through pathways or biological processes, can be found using DE analysis. The volcano plots in Figures 3.6a, 3.6b and 3.6c showed that a significant number of genes were DE amongst the different samples, and the most significant DE genes were used as gene signatures.

4.1 Identified Gene Signatures

4.1.1 Identified Primary CRC Gene Signature

The primary CRC tumour gene signature, Sig1, consisted of the following genes: *COL11A1, E1AF, INHBA, ADAM12, CLDN1, COL10A1, MMP1, FAP, CTHRC1, CASC19, MMP3, KRT17* and *FOXQ1* (Table 3.3). Collagen-related genes such as *COL11A1* and *COL10A1* were present in this signature. Collagen makes up a major component of the extracellular matrix (ECM), which has been shown to have an active role in many biological processes, including cell proliferation, differentiation and migration, and has also been shown early in cancer research to have a role in tumour progression (Boudreau and Bissell, 1998; Stracke *et al.*, 1994). *COL11A1*, has been shown to be commonly upregulated in cancers such as CRC, pancreatic cancer, breast and lung cancer, as well as being identified as a prognostic biomarker in ovarian cancer (Su *et al.*, 2019; Wu *et al.*, 2014). *COL11A1* has also been studied as a potential biomarker for CRC, with upregulation being observed in cancer stromal cells (Galván *et al.*, 2014). It has also been hypothesised that the expression of *COL11A1* could be the primary factor in the tumourigenesis of colon epithelial cells, while also being involved in the immune physiology of CRC, with a positive correlation between immune cell infiltration and high *COL11A1* expression observed in COAD (Wu and Xu, 2020). High expression of *COL10A1* has been correlated to poor prognosis, as well as invasion and metastasis in CRC (Huang *et al.*, 2018*b*; Li *et al.*, 2018).

Another set of genes from the same family found to be upregulated in primary CRC was *MMP1* and *MMP3*, genes encoding for matrix metalloprotease proteins (MMPs), of which many are involved in CRC (Zucker and Vacirca, 2004). MMPs act as collagenases, and are involved directly and indirectly in the degradation of surrounding ECM and its individual components such as

collagen, fibronectin and proteoglycans (Zucker and Vacirca, 2004). Studies have shown that MMPs may be involved in CRC progression, with overexpression of various MMPs, including the ones identified here, being observed in CRC tumours (Zucker and Vacirca, 2004). It was hypothesised that MMPs are initially produced by fibroblasts and inflammatory cells, and then dock at cancer cells providing the protease capability they need in order to perform invasion processes (Yu and Stamenkovic, 2000). High expression of MMP1 has been associated with poor prognosis in CRC patients (Murray *et al.*, 1996; Shiozawa *et al.*, 2000), with the high expression possibly occurring due to a mutation in the promoter region of the *MMP1* sequence (Sunami *et al.*, 2000). In CRC with low MSI, as is the case in the present study, a higher expression of MMP3 was observed (Morán *et al.*, 2002).

An interesting gene found upregulated in CRC and a member of the transforming growth factor (TGF-β) family was *INHBA*. TGF-β is commonly overexpressed in cancer in order to evade the immune system by inhibiting natural killer cells and cytotoxic T cells (Rook *et al.*, 1986; Thomas and Massagué, 2005). *INHBA* expression has been associated with CRC, and identified as a possible prognostic predictor for patients with COAD (Li *et al.*, 2020; Okano *et al.*, 2013).

Intriguingly, the high expression of the gene *CLDN1* was found to be a possible marker for better survival prognosis in CRC (Nakagawa *et al.*, 2011). However, separate studies found that an antibody targeting CLDN1 could be a possible therapeutic avenue for CRC treatment (Cherradi *et al.*, 2017) and that serum levels of *CLDN1* in CRC patients could be used as markers for differential diagnosis of cancer (Karabulut *et al.*, 2015). With the conflicting evidence, the role of *CLDN1* in cancer remains unclear, however given that the CLDN1 protein is involved with cell-cell adhesion, and with its presently observed high expression, it could still be regarded as a potential prognostic biomarker (Eftang *et al.*, 2013).

The cancer susceptibility 19 gene, or *CASC19*, a long non protein coding RNA gene (lncRNA), was found highly expressed in the primary CRC tumours. Non-coding RNAs are associated with regulatory roles in biological processes and epigenetics (Lee, 2012), with their dysregulation previously associated with cancers, including CRC (Wang *et al.*, 2017*b*; Liang *et al.*, 2018; Yang *et al.*, 2018). Another study found increased CRC *CASC19* expression to be associated with the promotion of proliferation and an inhibition of apoptosis *in vitro* (Wang *et al.*, 2019). This could be through a regulation of the cell

migration inducing hyaluronidase 1 (*CEMIP*) gene, which is possibly involved in the Wnt/β-catenin pathway (detailed below) as well as the degradation of hyaluronan, a protein expressed in the colon, leading to an angiogenic response (Fink *et al.*, 2015).

4.1.2 Identified Liver Metastases Gene Signature

The genes representing the liver metastases signature were: *CYP2E1, APOH, APOA1, APOA2, APOB, APOH, HPX, FGL1, ITIH1, ITIH2, ITIH3, AMBP, PLG, ORM1, HP,* and *PLG*. The most significantly DE of these was *CYP2E1*. The cytochrome P450 enzymes (CYP) are involved in the metabolism of drugs and are mainly expressed in the liver although have been observed in other tissues such as the lung and kidney (Coon, 2005; Gonzalez, 2005). However, CYP2E1 has been shown to produce toxic products including high amounts of ROS (Caro and Cederbaum, 2004). ROS has been associated with cancer cells, with elevated levels damaging DNA and proteins promoting cell instability and tumourigenesis (Moloney and Cotter, 2018). The exact justification for the expression of the CYP enzymes in the present study remains unclear, as it could also be that the increased expression is due to patient drug treatment.

The presence of apolipoprotein genes *APOH, APOA1* and *APOA2* as well as *APOB* in the upregulated genes was interesting as they are all involved in lipoprotein metabolism directly or indirectly (Borgquist *et al.*, 2016). A previous study done to identify key genes specifically in liver metastases of CRC found the following hub genes that were also found in the present study: *APOA1, APOB, APOH AMBP* and *PLG* (Zhang *et al.*, 2019*b*). APOA1 is involved with the efflux of cholesterol from cells to the liver, and was found to be decreased in late stage CRC patient colon cells (Peltier *et al.*, 2016), in contrast to the upregulated DE observed in the present study liver metastases cells. High levels of APOB has been associated to lung cancer incidence, with APOH associated with cancer promoting processes such as cell growth and angiogenesis in hepatocellular carcinoma (HCC) (Borgquist *et al.*, 2016; Jing *et al.*, 2015). AMBP is normally highly expressed in the liver and may be a false positive, however associations between increased AMBP expression in gastric cancer was linked to a poor response to paclitaxel-capecitabine chemotherapy (Huang *et al.*, 2013). PLG is involved in the regulation of immune response as well as cancer related processes angiogenesis, invasion and metastasis (Kumari and Malla, 2015). This is possibly due to the plasmin protein for which PLG encodes for being involved in the fibrinolytic system, of which certain compo-

nents are upregulated in cancer for the process of angiogenesis (McMahon and Kwaan, 2007).

4.2 Identified Biological Processes and Pathways in CRC

As the total number of significantly DE genes across the contrasts was so large (Table 3.2), it would not be viable to investigate the effects of genes individually as done above for some of the signature genes. Therefore, the GO analyses in Figure 3.8 allowed for a summative analysis of the genes and their BPs.

4.2.1 Gene Ontology Based Processes

The BPs most significantly upregulated in primary CRC included many terms related to the ECM, which plays a major role in tumour progression by directly affecting the tumour microenvironment (TME) (Figure 3.8a). There is an interplay between proliferating tumour cells and the surrounding ECM, helping cells migrate through surrounding tissues (Walker *et al.*, 2018). Therefore, seeing ECM-related processes upregulated is promising in looking for biomarkers of CRC, as these processes indicate one of the many factors enhancing tumour progression. Another notable process upregulated was 'angiogenesis', which is regarded as one of the hallmarks of cancer and cancer progression (Carmeliet and Jain, 2000). As cancer cells proliferate, not only do they require greater metabolic resources, but also an increase in blood supply. Angiogenesis accomplishes this by forming new blood vessels, involving genes previously observed in the signatures such as *PLG*. There were also processes related to morphogenesis, possibly involved in tumour progression and growth, and explicit terms labelled 'mesenchyme development' and 'mesenchyme morphogenesis' which could relate to the EMT process (Kalluri and Weinberg, 2009). Furthermore, cancer cells often show morphogenic capabilities similar to that observed in embryonic development, contributing to increased proliferation and differentiation processes (Brabletz *et al.*, 2002).

The most significantly upregulated processes for the DE genes in the liver metastases included the processes 'complement activation' and 'regulation of complement activation'. This was also found in the KEGG pathways, and the interaction between the complement system and CRC is discussed below.

Many metabolic processes were also upregulated in the liver metastases samples, including those involved with steroid and lipid metabolism. It has been found that higher levels of low density lipoprotein (LDL) cholesterol is associated with liver metastases of CRC, and could possibly account for the increased expression of these genes as the liver metabolises these products. (Wang *et al.*, 2017*a*). Metabolism pathways are further discussed below.

4.2.2 Identified Pathways Perturbed in CRC

The pathways upregulated in CRC, include inflammatory as well as gene expression control pathways of which two in particular will be discussed: 'WNT signalling' as it is related to the *APC* mutation commonly observed in CRC, and the 'ECM-receptor interaction' as ECM related processes were upregulated in the GO analyses.

The 'WNT signalling pathway' is a well studied pathway in CRC. Briefly, β-catenin is a protein involved in gene transcription of growth and proliferation genes (Logan and Nusse, 2004). Normally, a destruction complex degrades β-catenin. However, when WNT binds to the frizzled receptor, the destruction complex is inactivated leading to an increase in the levels of β-catenin and thus an increase in the transcription of growth and proliferation genes (Logan and Nusse, 2004). This is significant in CRC, as a common mutation discussed previously, the *APC* mutation, leads to a dysfunctional APC protein which is part of the destruction complex. Thus, the increased activation of the WNT signalling pathway, as well as *APC* mutations lead to an increase in β-catenin, and consequently an increase in growth and proliferation genes that help in the progression of CRC (Bienz and Clevers, 2000).

As described previously the ECM and relation to the TME plays a major role in the development of CRC, and GO processes in ECM organisation and structure were seen upregulated. 'ECM-receptor interaction' is described as interactions mediated by transmembrane molecules which in turn have direct and indirect control on cellular activities such as differentiation, proliferation, and apoptosis. The upregulation of this pathway further emphasis the role the ECM and TME have on CRC development.

Figures 3.9b & 3.9c show the KEGG pathways upregulated by liver metastases of CRC. The 'complement and coagulation cascades' was a common pathway upregulated, similar to the upregulated GO term 'complement activation' discussed earlier. The complement system is involved in immune response, and aids in the removal of pathogens and damaged cells (Sarma

and Ward, 2011). The TME contains stromal cells such as tumour associated macrophages (TAMs), tumour associated neutrophils (TANs) and myeloid derived suppressor cells (MDSCs) to name a few, which are immunosuppressive cells (Afshar-Kharghan and Others, 2017). Complement activation within the TME has been shown to enhance tumour growth and increase metastasis, described in Figure 4.1. Evidence has shown complement activation promoting EMT and angiogenesis (Cho *et al.*, 2016; Zhang *et al.*, 2019*a*). Furthermore, the coagulation cascade could be a pathway inherited by the metastases cells that used coagulation to protect circulating tumour cells from shear stress, prevent natural killer cell targeting, and facilitate angiogenesis (Gil-Bernabé *et al.*, 2013). Consequently, anti-coagulant drugs have been shown to hamper metastatic performance of cancer (Algra and Rothwell, 2012).

An interesting pathway upregulated was 'steroid hormone biosynthesis', relating to increased production of hormones oestrogen and progesterone which are associated with the regulation of cell proliferation survival and development (Madhunapantula *et al.*, 2010). Cholesterol acts as a precursor for the synthesis of these hormones, and the changes in apolipoprotein gene expression observed above demonstrate the role cholesterol may play in liver metastases (Gu *et al.*, 2019).

Other metabolism pathways were upregulated, such as the 'PPAR signalling pathway' which is the peroxisome proliferator-activated receptor (PPAR) pathway regulating gene expression of genes involved in lipid metabolism as well as cell proliferation and differentiation (Liu *et al.*, 2018). Drug metabolism pathways were upregulated as well, possibly due to the liver being the main organ for drug metabolism, or alternatively lending to a hypothesis that liver metastases cells use drug metabolic molecules and transporters to aid in clearance of the drug before it affects the cancer cells, thus inducing drug resistance. However, with patient treatment data being unavailable due to confidentiality, it could be that the patients from which the liver metastases samples were extracted and sequenced could have been undergoing a drug therapy, leading to a subsequent increase in the metabolism pathways of the liver.

The KEGG analysis overall showed many pathways involved in cancer and further corroborate the GO analysis results, linking the significantly DE genes and gene signatures to processes and pathways involved with CRC development and progression. These results support the analysis workflow used in the present study and highlight processes and pathways contributing to CRC and metastases.

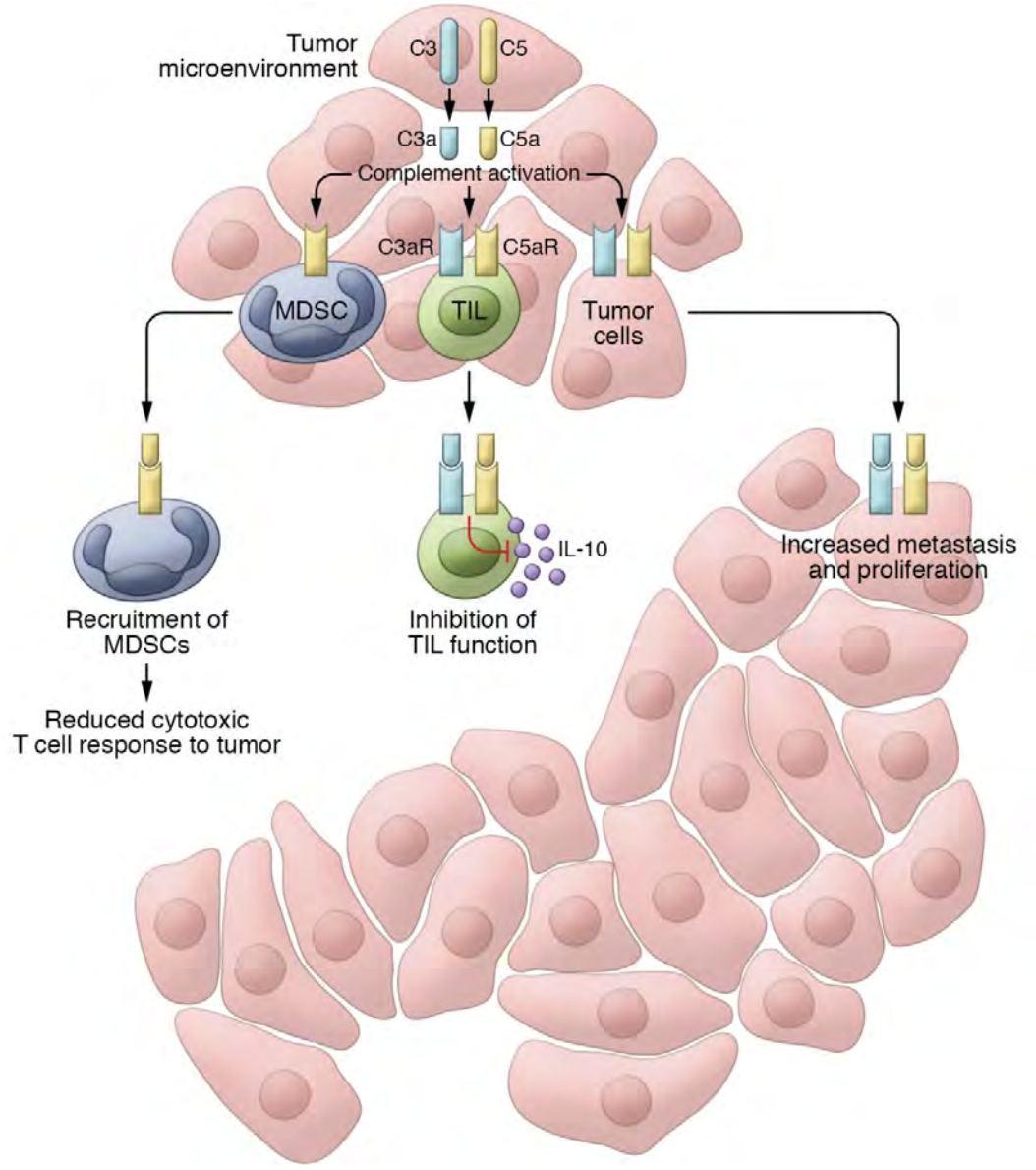


Figure 4.1: Figure showing how activation of the complement system within the TME enhances cancer cell metastasis and proliferation. The release of complement proteins and activation of their receptors lead to reduced cytotoxic T-cell ability through myeloid derived suppressor cells (MDSCs), as well as inhibiting interleukin-10 (IL-10) release from tumour infiltrating lymphocytes. Figure taken from Afshar-Kharghan and Others (2017).

One key benefit of the analysis performed in the present study supported the identification of hub genes which represent key genes in the module, with high connectivity to the other genes and often plays a role in the phenotype of the sample group (Langfelder and Horvath, 2008). The hub gene of the yellow module was beta-ureidopropionase 1 (*UPB1*). UPB1 has biological functions in metabolism of beta alanine, as well as the biosynthesis of coenzyme A (CoA), and has been identified as a possible circulating biomarker for HCC (Awan *et al.*, 2015). *UPB1* was also investigated as a prognostic gene signature for HCC, alongside the genes *SOCS2* (suppressor of cytokine signalling 2) and *RTN3* (reticulon 3) (Li *et al.*, 2017). Because the yellow module was most correlated to liver metastases, the association of this particular hub gene to HCC is partly expected. The co-expression analysis performed here acts as a point of departure for a more in depth co-expression study on CRC and liver metastases.

4.3 *In silico* Validation of the Identified Gene Signatures

In order to determine how influential the genes were on CRC prognosis, survival analysis was performed in a signature batch analysis as well as individually with each gene. High expression of the primary CRC tumour genes resulted in decreased survival with an almost significant p-value of 0.053 (Figure 3.14a). High expression of the two metastases gene signatures had decreased survival both with a significant p-value of 0.00015 (Figures 3.14b & 3.14c). These results emphasise the role the identified genes have on CRC patient prognosis, showing the potential for these genes to be used as prognostic markers.

Of the individual primary CRC genes stage expression analysis (Appendix C), *CLDN1* had high expression across all stages. *COL10A1, COL11A1, FAP, FOXQ1* and *INHBA* had increasing expression with stage progression. *ETV4/E1AF* had a notable increase in average expression in the later stages. High expression in different stages of CRC is important when identifying a potential biomarker as those expressed in later stages could correlate to the severity of the cancer in an individual.

The liver metastases genes had lower expression counts than the primary CRC genes when mapped to COAD, making it difficult to identify genes with correlation to CRC stage expression. When mapped to TCGA LIHC samples, the genes had clearer stage correlations. Many of the genes were commonly

found in HCC, explaining the high expression across stages in LIHC. Similar to the metastasis genes only being significant in liver cancer studies in Oncomine, the stage expression plots emphasise the specificity of the gene expression to the liver. This data suggests that the identified biomarkers show variable expression profiles across different stages of CRC, thus providing support for their further exploration in the laboratory setting or in CRC patients' samples for potential use as biomarkers.

Of the 12 genes representing the primary CRC gene signature, 8 were found in CoReCG studies and 10 were found in Oncomine CRC studies (Table 3.6). The presence of the genes in these databases demonstrates that these genes may have high potential to act as biomarkers. Genes present in both CoReCG and Oncomine were particularly noted, and were: *COL11A1, ETV4, INHBA, CLDN1, MMP1, CTHRC1, MMP3, FOXQ1*. Briefly discussed above, these genes have previously been studied in CRC. Furthermore, these genes were found in the top 1% of DE genes in TCGA studies in Oncomine, with high fold changes and significance (Appendix Table D.2). *CASC19* was not found in any database and could be due to it being an lncRNA gene and not an mRNA gene.

Of the liver metastases gene signature, many were found in Oncomine under liver cancer studies, whereas very few were found under CRC studies or in CoReCG (Table 3.7 & 3.8). The only genes found in both databases were: *APOB, FGB, IGFBP1* and *PLG*. This could possibly lead to two scenarios; either these are new potential biomarkers that should be explored further, or the lack of confirmation in previous studies and databases makes it difficult to consider them as liver metastases biomarkers. With the latter in mind, biomarker testing was only done on the primary CRC genes that were found in previous literature associated with CRC.

4.4 Identified Potential CRC Biomarkers

An ideal biomarker is one that can be measured with minimal invasion, while still having high levels of sensitivity, specificity and precision. CRC is not currently clinically identified using the expression of specific genes, but is commonly identified by the presence of certain gene mutations, such as *BRAF* or *APC* mutations, and microsatellite instability (MSI) as discussed earlier (Vega *et al.*, 2015; Nakanishi *et al.*, 2019; Zoratto *et al.*, 2014). The ability to detect CRC presence by the expression of genes alone would add to the arsenal of

diagnostic tools in the clinical setting. With the above-described analyses and results in mind, the selected CRC genes were tested as biomarkers via two methods: blood biomarker testing and tissue biomarker testing.

4.4.1 Possible Blood Biomarkers for CRC

One way to test for biomarkers is using blood samples. Cell free DNA (cfDNA) consists of fragments of DNA that originate from the primary CRC tumour or liver metastases that circulate in the plasma and can be detected in peripheral blood (Tan *et al.*, 2016). Circulating mRNA may also be detectable in blood, and along with RT-PCR may provide a new diagnostic avenue for CRC biomarker detection (Siravegna and Bardelli, 2016; Gingras *et al.*, 2015). A previous study was able to develop a seven gene blood based biomarker panel for detection of CRC, which had a sensitivity of 72% and a specificity of 70% (Marshall *et al.*, 2010; Yip *et al.*, 2010). A blood-based biomarker test has a range of benefits including enhancing patient participation by providing an alternative to colonoscopy, identifying patients with greater risk, and reducing healthcare costs by prioritising additional colonoscopy procedures for these patients. This would improve efficiency and aid in physician and patient decision making. A combination of pre-screening with colonoscopy has shown a 2-4 times improvement in the detection of cancer (Marshall *et al.*, 2010).

In order to test the blood biomarker potential for the above identified genes, the blood expression levels of the genes were assessed using the BBCancer database. As explained above, a biomarker needs to be distinguishable as well as detectable. The genes that were both distinguishable and detectable in blood were *MMP1, CTHRC1, KRT17* and *IGFBP1* (Table 3.11). This was fewer genes than expected, given the significant overexpression of these genes following the DE analysis and with previous studies having shown the majority of the signature genes to be overexpressed in CRC. This could be attributed to the limited datasets available in the BBCancer which hosts 524 samples of CRC, with only 124 having been analysed using RNA-Seq with the rest analysed as microarray data (Zuo *et al.*, 2020). Of all the RNAs measured in CRC, less than 200 samples were analysed for mRNA with the majority of samples being analysed on miRNA. With this in mind it could be that the studies hosted in BBCancer do not represent a comprehensive overview of the accurate mRNA gene expression in blood. The creators of BBCancer also noted that in CRC the most detectable form of RNA (with the highest expression abundance) was transfer RNA fragments (tRFRNAs), and

suggested tRFRNAs as more effective blood biomarkers for early detection of CRC (Zuo *et al.*, 2020).

4.4.2 Possible Tissue Biomarkers for CRC

While the blood based biomarker testing resulted in only a few genes being distinguishable and detectable, the other genes were tested as pure biomarkers of CRC using tissue samples. The genes identified to have high (90%+) sensitivity, specificity and precision were *COL11A1, INHBA* and *CLDN1* (Table 3.10). *ETV4* and *FOXQ1* represented the next best genes as potential biomarkers (85%+).

4.5 Conclusion

This study had two aims, a primary aim to identify potential biomarkers of CRC and a secondary aim to develop a workflow for the identification of DE genes in CRC. A workflow was developed for the DE analysis of RNA-Seq data, as well as methods for functional analysis, survival analysis, literature validation and biomarker testing, achieving the secondary aim. The entire workflow identified many processes and pathways involved in CRC. Notably, processes involved in altering of the ECM through collagen played a significant role in the onset of CRC. The related genes *COL10A1* and *COL11A1* along with protease genes *MMP1* and *MMP3* were found to be upregulated assisting in the progression of CRC from initial polyps. In terms of further progression of CRC from primary to liver metastases, processes and pathways involved in metabolism and complement activation had the greatest association.

The workflow then, using a variety of software, tested the RNA gene signatures as biomarkers, achieving the primary aim and yielding two types: non-invasives that could be measured in blood, as well as invasives that could be measured from tissue. The genes *MMP1, CTHRC1, KRT17* and *IGFBP1* represent the non-invasive biomarkers recommended here that should be tested further *KRT17* was not found in the validation databases, and could signal a novel gene marker. These blood based biomarkers were involved in the processes discussed, *MMP1* with ECM processes, *CTHRC1* with the Wnt pathway, *KRT17* with tumour proliferation and *IGFBP1* involved with growth. The genes: *COL11A1, INHBA*, and *CLDN1*, represent the invasive bio- markers recommended here that should be tested further. Here, *COL11A1* being involved in the ECM processes, *INHBA* involved in immunosuppression, and

CLDN1 having both positive and negative implications with CRC. The study workflow is summarised in Figure 4.2.

These biomarkers could be used as less invasive alternative tool reducing the dependency on invasive procedures, increasing potential for early diagnosis and therefore increasing CRC patient survival rate.

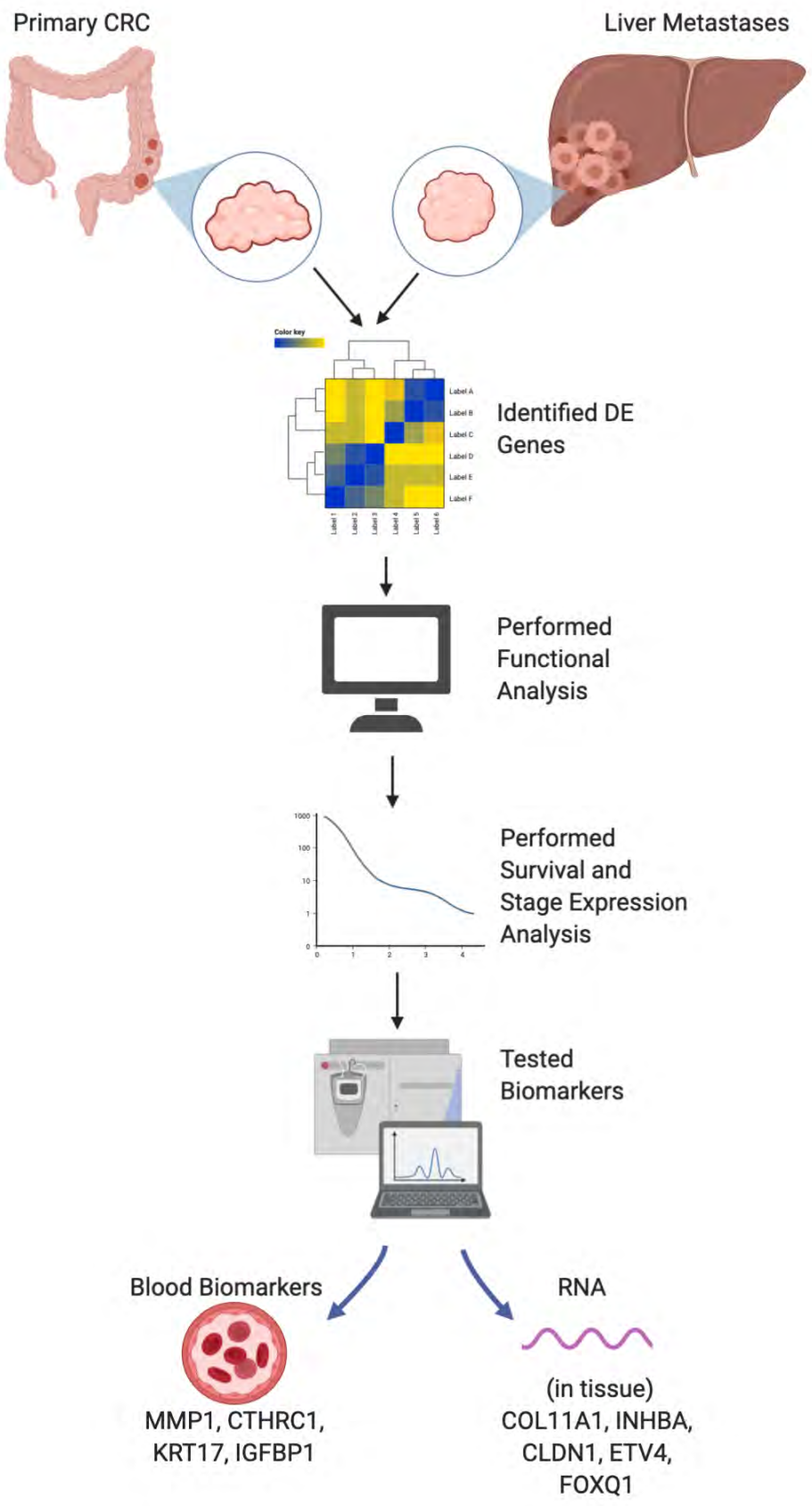


Figure 4.2: Diagram showing the analysis workflow used in this study.

4.6 Limitations of the Study and Future Recommendations

The present study has several limitations. The first limitation is regarding the GSE50760 dataset from Kim *et al.* (2014*b*) as it had no healthy liver controls, but contained liver metastases samples. This meant genes normally expressed in the liver could not be identified and excluded from the analyses. Secondly, patient data was unavailable at the time of writing due to confidentiality. This meant that patient history as well as patient medication was not known and it could be that the patients were undergoing a form of chemotherapy or drug therapy when samples were extracted, which could have lead to the expression of certain genes and pathways related to drug metabolism or adverse effects of the drug.

Additionally, identifying liver metastases genes as metastasis prognostic biomarkers represented a challenge due to the lack of databases on specific CRC metastasis genes. Furthermore there is little literature focused on liver metastases of CRC. Moreover, the CoReCG database used for literature validation presented a limitation as it only mined data up to 2015, lacking data for more recent CRC studies.

In this study, the selection of genes for further biomarker analysis was determined through significance (p-value) and log_2FC only. In future studies, a protein-protein interaction clustering method is recommended. Clustering the most significant interacting proteins of the DE genes could provide more insight into the development of CRC as well as make use of a more integrated bioinformatics approach. From there, the genes coding for the clustered proteins could be investigated further, alongside the significant DE genes. Additionally, further investigation into the co-expression analysis is recommended as only a single hub gene was extracted in the present study. The large amount of gene modules should be investigated to discover potential gene batches that are co- expressed in CRC which can then be used in understanding CRC progression or tested as a biomarker panel. This can be addressed by using the commercially available QIAGEN Ingenuity Pathway Analysis tool which has the capability to provide in-depth and extensive functional analysis results.

Future recommendations of a study of this nature involve using more than one CRC dataset, as well as more datasets consisting of CRC liver metastases. Additionally, the identification of the lncRNA *CASC19* provides an interesting point of departure into a lncRNA or miRNA study. As described, dysregulation

of lncRNAs are involved in various cancers, and miRNAs have been identified as potential biomarkers (Liang *et al.*, 2018; Yang *et al.*, 2018).

Another recommendation would be to use more programs in QC, DE and functional analysis. For example, using RSeQC can provide greater insight into the quality of the RNA-Seq files, while using another DE program such as edgeR could yield different results.

Additionally, a future study is recommended to test these biomarkers in a wet lab setting. The biomarkers were identified *in silico*, however to test their clinical significance an *in vitro* study is required and could provide the foundation necessary to establish the genes as a standard biomarker set for CRC diagnosis and prognosis.

Appendices

Appendix A

Additional DE and Co-Expression Results

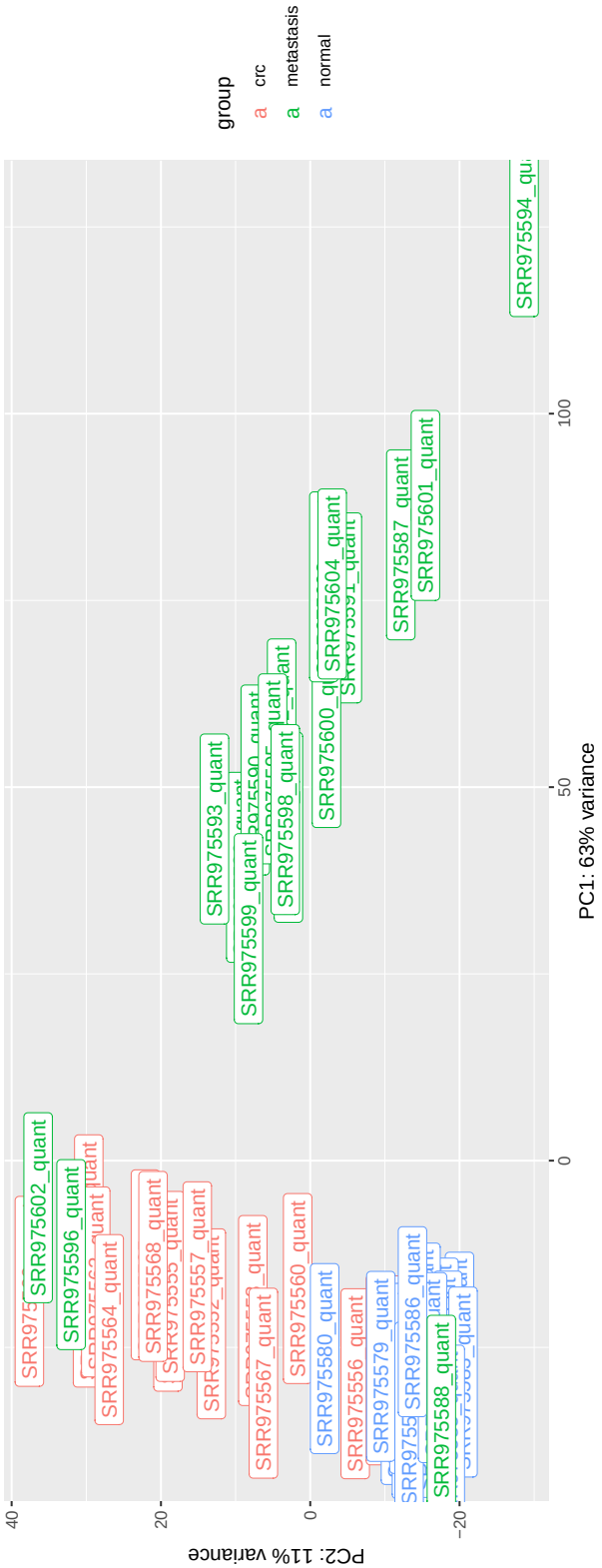


Figure A.1: Figure showing the PCA plot using sample names instead of dots, from original Figure 3.4.

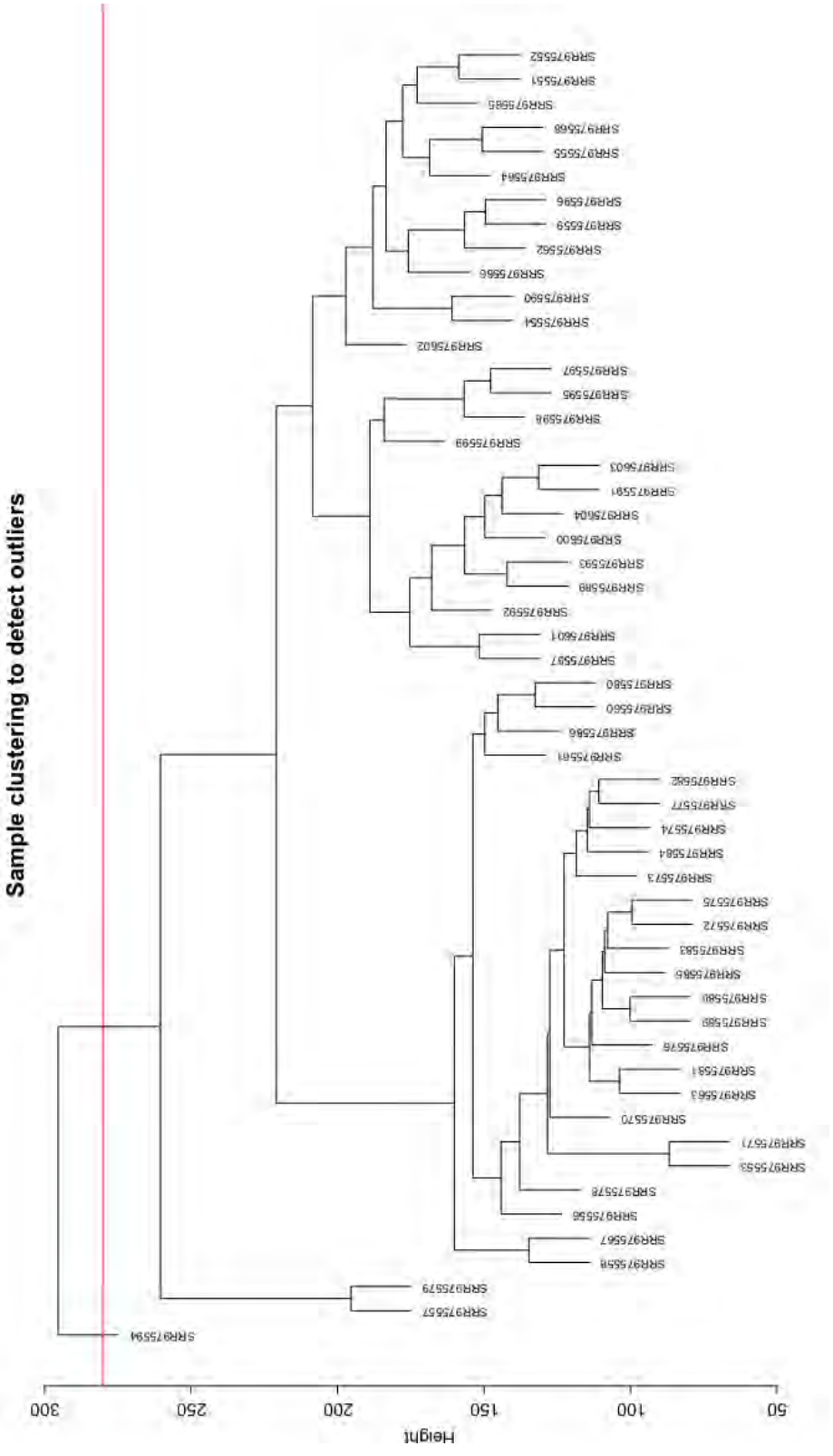
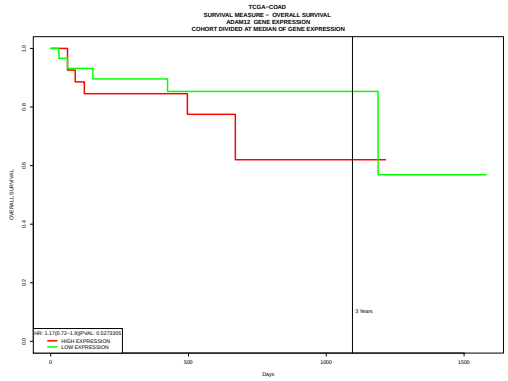


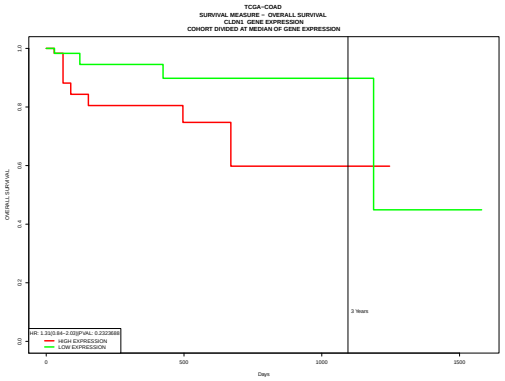
Figure A.2: Figure showing the cluster dendrogram with the red line indicating the cut-off point for outliers, leading to Figure 3.11. Here, sample SRR75594 was removed from analysis.

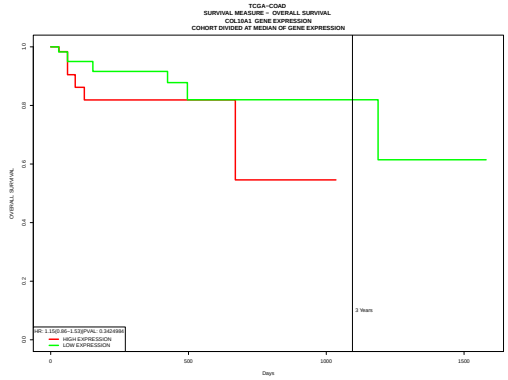
Appendix B

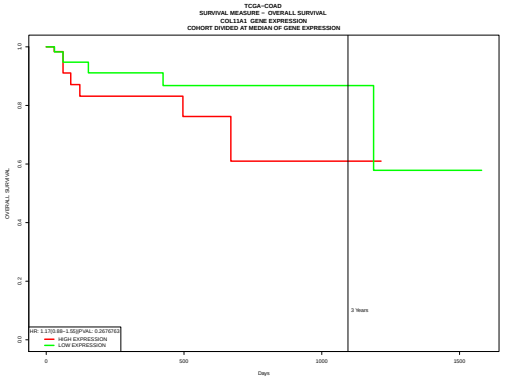
Individual Survival Analysis Plots

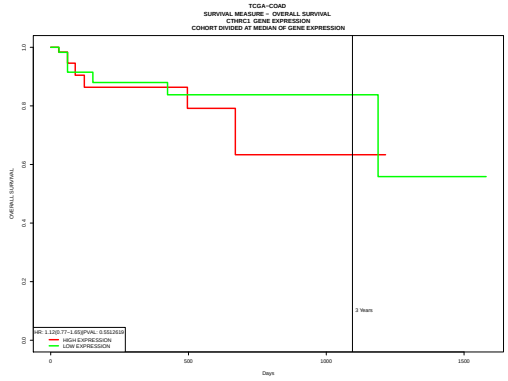
This Appendix includes the survival analysis plots for each individual gene from the Tables 3.3, 3.4 & 3.5.

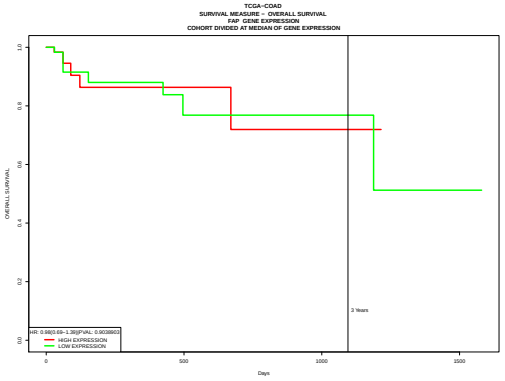

TCGA−COAD
SURVIVAL MEASURE − OVERALL SURVIVAL
ADAM12 GENE EXPRESSION
COHORT DIVIDED AT MEDIAN OF GENE EXPRESSION
OVERALL SURVIVAL
Days
3 Years
HIGH EXPRESSION
LOW EXPRESSION

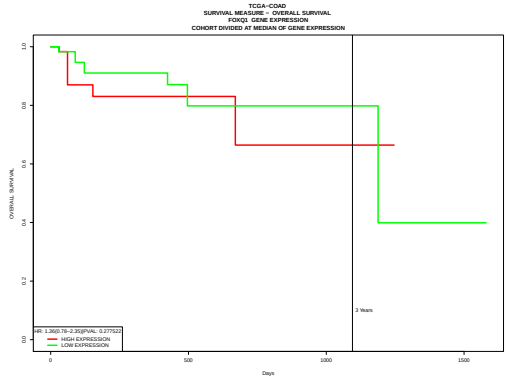

TCGA−COAD
SURVIVAL MEASURE − OVERALL SURVIVAL
CLDN1 GENE EXPRESSION
COHORT DIVIDED AT MEDIAN OF GENE EXPRESSION
OVERALL SURVIVAL
Days
3 Years
HIGH EXPRESSION
LOW EXPRESSION

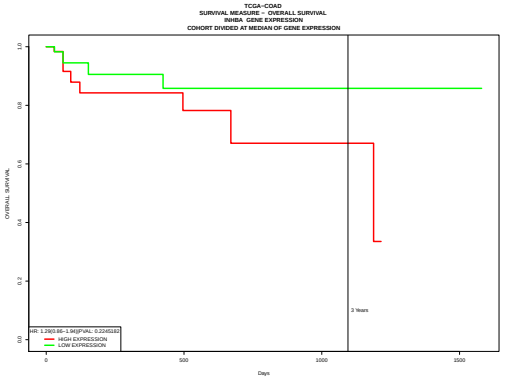

TCGA−COAD
SURVIVAL MEASURE − OVERALL SURVIVAL
COL10A1 GENE EXPRESSION
COHORT DIVIDED AT MEDIAN OF GENE EXPRESSION
OVERALL SURVIVAL
Days
3 Years
HIGH EXPRESSION
LOW EXPRESSION

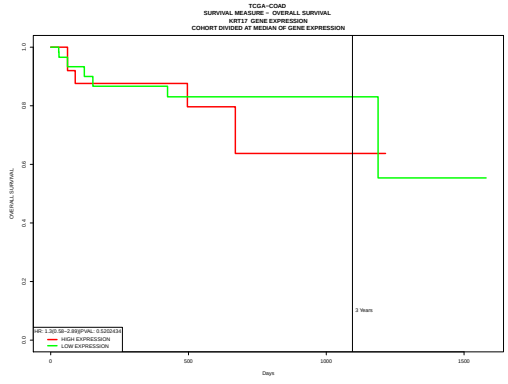

TCGA−COAD
SURVIVAL MEASURE − OVERALL SURVIVAL
COL11A1 GENE EXPRESSION
COHORT DIVIDED AT MEDIAN OF GENE EXPRESSION
OVERALL SURVIVAL
Days
3 Years
HIGH EXPRESSION
LOW EXPRESSION

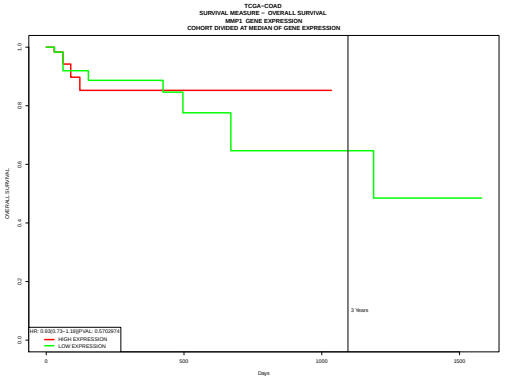

TCGA−COAD
SURVIVAL MEASURE − OVERALL SURVIVAL
CTHRC1 GENE EXPRESSION
COHORT DIVIDED AT MEDIAN OF GENE EXPRESSION
OVERALL SURVIVAL
Days
3 Years
HIGH EXPRESSION
LOW EXPRESSION


TCGA−COAD
SURVIVAL MEASURE − OVERALL SURVIVAL
FAP GENE EXPRESSION
COHORT DIVIDED AT MEDIAN OF GENE EXPRESSION
OVERALL SURVIVAL
Days
3 Years
HIGH EXPRESSION
LOW EXPRESSION


TCGA−COAD
SURVIVAL MEASURE − OVERALL SURVIVAL
FOXQ1 GENE EXPRESSION
COHORT DIVIDED AT MEDIAN OF GENE EXPRESSION
OVERALL SURVIVAL
Days
3 Years
HIGH EXPRESSION
LOW EXPRESSION


TCGA−COAD
SURVIVAL MEASURE − OVERALL SURVIVAL
INHBA GENE EXPRESSION
COHORT DIVIDED AT MEDIAN OF GENE EXPRESSION
OVERALL SURVIVAL
Days
3 Years
HIGH EXPRESSION
LOW EXPRESSION


TCGA−COAD
SURVIVAL MEASURE − OVERALL SURVIVAL
KRT17 GENE EXPRESSION
COHORT DIVIDED AT MEDIAN OF GENE EXPRESSION
OVERALL SURVIVAL
Days
3 Years
HIGH EXPRESSION
LOW EXPRESSION


TCGA−COAD
SURVIVAL MEASURE − OVERALL SURVIVAL
MMP1 GENE EXPRESSION
COHORT DIVIDED AT MEDIAN OF GENE EXPRESSION
OVERALL SURVIVAL
Days
3 Years
HIGH EXPRESSION
LOW EXPRESSION

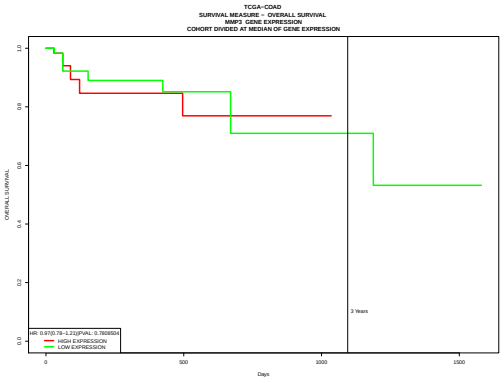
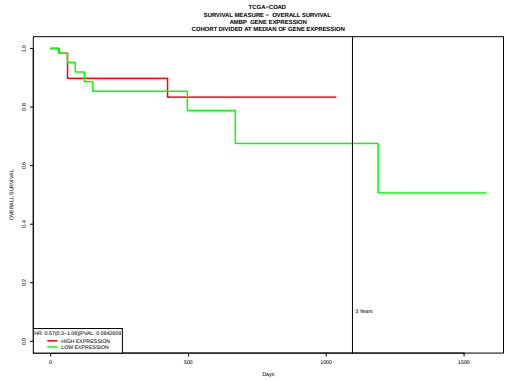
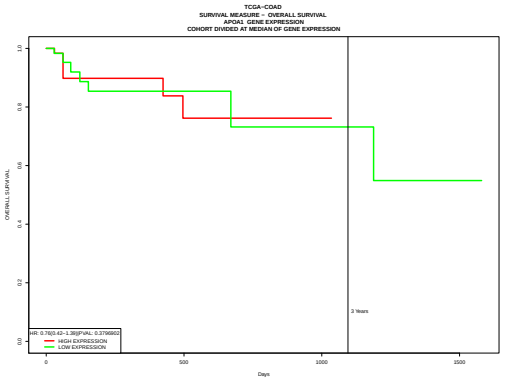
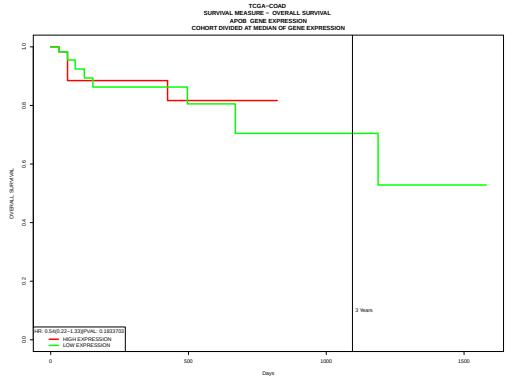
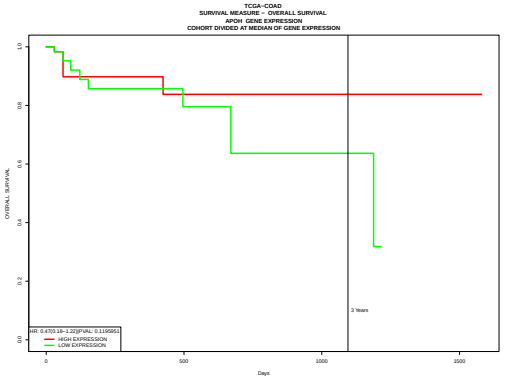


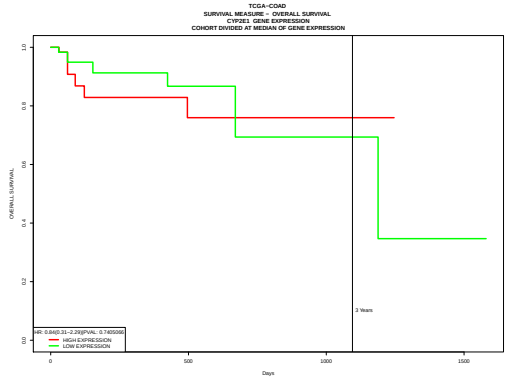
Figure B.1: Individual survival plots for Normal vs CRC Significant genes in Table 3.3. Red line indicates high expression. Green line indicates low expression.

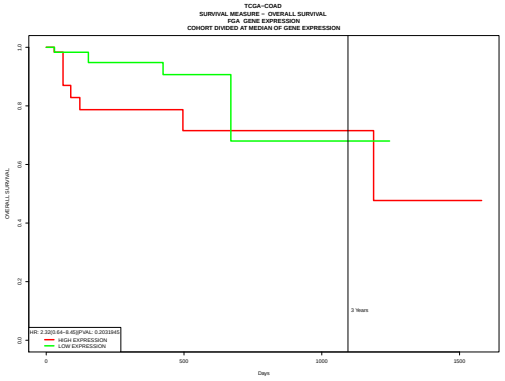

TCGA-COAD
SURVIVAL MEASURE - OVERALL SURVIVAL
AMBP GENE EXPRESSION
COHORT DIVIDED AT MEDIAN OF GENE EXPRESSION
OVERALL SURVIVAL
Days
3 Years
HIGH EXPRESSION
LOW EXPRESSION

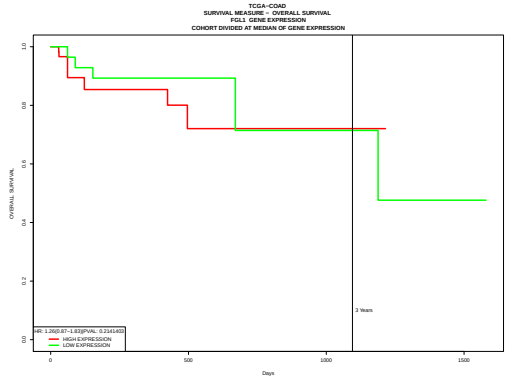

TCGA-COAD
SURVIVAL MEASURE - OVERALL SURVIVAL
APOA1 GENE EXPRESSION
COHORT DIVIDED AT MEDIAN OF GENE EXPRESSION
OVERALL SURVIVAL
Days
3 Years
HIGH EXPRESSION
LOW EXPRESSION

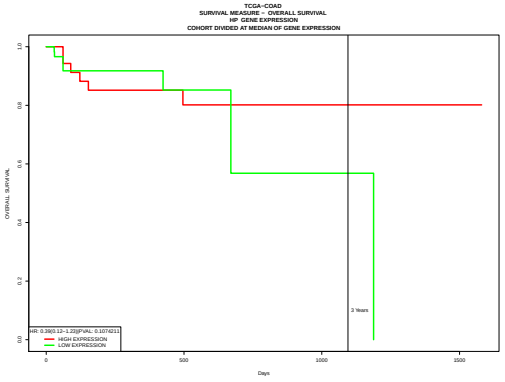

TCGA-COAD
SURVIVAL MEASURE - OVERALL SURVIVAL
APOB GENE EXPRESSION
COHORT DIVIDED AT MEDIAN OF GENE EXPRESSION
OVERALL SURVIVAL
Days
3 Years
HIGH EXPRESSION
LOW EXPRESSION

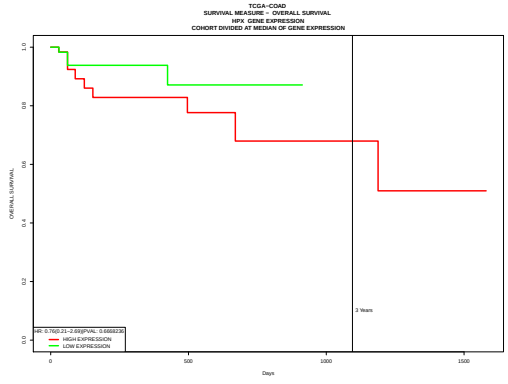

TCGA-COAD
SURVIVAL MEASURE - OVERALL SURVIVAL
APOH GENE EXPRESSION
COHORT DIVIDED AT MEDIAN OF GENE EXPRESSION
OVERALL SURVIVAL
Days
3 Years
HIGH EXPRESSION
LOW EXPRESSION

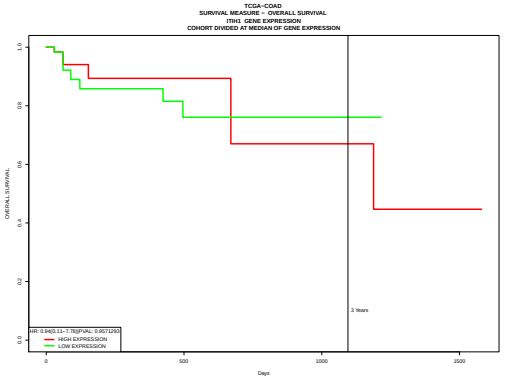

TCGA-COAD
SURVIVAL MEASURE - OVERALL SURVIVAL
CYP2E1 GENE EXPRESSION
COHORT DIVIDED AT MEDIAN OF GENE EXPRESSION
OVERALL SURVIVAL
Days
3 Years
HIGH EXPRESSION
LOW EXPRESSION


TCGA-COAD
SURVIVAL MEASURE - OVERALL SURVIVAL
FGA GENE EXPRESSION
COHORT DIVIDED AT MEDIAN OF GENE EXPRESSION
OVERALL SURVIVAL
Days
3 Years
HIGH EXPRESSION
LOW EXPRESSION


TCGA-COAD
SURVIVAL MEASURE - OVERALL SURVIVAL
FGL1 GENE EXPRESSION
COHORT DIVIDED AT MEDIAN OF GENE EXPRESSION
OVERALL SURVIVAL
Days
3 Years
HIGH EXPRESSION
LOW EXPRESSION


TCGA-COAD
SURVIVAL MEASURE - OVERALL SURVIVAL
HP GENE EXPRESSION
COHORT DIVIDED AT MEDIAN OF GENE EXPRESSION
OVERALL SURVIVAL
Days
3 Years
HIGH EXPRESSION
LOW EXPRESSION


TCGA-COAD
SURVIVAL MEASURE - OVERALL SURVIVAL
HPX GENE EXPRESSION
COHORT DIVIDED AT MEDIAN OF GENE EXPRESSION
OVERALL SURVIVAL
Days
3 Years
HIGH EXPRESSION
LOW EXPRESSION


TCGA-COAD
SURVIVAL MEASURE - OVERALL SURVIVAL
ITIH1 GENE EXPRESSION
COHORT DIVIDED AT MEDIAN OF GENE EXPRESSION
OVERALL SURVIVAL
Days
3 Years
HIGH EXPRESSION
LOW EXPRESSION

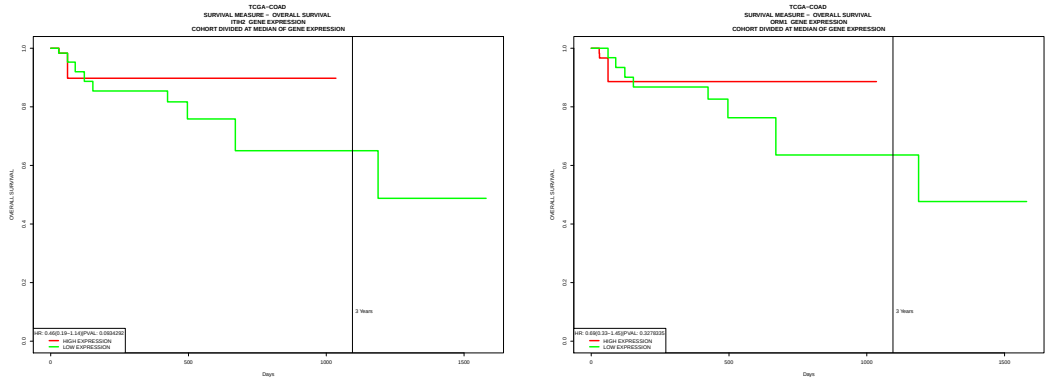
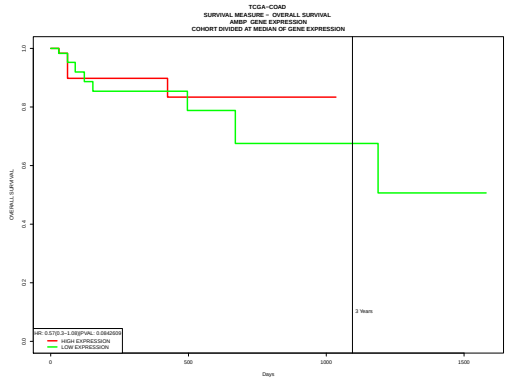
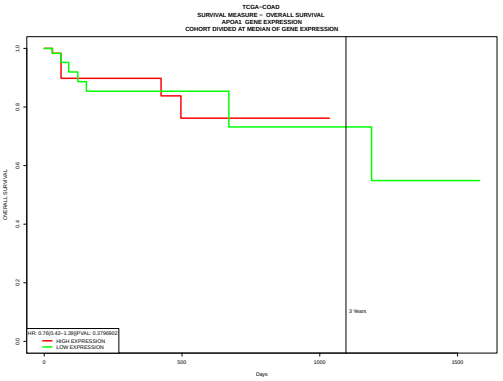
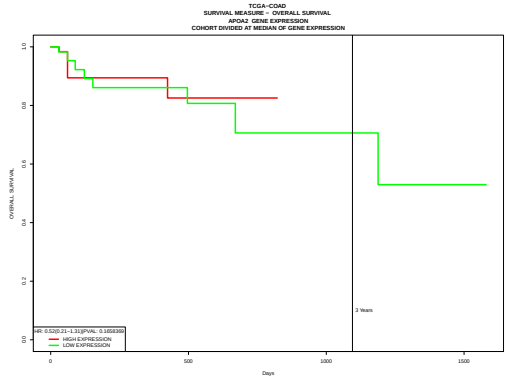
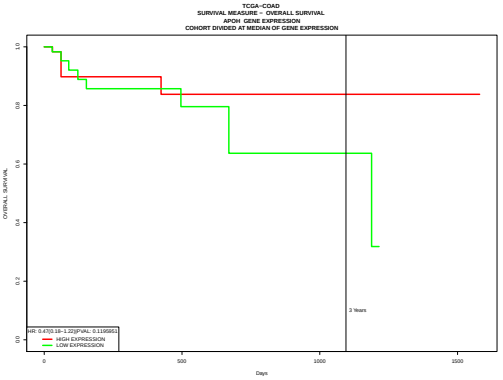


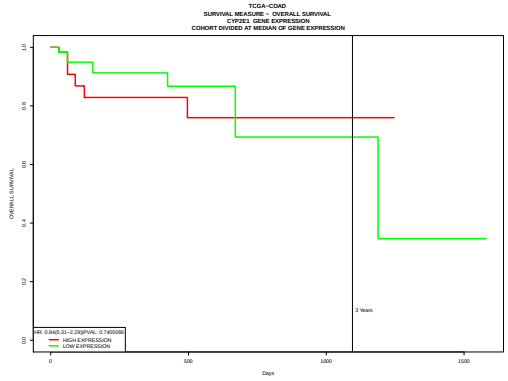
Figure B.2: Individual survival plots for CRC vs Metastasis Significant genes in Table 3.4. Red line indicates high expression. Green line indicates low expression.

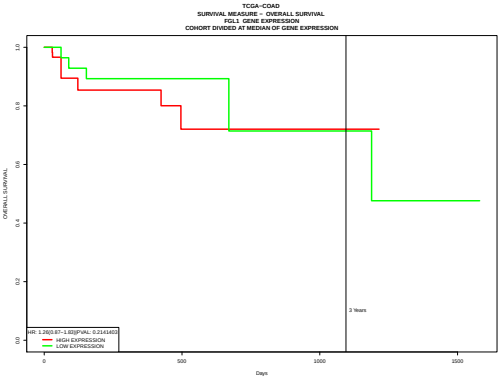

TCGA−COAD
SURVIVAL MEASURE − OVERALL SURVIVAL
AMBP GENE EXPRESSION
COHORT DIVIDED AT MEDIAN OF GENE EXPRESSION
OVERALL SURVIVAL
Days
3 Years
HIGH EXPRESSION
LOW EXPRESSION

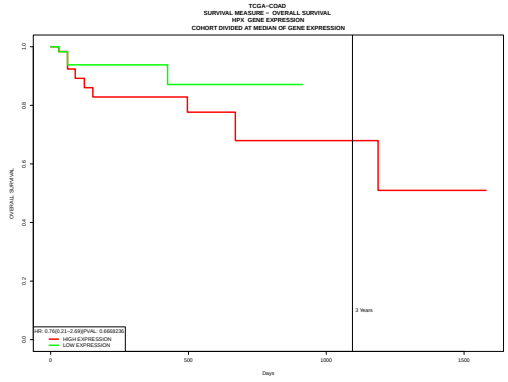

TCGA−COAD
SURVIVAL MEASURE − OVERALL SURVIVAL
APOA1 GENE EXPRESSION
COHORT DIVIDED AT MEDIAN OF GENE EXPRESSION
OVERALL SURVIVAL
Days
3 Years
HIGH EXPRESSION
LOW EXPRESSION

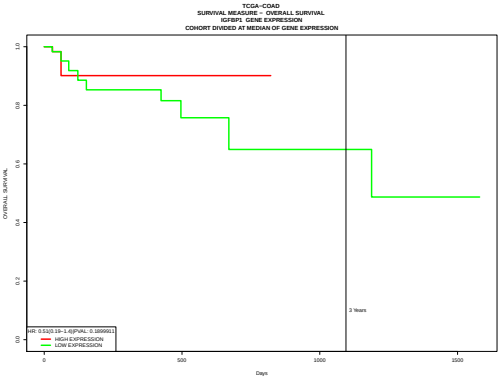

TCGA−COAD
SURVIVAL MEASURE − OVERALL SURVIVAL
APOA2 GENE EXPRESSION
COHORT DIVIDED AT MEDIAN OF GENE EXPRESSION
OVERALL SURVIVAL
Days
3 Years
HIGH EXPRESSION
LOW EXPRESSION

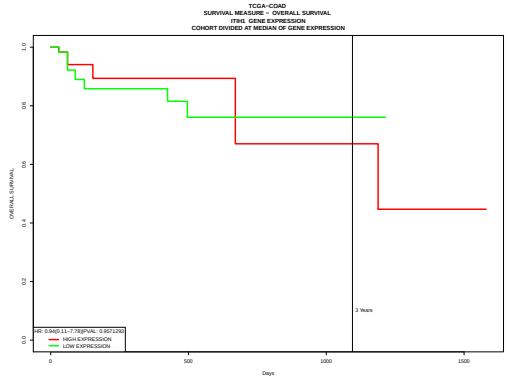

TCGA−COAD
SURVIVAL MEASURE − OVERALL SURVIVAL
APOH GENE EXPRESSION
COHORT DIVIDED AT MEDIAN OF GENE EXPRESSION
OVERALL SURVIVAL
Days
3 Years
HIGH EXPRESSION
LOW EXPRESSION

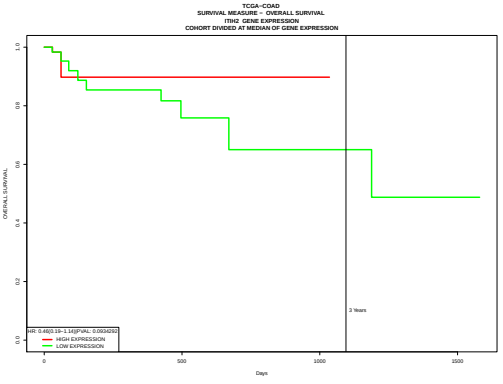

TCGA−COAD
SURVIVAL MEASURE − OVERALL SURVIVAL
CYP2E1 GENE EXPRESSION
COHORT DIVIDED AT MEDIAN OF GENE EXPRESSION
OVERALL SURVIVAL
Days
3 Years
HIGH EXPRESSION
LOW EXPRESSION


TCGA−COAD
SURVIVAL MEASURE − OVERALL SURVIVAL
FGL1 GENE EXPRESSION
COHORT DIVIDED AT MEDIAN OF GENE EXPRESSION
OVERALL SURVIVAL
Days
3 Years
HIGH EXPRESSION
LOW EXPRESSION


TCGA−COAD
SURVIVAL MEASURE − OVERALL SURVIVAL
HPX GENE EXPRESSION
COHORT DIVIDED AT MEDIAN OF GENE EXPRESSION
OVERALL SURVIVAL
Days
3 Years
HIGH EXPRESSION
LOW EXPRESSION


TCGA−COAD
SURVIVAL MEASURE − OVERALL SURVIVAL
IGFBP1 GENE EXPRESSION
COHORT DIVIDED AT MEDIAN OF GENE EXPRESSION
OVERALL SURVIVAL
Days
3 Years
HIGH EXPRESSION
LOW EXPRESSION


TCGA−COAD
SURVIVAL MEASURE − OVERALL SURVIVAL
ITIH1 GENE EXPRESSION
COHORT DIVIDED AT MEDIAN OF GENE EXPRESSION
OVERALL SURVIVAL
Days
3 Years
HIGH EXPRESSION
LOW EXPRESSION


TCGA−COAD
SURVIVAL MEASURE − OVERALL SURVIVAL
ITIH2 GENE EXPRESSION
COHORT DIVIDED AT MEDIAN OF GENE EXPRESSION
OVERALL SURVIVAL
Days
3 Years
HIGH EXPRESSION
LOW EXPRESSION

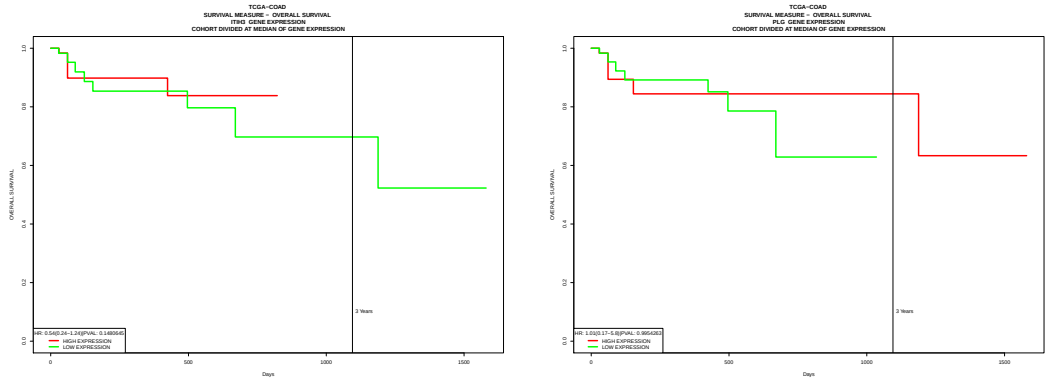
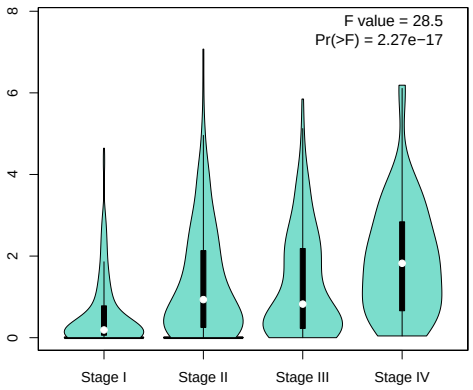


Figure B.3: Individual survival plots for Normal vs Metastasis Significant genes in Table 3.5. Red line indicates high expression. Green line indicates low expression.

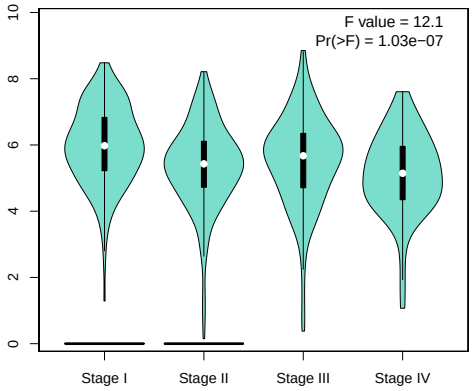
Appendix C

Individual Stage Expression Plots

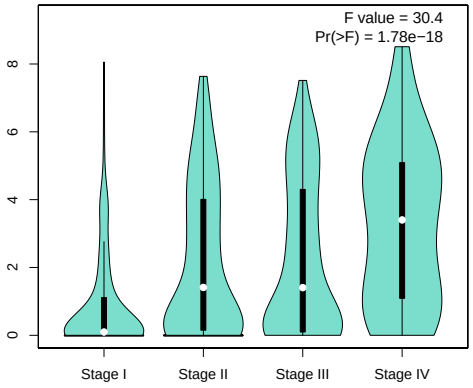
This Appendix includes the stage expression plots for each individual gene from the Tables 3.3, 3.4 & 3.5. Each gene was first mapped to TCGA COAD and READ, and then separately to TCGA LIHC, resulting in six figure groups.



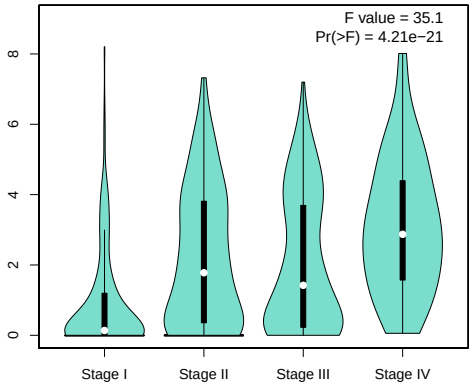
(a) ADAM12



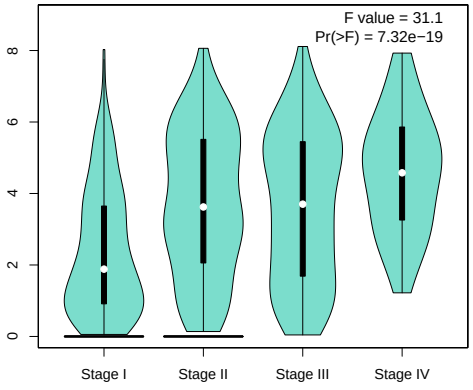
(b) CLDN1



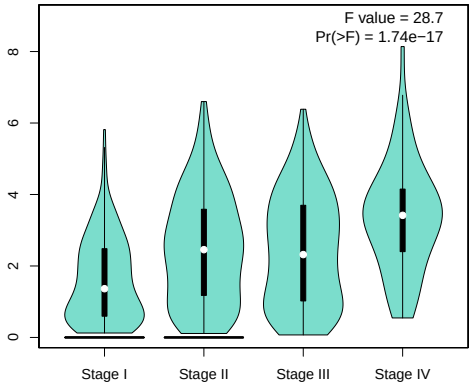
(c) COL10A1



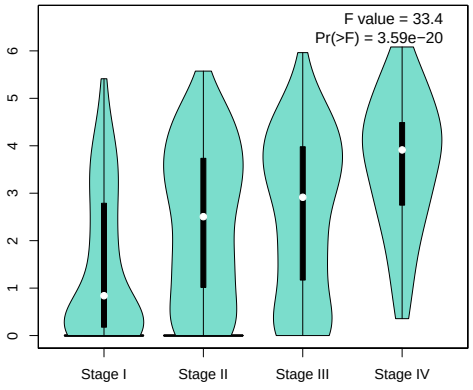
(d) COL11A1



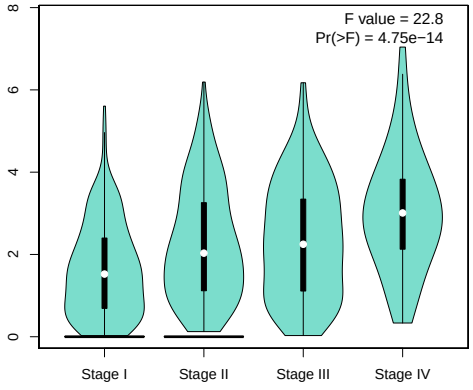
(e) CTHRC1



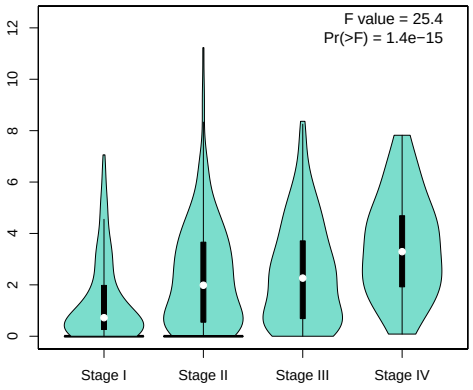
(f) FAP



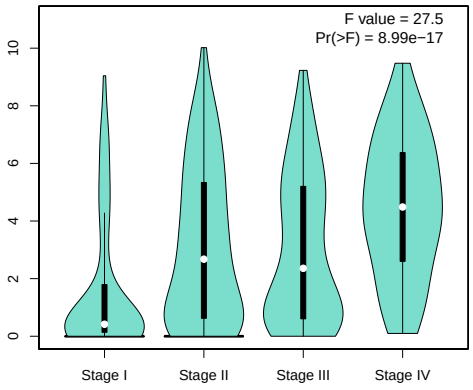
(g) FOXQ1



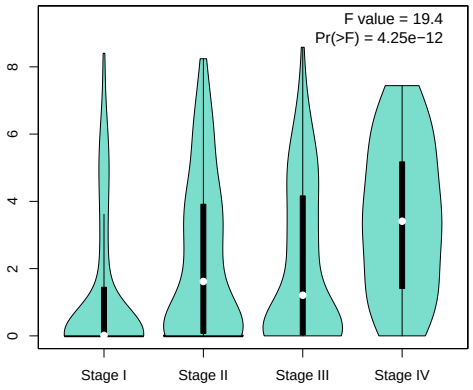
(h) INHBA



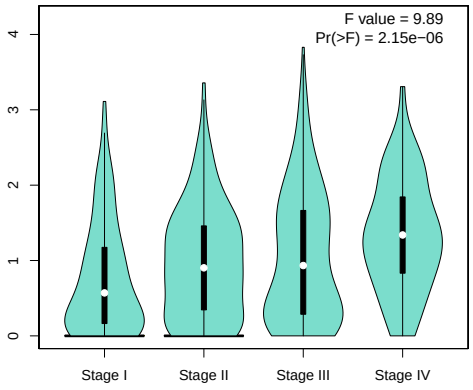
(i) KRT17



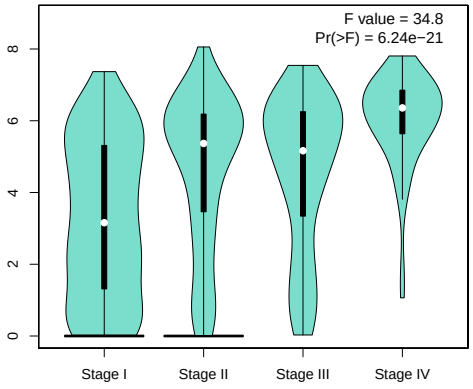
(j) MMP1



(k) MMP3

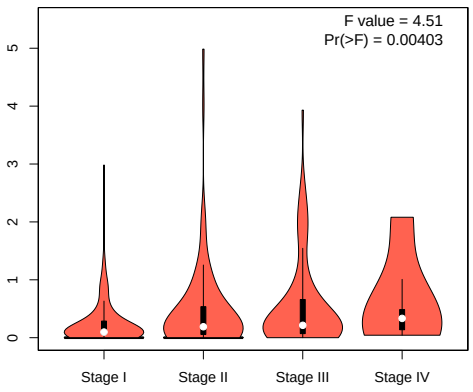


(l) CASC19

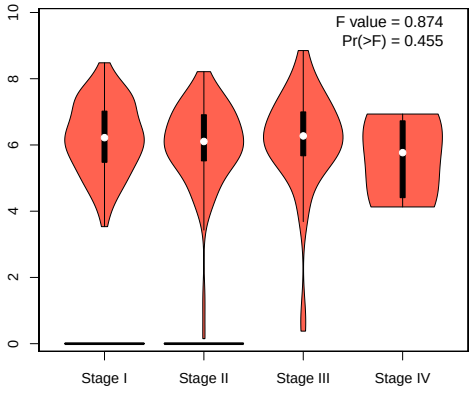


(m) E1AF

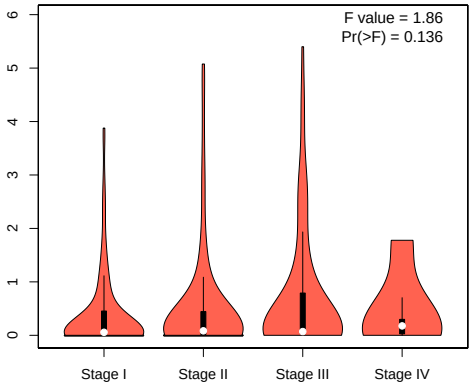
Figure C.1: Individual stage expression plots for Normal vs CRC Significant genes in Table 3.3 using GEPIA2. The genes were mapped against TCGA COAD and READ.



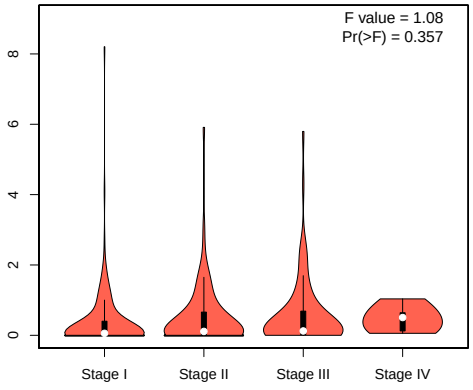
(a) ADAM12



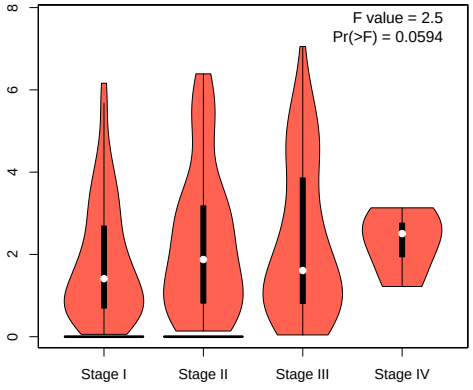
(b) CLDN1



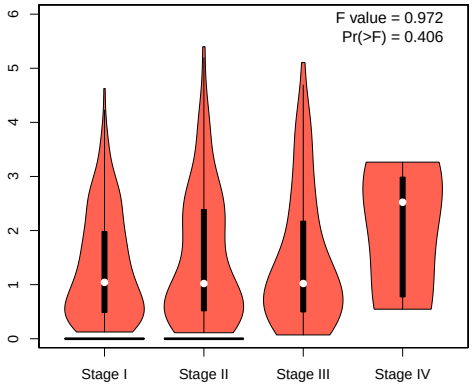
(c) COL10A1



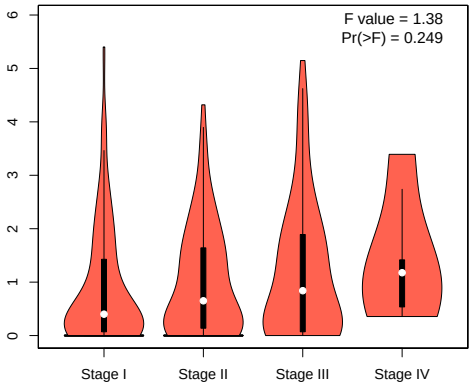
(d) COL11A1



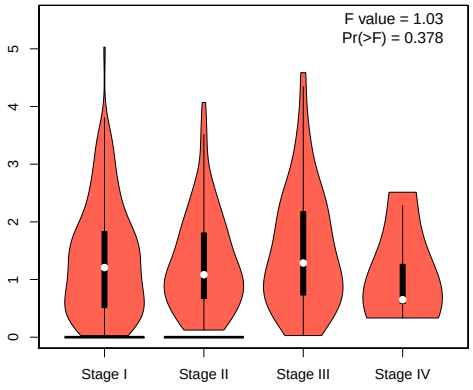
(e) CTHRC1



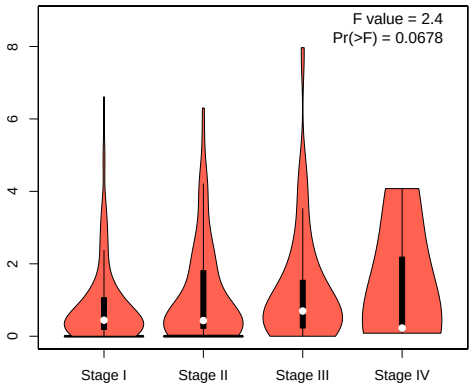
(f) FAP



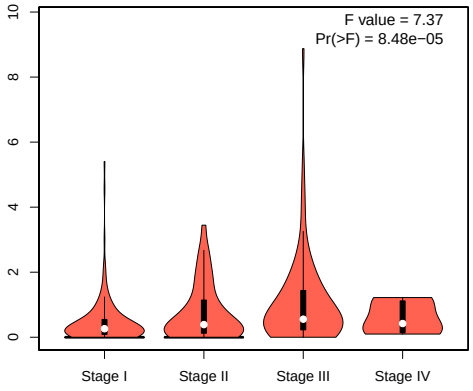
(g) FOXQ1



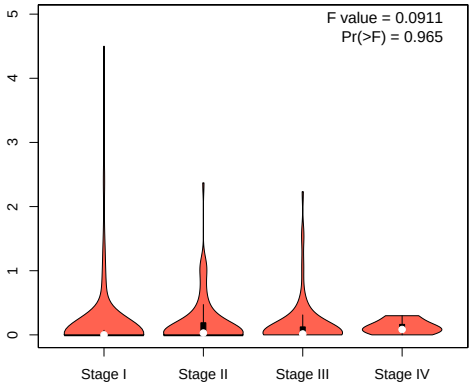
(h) INHBA



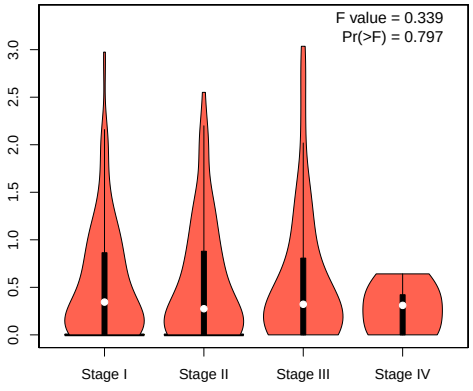
(i) KRT17



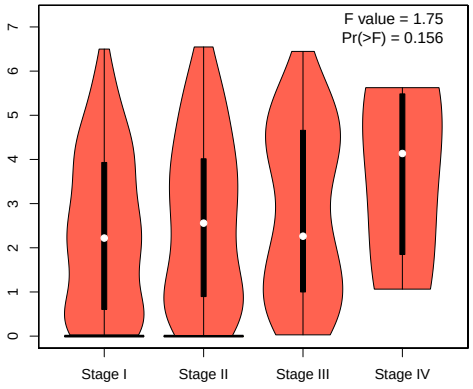
(j) MMP1



(k) MMP3

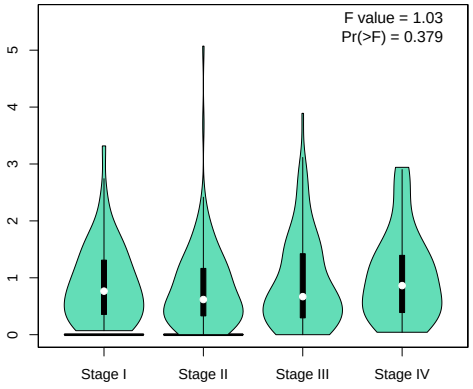


(l) CASC19

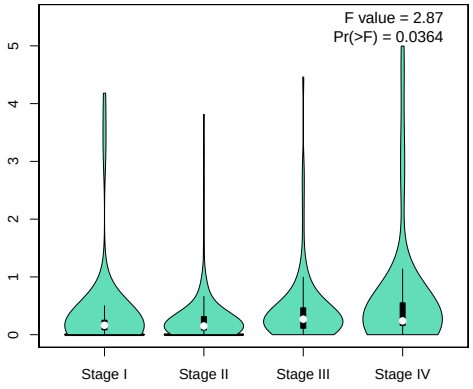


(m) E1AF

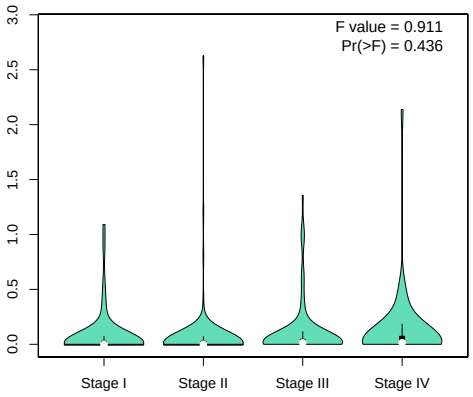
Figure C.2: Individual stage expression plots for Normal vs CRC Significant genes in Table 3.3 using GEPIA2. The genes were mapped against TCGA LIHC to indivate their presence in liver metastases.



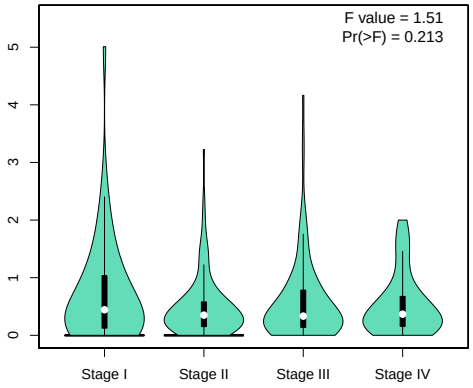
(a) AMBP



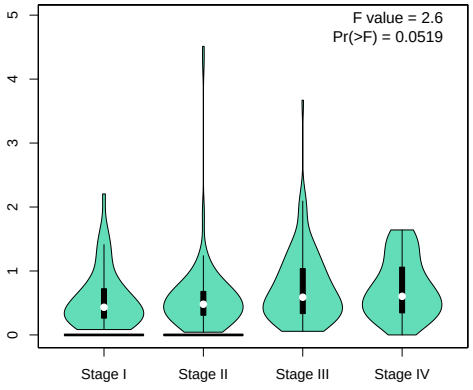
(b) APOA1



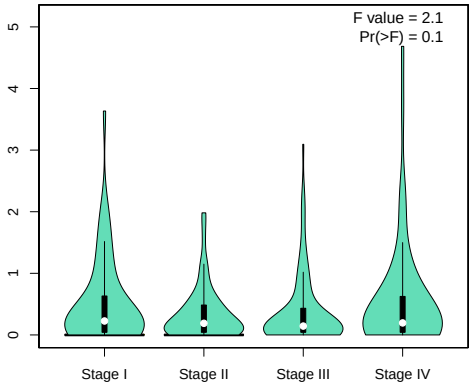
(c) APOB



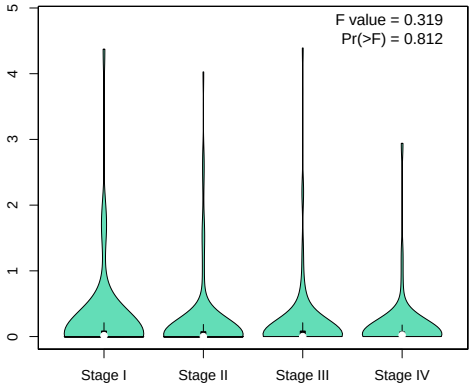
(d) APOH



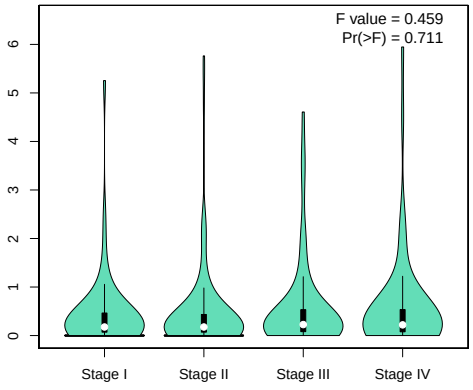
(e) CYP2E1



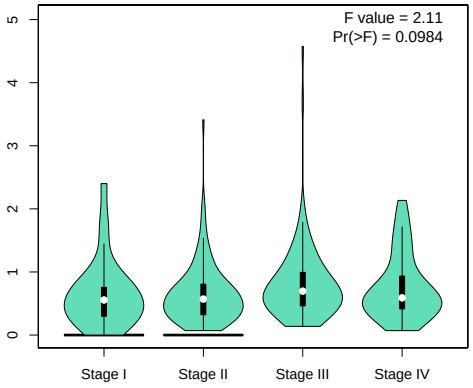
(f) FGA



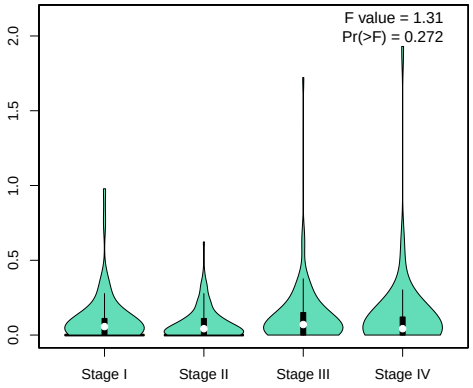
(g) FGL1



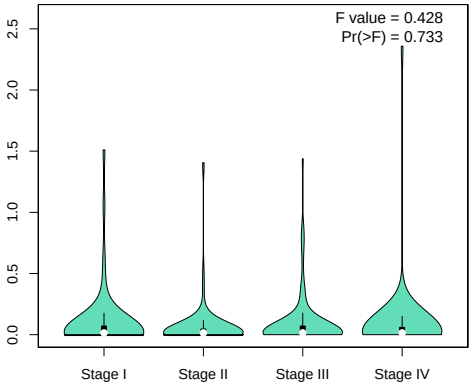
(h) HP



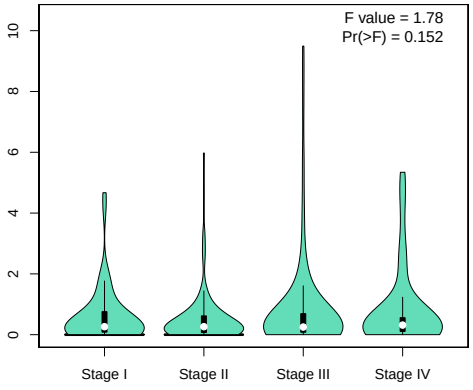
(i) HPX



(j) ITIH1

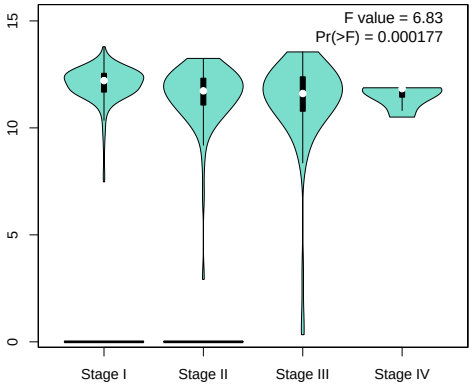


(k) ITIH2

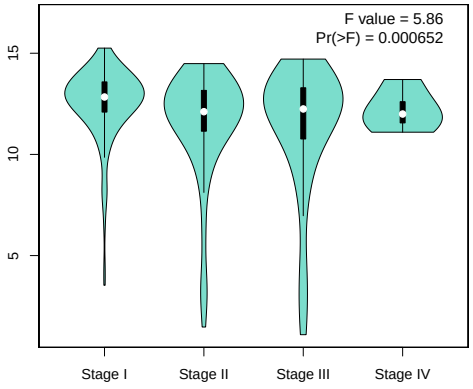


(l) ORM1

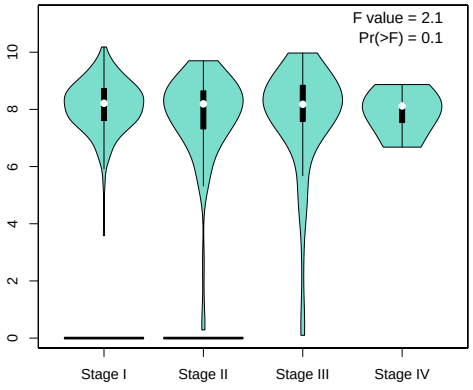
Figure C.3: Individual stage expression plots for CRC vs Metastasis Significant genes in Table 3.4 using GEPIA2. The genes were mapped against COAD and READ.



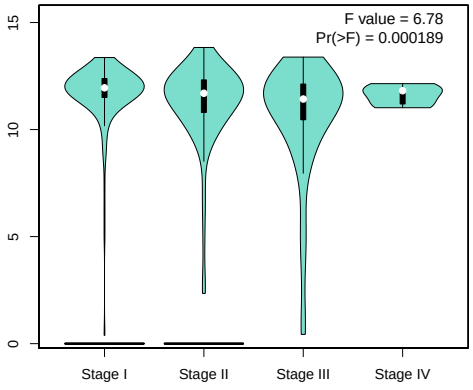
(a) AMBP



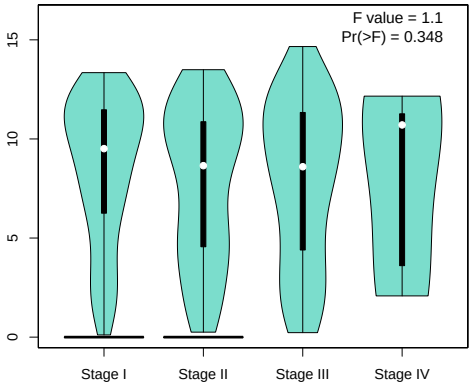
(b) APOA1



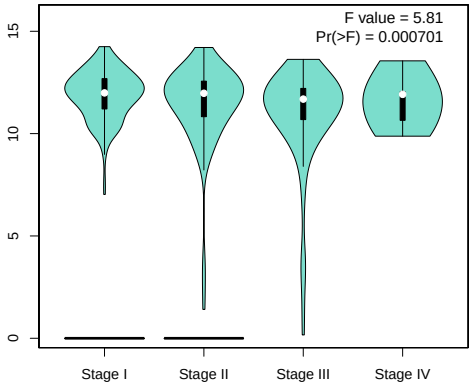
(c) APOB



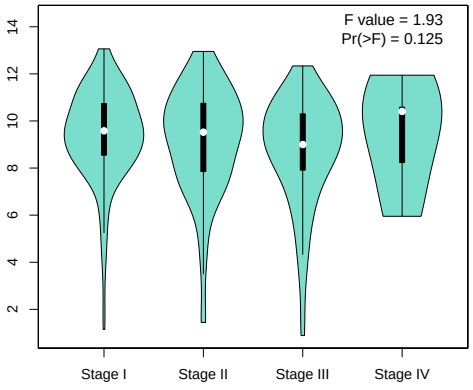
(d) APOH



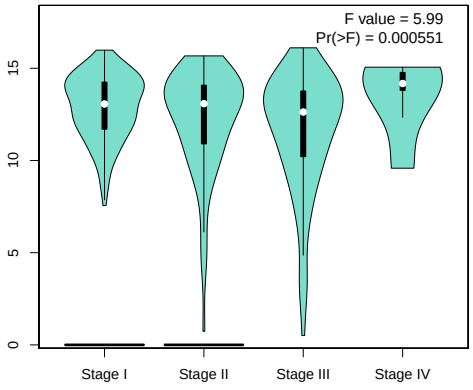
(e) CYP2E1



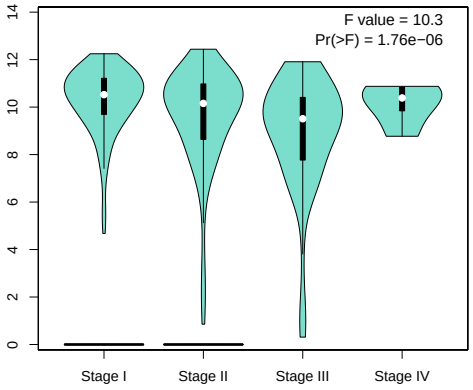
(f) FGA



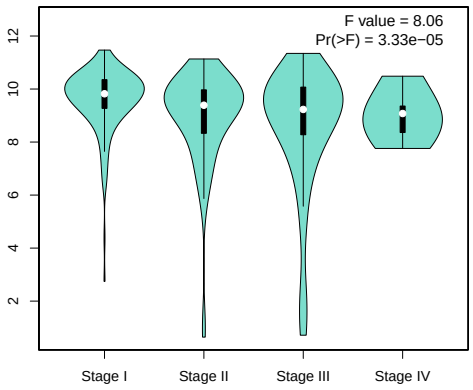
(g) FGL1



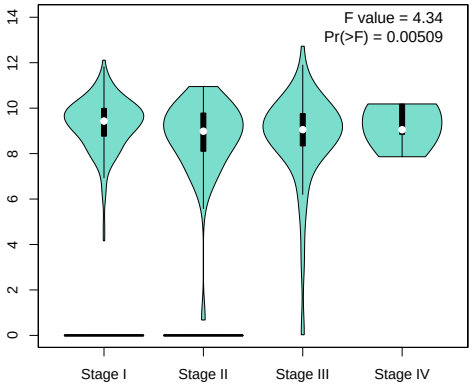
(h) HP



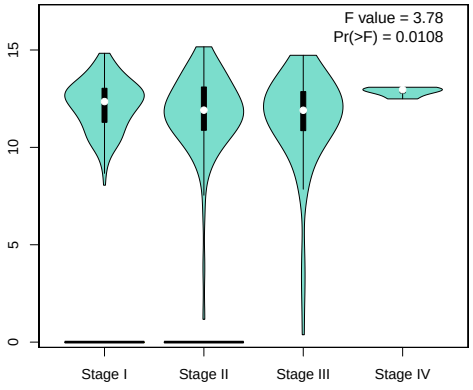
(i) HPX



(j) ITIH1

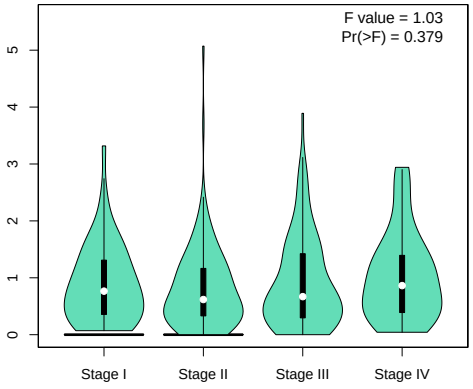


(k) ITIH2

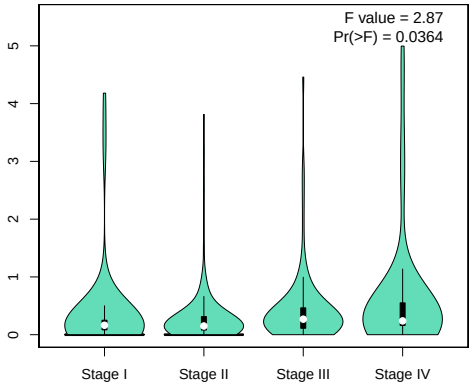


(l) ORM1

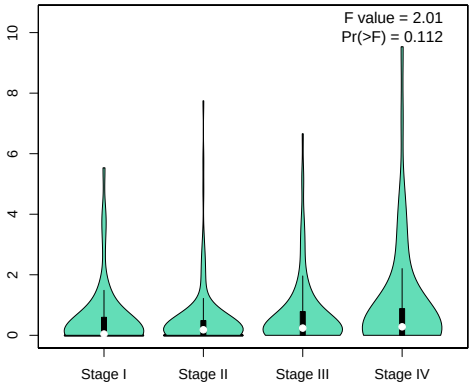
Figure C.4: Individual stage expression plots for CRC vs Metastasis Significant genes in Table 3.4 using GEPIA2. The genes were mapped against LIHC.



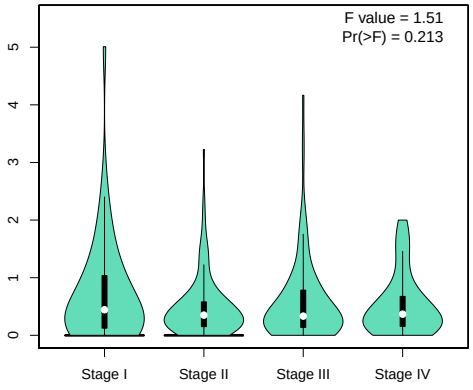
(a) AMBP



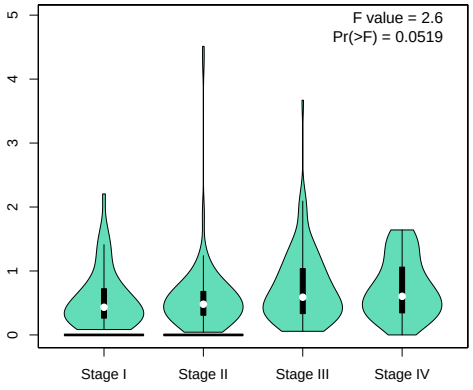
(b) APOA1



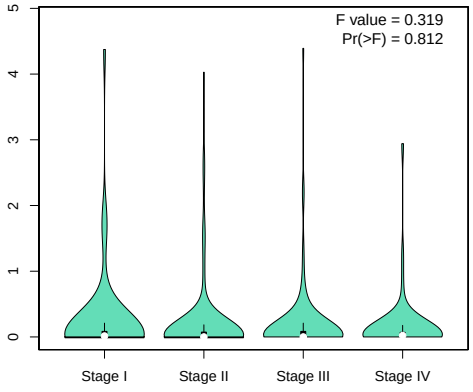
(c) APOA2



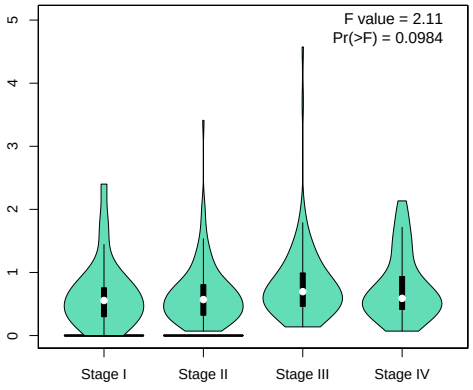
(d) APOH



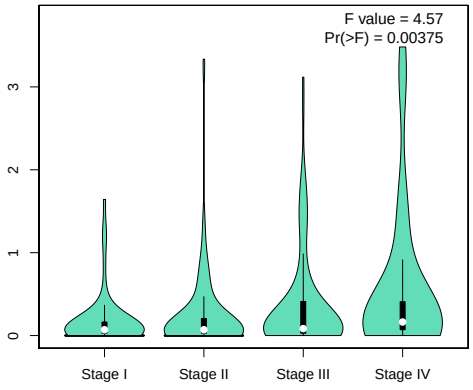
(e) CYP2E1



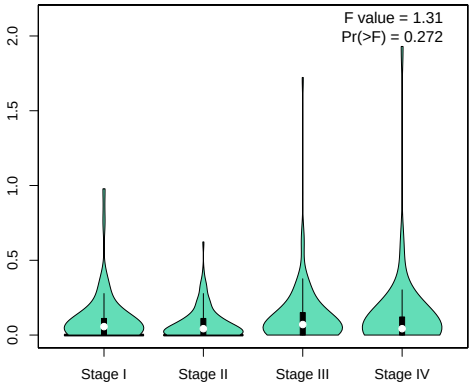
(f) FGL1



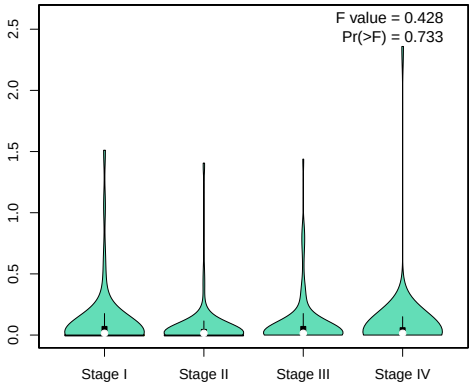
(g) HPX



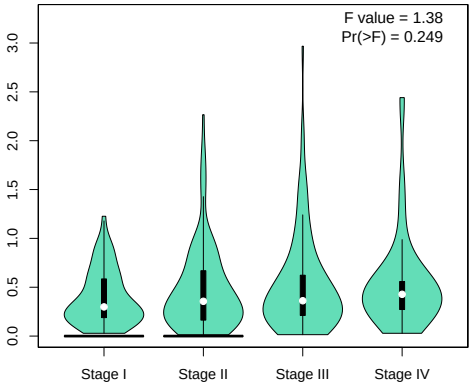
(h) IGFBP1



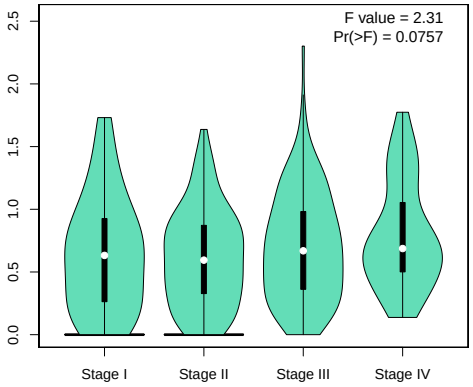
(i) ITIH1



(j) ITIH2

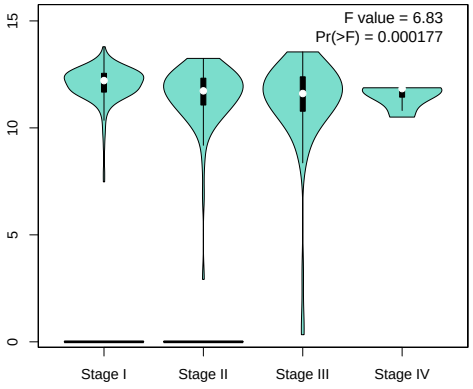


(k) ITIH3

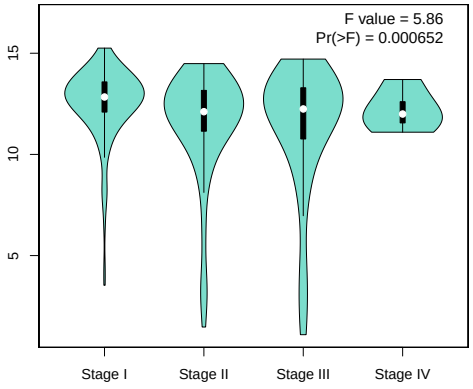


(l) PLG

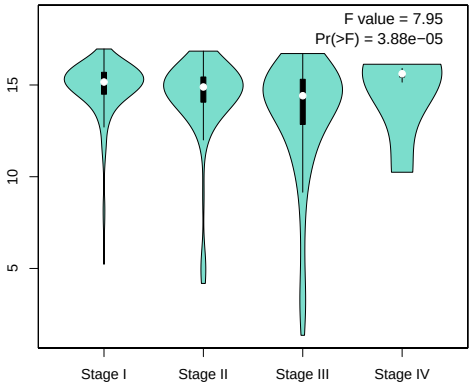
Figure C.5: Individual stage expression plots for Normal vs Metastasis Significant genes in Table 3.5 using GEPIA2. Genes were mapped against TCGA COAD and READ.



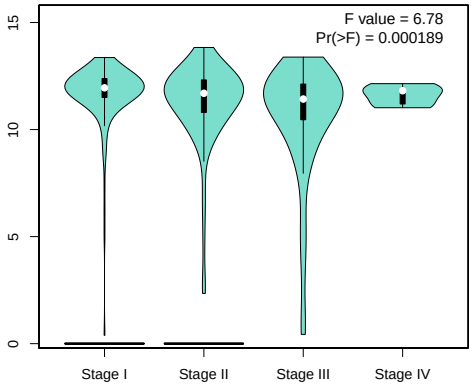
(a) AMBP



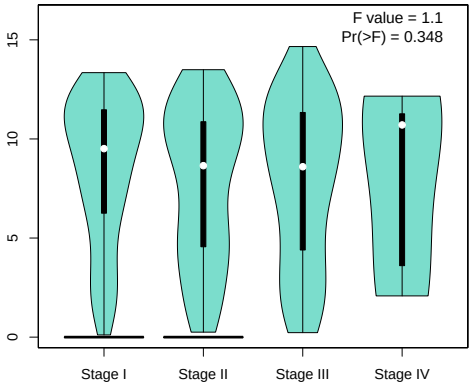
(b) APOA1



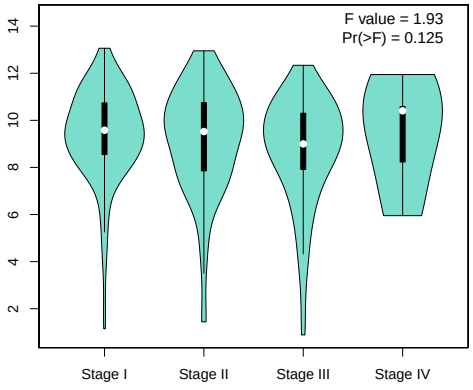
(c) APOA2



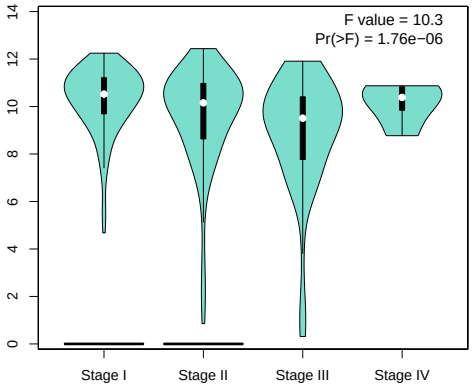
(d) APOH



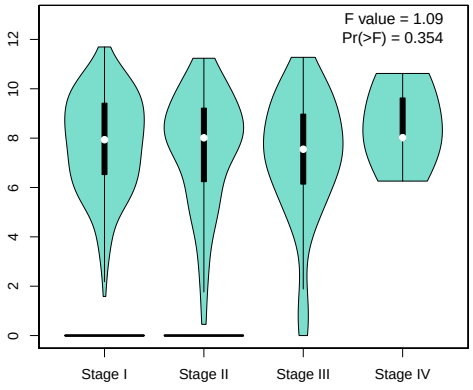
(e) CYP2E1



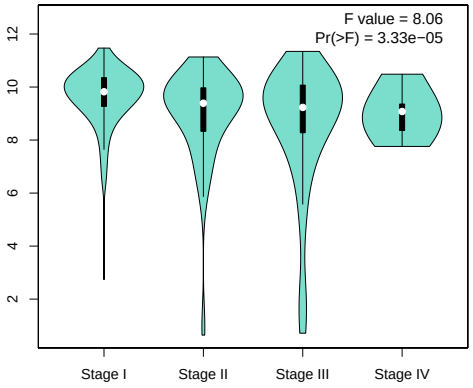
(f) FGL1



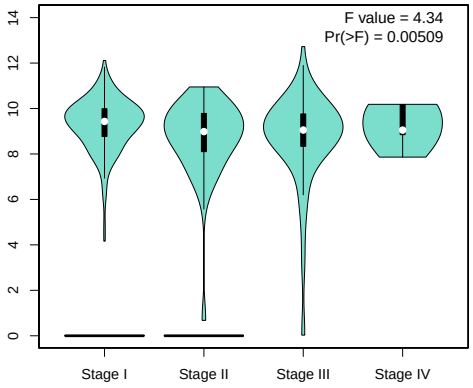
(g) HPX



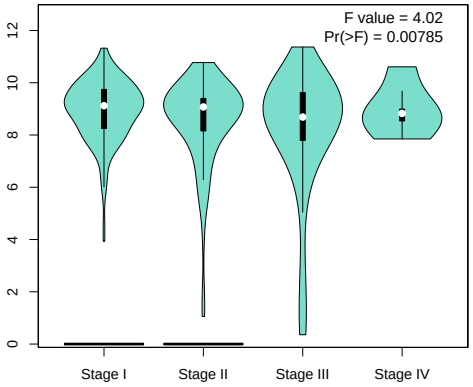
(h) IGFBP1



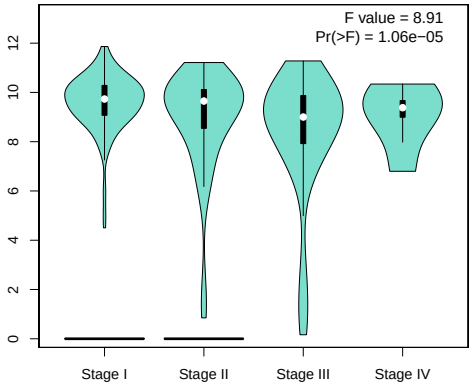
(i) ITIH1



(j) ITIH2



(k) ITIH3



(l) PLG

Figure C.6: Individual stage expression plots for Normal vs Metastasis Significant genes in Table 3.5 using GEPIA2. Genes were mapped against TCGA LIHC.

Appendix D

Additional Validation Results

D.1 CoReCG and Oncomine

This Appendix includes the CoReCG gene descriptions and Oncomine validation statistics for each individual gene from the Tables 3.3, 3.4 & 3.5.

Gene Name	CoReCG Description	Source
COL11A1	Stromal expression associated with malignancy in CRC	(Fischer *et al.*, 2001)
ETV4	Associated with histopathology, proliferation and invasion	(Nosho *et al.*, 2005; Moss *et al.*, 2006)
INHBA	Useful as a predictive marker for CRC prognosis	(Okano *et al.*, 2013)
ADAM12	Not found	N/A
CLDN1	Expression highly unregulated in CRC and promotes tumour progression and metastasis	(Singh *et al.*, 2012)
COL10A1	Not found	N/A
MMP1	Polymorphisms associated with increased risk of CRC	(Liu *et al.*, 2011; Lu *et al.*, 2015)
FAP	Not found	N/A
CTHRC1	High levels are associated with poor clinical outcomes	(Kim *et al.*, 2014*a*)
CASC19	Not found	N/A
MMP3	In sporadic CR tumours, expression is highly specific to tumours transitions	(Sipos *et al.*, 2014)
KRT17	Demonstrated tendency to increased expression in CRC stage progression	(Kim *et al.*, 2012)
FOXQ1	Potential biomarker of metastasis, increased expression associated with migration and invasion	(Abba *et al.*, 2013)

Table D.1: Table showing the primary CRC genes (Table 3.3) validated on CoReCG along with the mined summaries of the genes interaction with CRC and the sources from where the descriptions were mined.

Gene Name	Fold Change	P-Value	Rank	Subset of TCGA
COL11A1	32.796	2.19E-44	3 (in top 1%)	COAD
E1AF/ETV4	8.564	3.71E-41	2 (in top 1%)	READ
INHBA	22.920	7.32E-23	2 (in top 1%)	COMAD
ADAM12	5.319	1.14E-29	76 (in top 1%)	COAD
CLDN1	22.267	8.63E-16	2 (in top 1%)	REMAD
COL10A1	15.755	1.22E-40	13 (in top 1%)	COAD
MMP1	40.543	1.63E-14	119 (in top 1%)	COAD
FAP	15.072	2.04E-13	192 (in top 1%)	COMAD
CTHRC1	6.222	1.66E-29	79 (in top 1%)	COAD
CASC19	-	-	-	-
MMP3	58.870	4.55E-15	84 (in top 1%)	COMAD
KRT17	-	-	-	-
FOXQ1	50.577	3.10E-37	4 (in top 1%)	READ

Table D.2: Table showing the detailed Oncomine validation results for Normal vs CRC.

Gene Name	CoReCG Description	Source
CYP2E1	Association with 96-bp insertion polymorphism and risk of CRC	(Sameer *et al.*, 2011)
APOB	Expressions was lower in patients with CRC	(Zhang *et al.*, 2014)
FGB	Found differentially expressed in CRC tissue	(Zhao *et al.*, 2010)
HP	Upregulated in CRC liver metastases, possible biomarker	(Sun *et al.*, 2012; Chalkias *et al.*, 2011)

Table D.3: Table showing the liver metastases (Table 3.4) that were validated on CoReCG along with the mined summaries of the gene's interaction with CRC and the sources from where the descriptions were mined.

Gene Name	Fold Change	P-Value	Rank	Source
CYP2E1	-	-	-	-
HPX	214.761	7.23E-5	235 (in top 3%)	Su (Liver Cancer)
APOH	9.425	5.57E-5	513 (in top 3%)	Barretina (liver)
APOA1	1965.894	1.05E-5	135 (in top 2%)	Su (Liver Cancer)
APOB	189.986	9.37E-5	259 (in top 4%)	Su (Liver Cancer)
FGL1	57.707	8.57E-5	171 (in top 1%)	Bittner (liver)
FGB	12.265	3.73E-6	293 (in top 2%)	Barretina (liver)
ITIH2	108.585	7.69E-5	241 (in top 3%)	Su (Liver Cancer)
HP	2.044	1.34E-14	42 (in top 1%)	COAD
FGA	641.062	2.94E-5	174 (in top 3%)	Su (Liver Cancer)
ITIH1	-	-	-	-
AMBP	301.181	4.53E-5	113 (in top 1%)	Barretina (liver)
ORM1	11.165	6.23E-5	527 (in top 3%)	Barretina (liver)

Table D.4: Table showing the detailed Oncomine validation results for CRC vs Metastasis.

Gene Name	CoReCG Description	Source
CYP2E1	Association with 96-bp insertion polymorphism and risk of CRC	(Sameer *et al.*, 2011)
IGFBP1	Association with insulin in COAD	(Le Marchand *et al.*, 2010)
PLG	Differentially expressed in COAD and CRC	(Albrethsen *et al.*, 2010)

Table D.5: Table showing the liver metastases genes (Table 3.5) validated on CoReCG along with the mined summaries of the genes interaction with CRC and the sources from where the descriptions were mined.

Gene Name	Fold Change	P-Value	Rank	Source
CYP2E1	-	-	-	-
APOH	9.425	5.57E-5	513 (in top 3%)	Barretina (liver)
APOA1	1965.894	1.05E-5	135 (in top 2%)	Su (Liver Cancer)
HPX	214.761	7.23E-5	235 (in top 3%)	Su (Liver Cancer)
FGL1	57.707	8.57E-5	171 (in top 1%)	Bittner (liver)
ITIH1	-	-	-	-
IGFBP1	143.713	7.24E-5	157 (in top 1%)	Bittner (liver)
ITIH2	108.585	7.69E-5	241 (in top 3%)	Su (Liver Cancer)
AMBP	301.181	4.53E-5	113 (in top 1%)	Barretina (liver)
ITIH3	-	-	-	-
APOA2	105.848	3.83E-5	88 (in top 1%)	Yu (liver)
PLG	2.624	4.49E-12	72 (in top 1%)	-

Table D.6: Table showing the detailed Oncomine validation results for CRC vs Metastasis.

D.2 Blood Biomarker Results

This Appendix includes the meta scores from BBCancer for each individual gene from the Tables 3.3, 3.4 & 3.5.

Gene	CRC (blood)	CRC (EVs)
COL11A1	-	0.620719
ETV4	-0.7786803	0.656177
INHBA	0.644337	1.37514
ADAM12	-0.276544	0.50265
CLDN1	-0.137382	-0.554376
COL10A1	-0.187836	-1.49867
MMP1	1.440776	-2.1626
FAP	-	-
CTHRC1	0.702162	0.525781
CASC19	N/A	N/A
MMP3	0.459805	1.55269
KRT17	0.784388	0.749209
FOXQ1	-0.616123	2.43687

Table D.7: Table showing the meta scores for differential gene expression taken from different studies comparing the blood from CRC patients to a normal patients. The genes used were from the "CRC vs Normal" contrast. The second column indicates expression from peripheral blood, whereas the third column indicates expression for extracellular vesicles in the blood. "-" indicates genes that were found but had no expression data. "N/A" indicates genes that were not found on the database.

Gene	CRC (blood)	CRC (EVs)
CYP2E1	-1.03678	0.898923
HPX	-1.25042	1.46703
APOH	-0.0289761	-0.35751
APOA1	0.423458	0.144758
APOB	-0.340088	0.791403
FGL1	-0.589689	-0.0202076
FGB	0.554867	-0.0532516
ITIH2	-	-0.658083
HP	2.40197	-0.825083
FGA	-0.587274	0.109299
ITIH1	-0.248711	0.446475
AMBP	0.0463872	0.0993862
ORM1	-0.210331	-

Table D.8: Table showing the meta scores for differential gene expression taken from different studies comparing the blood from CRC patients to a normal patients. The genes used were from the "Metastasis vs CRC" contrast. The second column indicates expression from peripheral blood, whereas the third column indicates expression for extracellular vesicles in the blood. "-" indicates genes that were found but had no expression data.

Gene	CRC (blood)	CRC (EVs)
CYP2E1	-1.03678	0.898923
HPX	-1.25042	1.46703
APOH	-0.0289761	-0.35751
APOA1	0.423458	0.144758
FGL1	-0.589689	-0.0202076
IGFBP1	1.39187	-0.282653
ITIH2	-	-0.658083
ITIH3	-0.978324	0.538744
APOA2	0.813999	-0.134845
ITIH1	-0.248711	0.446475
AMBP	0.0463872	0.0993862
PLG	-1.74683	-0.21021

Table D.9: Table showing the meta scores for differential gene expression taken from different studies comparing the blood from CRC patients to a normal patients. The genes used were from the "Metastasis vs Normal" contrast. The second column indicates expression from peripheral blood, whereas the third column indicates expression for extracellular vesicles in the blood. "-" indicates genes that were found but had no expression data.

Gene	CRC (blood)	CRC (EVs)
COL11A1	-	-2.3034
ETV4	0.17195	-0.952092
INHBA	-0.682845	-0.601174
ADAM12	-0.0242739	-0.636253
CLDN1	-1.59686	-0.945864
COL10A1	-0.103577	-0.217084
MMP1	0.0787946	-0.600793
FAP	-	-1.21175
CTHRC1	0.500092	-0.623925
CASC19	N/A	N/A
MMP3	-0.781846	-1.86707
KRT17	0.431464	-1.05326
FOXQ1	0.368428	-1.02314

Table D.10: Table showing the meta scores for gene expression abundance taken from different studies comparing the blood from CRC patients to a normal patients. The genes used were from the "CRC vs Normal" contrast. The second column indicates expression from peripheral blood, whereas the third column indicates expression for extracellular vesicles in the blood. "-" indicates genes that were found but had no expression data. "N/A" indicates genes that were not found on the database.

Gene	CRC (blood)	CRC (EVs)
CYP2E1	-0.0987475	-0.824358
HPX	-1.77707	-0.789152
APOH	-1.85561	-0.46251
APOA1	-0.203214	-
APOB	-1.03717	-0.991874
FGL1	-2.56998	-
FGB	-0.141322	-0.763351
ITIH2	-	-0.778349
HP	-0.997006	-0.279489
FGA	-1.43266	-0.602953
ITIH1	-.022609	-0.952855
AMBP	-1.6327	-0.684296
ORM1	-1.13439	-0.0829951

Table D.11: Table showing the meta scores for gene expression abundance taken from different studies comparing the blood from CRC patients to a normal patients. The genes used were from the "Metastasis vs CRC" contrast. The second column indicates expression from peripheral blood, whereas the third column indicates expression for extracellular vesicles in the blood. "-" indicates genes that were found but had no expression data.

Gene	CRC (blood)	CRC (EVs)
CYP2E1	-0.0987475	-0.824358
HPX	-1.77707	-0.789152
APOH	-1.85561	-0.46251
APOA1	-0.203214	-
FGL1	-2.56998	-
IGFBP1	0.505429	-0.577915
ITIH2	-	-0.778349
ITIH3	-0.0379994	-1.09012
APOA2	-0.432481	-0.195858
ITIH1	-.022609	-0.952855
AMBP	-1.6327	-0.684296
PLG	-1.67146	-2.27912

Table D.12: Table showing the meta scores for gene expression abundance taken from different studies comparing the blood from CRC patients to a normal patients. The genes used were from the "Metastasis vs Normal" contrast. The second column indicates expression from peripheral blood, whereas the third column indicates expression for extracellular vesicles in the blood. "-" indicates genes that were found but had no expression data.

Bibliography

Abba, M., Patil, N., Rasheed, K., Nelson, L.D., Mudduluru, G., Leupold, J.H. and Allgayer, H. (2013). Unraveling the role of FOXQ1 in colorectal cancer metastasis. *Molecular Cancer Research*, vol. 11, no. 9, pp. 1017–1028.

Afshar-Kharghan, V. and Others (2017). The role of the complement system in cancer. *The Journal of clinical investigation*, vol. 127, no. 3, pp. 780–789.

Agarwal, R., Kumar, B., Jayadev, M., Raghav, D. and Singh, A. (2016*a*). CoReCG: A comprehensive Database of genes associated with colon-rectal cancer. *Database*, vol. 2016, pp. 1–9.

Agarwal, R., Kumar, B., Jayadev, M., Raghav, D. and Singh, A. (2016*b*). CoReCG: a comprehensive database of genes associated with colon-rectal cancer. *Database*, vol. 2016, pp. 1–9.

Albrethsen, J., Knol, J.C., Piersma, S.R., Pham, T.V., de Wit, M., Mongera, S., Carvalho, B., Verheul, H.M.W., Fijneman, R.J.A., Meijer, G.A. and Others (2010). Subnuclear proteomics in colorectal cancer: identification of proteins enriched in the nuclear matrix fraction and regulation in adenoma to carcinoma progression. *Molecular & Cellular Proteomics*, vol. 9, no. 5, pp. 988–1005.

Algra, A.M. and Rothwell, P.M. (2012). Effects of regular aspirin on long-term cancer incidence and metastasis: a systematic comparison of evidence from observational studies versus randomised trials. *The lancet oncology*, vol. 13, no. 5, pp. 518–527.

Allegra, C.J., Jessup, J.M., Somerfield, M.R., Hamilton, S.R., Hammond, E.H., Hayes, D.F., McAllister, P.K., Morton, R.F. and Schilsky, R.L. (2009). American Society of Clinical Oncology provisional clinical opinion: testing for KRAS gene mutations in patients with metastatic colorectal carcinoma to predict response to anti–epidermal growth factor receptor monoclonal antibody therapy. *Journal of clinical oncology*, vol. 27, no. 12, pp. 2091–2096.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Nature Precedings*, p. 1.

Anders, S., Pyl, P.T. and Huber, W. (2014). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, vol. 31, no. 2, pp. 166–169.

Andrews, S., Krueger, F., Segonds-Pichon, A., Biggins, L., Krueger, C. and Wingett, S. (2012 jan). FastQC. Babraham Institute.

Ascierto, P.A., Kirkwood, J.M., Grob, J.-J., Simeone, E., Grimaldi, A.M., Maio, M., Palmieri, G., Testori, A., Marincola, F.M. and Mozzillo, N. (2012). The role of BRAF V600 mutation in melanoma. *Journal of translational medicine*, vol. 10, no. 1, p. 85.

Awan, F.M., Naz, A., Obaid, A., Ali, A., Ahmad, J., Anjum, S. and Janjua, H.A. (2015). Identification of Circulating Biomarker Candidates for Hepatocellular Carcinoma (HCC): An Integrated Prioritization Approach. *PLOS ONE*, vol. 10, no. 9, pp. 1–26.

Bardelli, A. and Siena, S. (2010). Molecular mechanisms of resistance to cetuximab and panitumumab in colorectal cancer. *Journal of clinical oncology*, vol. 28, no. 7, pp. 1254–1261.

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M. and Others (2012). NCBI GEO: archive for functional genomics data sets—update. *Nucleic acids research*, vol. 41, no. D1, pp. D991—-D995.

Baxter, N.N., Goldwasser, M.A., Paszat, L.F., Saskin, R., Urbach, D.R. and Rabeneck, L. (2009). Association of colonoscopy and death from colorectal cancer. *Annals of internal medicine*, vol. 150, no. 1, pp. 1–8.

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)*, vol. 57, no. 1, pp. 289–300.

Berger, M.F., Levin, J.Z., Vijayendran, K., Sivachenko, A., Adiconis, X., Maguire, J., Johnson, L.A., Robinson, J., Verhaak, R.G., Sougnez, C. and Others (2010). Integrative analysis of the melanoma transcriptome. *Genome research*, vol. 20, no. 4, pp. 413–427.

Beydoun, H.A. and Beydoun, M.A. (2008). Predictors of colorectal cancer screening behaviors among average-risk older adults in the United States. *Cancer Causes & Control*, vol. 19, no. 4, pp. 339–359.

Bienz, M. and Clevers, H. (2000). Linking colorectal cancer to Wnt signaling. *Cell*, vol. 103, no. 2, pp. 311–320.

Blankenberg, D., Kuster, G.V., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A. and Taylor, J. (2010). Galaxy: a web-based genome analysis tool for experimentalists. *Current protocols in molecular biology*, vol. 89, no. 1, pp. 10–19.

Bokemeyer, C., Bondarenko, I., Hartmann, J.T., De Braud, F.G., Volovat, C., Nippgen, J., Stroh, C., Celik, I. and Koralewski, P. (2008). KRAS status and efficacy of first-line treatment of patients with metastatic colorectal cancer (mCRC) with FOLFOX with or without cetuximab: The OPUS experience. *Journal of Clinical Oncology*, vol. 26, no. 15_suppl, p. 4000.

Boland, C.R. and Goel, A. (2010). Microsatellite instability in colorectal cancer. *Gastroenterology*, vol. 138, no. 6, pp. 2073–2087.

Borgquist, S., Butt, T., Almgren, P., Shiffman, D., Stocks, T., Orho-Melander, M., Manjer, J. and Melander, O. (2016). Apolipoproteins, lipids and risk of cancer. *International journal of cancer*, vol. 138, no. 11, pp. 2648–2656.

Boudreau, N. and Bissell, M.J. (1998). Extracellular matrix signaling: integration of form and function in normal and malignant cells. *Current opinion in cell biology*, vol. 10, no. 5, p. 640.

Brabletz, T., Jung, A. and Kirchner, T. (2002). Beta-Catenin and the morphogenesis of colorectal cancer. *Virchows Archiv*, vol. 441, no. 1, pp. 1–11.

Brand, M., Gaylard, P. and Ramos, J. (2018). Colorectal cancer in South Africa: An assessment of disease presentation, treatment pathways and 5-year survival. *South African Medical Journal*, vol. 108, no. 2, pp. 118–122.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A. and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, vol. 68, no. 6, pp. 394–424.

Bray, F., Jemal, A., Grey, N., Ferlay, J. and Forman, D. (2012). Global cancer transitions according to the Human Development Index (2008–2030): a population-based study. *The Lancet Oncology*, vol. 13, no. 8, pp. 790–801.

Brocardo, M. and Henderson, B.R. (2008). APC shuttling to the membrane, nucleus and beyond. *Trends in cell biology*, vol. 18, no. 12, pp. 587–596.

Carlson, M., Falcon, S., Pages, H. and Li, N. (2017). GO. db: A set of annotation maps describing the entire Gene Ontology. *R package version*, vol. 3, no. 1, pp. 10–18.

Carmeliet, P. and Jain, R.K. (2000). Angiogenesis in cancer and other diseases. *nature*, vol. 407, no. 6801, pp. 249–257.

Caro, A.A. and Cederbaum, A.I. (2004). Oxidative stress, toxicology, and pharmacology of CYP2E1. *Annu. Rev. Pharmacol. Toxicol.*, vol. 44, pp. 27–42.

Chalkias, A., Nikotian, G., Koutsovasilis, A., Bramis, J., Manouras, A., Mystrioti, D. and Katergiannakis, V. (2011). Patients with colorectal cancer are characterized by increased concentration of fecal hb-hp complex, myeloperoxidase, and secretory IgA. *American journal of clinical oncology*, vol. 34, no. 6, pp. 561–566.

Cherradi, S., Ayrolles-Torro, A., Vezzo-Vié, N., Gueguinou, N., Denis, V., Combes, E., Boissiere, F., Busson, M., Canterel-Thouennon, L., Mollevi, C. and Others (2017). Antibody targeting of claudin-1 as a potential colorectal cancer therapy. *Journal of Experimental & Clinical Cancer Research*, vol. 36, no. 1, p. 89.

Cho, M.S., Rupaimoole, R., Choi, H.-J., Noh, K., Chen, J., Hu, Q., Sood, A.K. and Afshar-Kharghan, V. (2016). Complement component 3 is regulated by TWIST1 and mediates epithelial–mesenchymal transition. *The Journal of Immunology*, vol. 196, no. 3, pp. 1412–1418.

Cittadini, G. (2012). *Double contrast barium enema: the Genoa approach*. Springer Science & Business Media.

Consortium, G.O. (2019). The gene ontology resource: 20 years and still GOing strong. *Nucleic acids research*, vol. 47, no. D1, pp. D330—-D338.

Coon, M.J. (2005). Cytochrome P450: nature's most versatile biological catalyst. *Annu. Rev. Pharmacol. Toxicol.*, vol. 45, pp. 1–25.

Coskun, M., Bjerrum, J.T., Seidelin, J.B. and Nielsen, O.H. (2012). MicroRNAs in inflammatory bowel disease-pathogenesis, diagnostics and therapeutics. *World journal of gastroenterology: WJG*, vol. 18, no. 34, p. 4629.

Cotto, K.C., Wagner, A.H., Feng, Y.-Y., Kiwala, S., Coffman, A.C., Spies, G., Wollam, A., Spies, N.C., Griffith, O.L. and Griffith, M. (2018). DGIdb 3.0: a redesign and expansion of the drug–gene interaction database. *Nucleic acids research*, vol. 46, no. D1, pp. D1068—-D1073.

Danielsen, S.A., Eide, P.W., Nesbakken, A., Guren, T., Leithe, E. and Lothe, R.A. (2015). Portrait of the PI3K/AKT pathway in colorectal cancer. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, vol. 1855, no. 1, pp. 104–121.

der Jeught, K., Xu, H.-C., Li, Y.-J., Lu, X.-B. and Ji, G. (2018). Drug resistance and new therapies in colorectal cancer. *World journal of gastroenterology*, vol. 24, no. 34, p. 3834.

Dudoit, S., Yang, Y.H., Callow, M.J. and Speed, T.P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica sinica*, pp. 111–139.

Easton, D.F., Ford, D. and Bishop, D.T. (1995). Breast and ovarian cancer incidence in BRCA1-mutation carriers. Breast Cancer Linkage Consortium. *American journal of human genetics*, vol. 56, no. 1, p. 265.

Eftang, L.L., Esbensen, Y., Tannæs, T.M., Blom, G.P., Bukholm, I.R.K. and Bukholm, G. (2013). Up-regulation of CLDN1 in gastric cancer is correlated with reduced survival. *BMC cancer*, vol. 13, no. 1, p. 586.

Ewels, P., Magnusson, M., Lundin, S. and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, vol. 32, no. 19, pp. 3047–3048.

Ewing, B. and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome research*, vol. 8, no. 3, pp. 186–194.

Fadaka, A.O., Klein, A. and Pretorius, A. (2019). In silico identification of microRNAs as candidate colorectal cancer biomarkers. *Tumor Biology*, vol. 41, no. 11, p. 101.

Fearon, E.R. (2011). Molecular genetics of colorectal cancer. *Annual Review of Pathology: Mechanisms of Disease*, vol. 6, pp. 479–507.

Felipe De Sousa, E.M., Wang, X., Jansen, M., Fessler, E., Trinh, A., De Rooij, L.P.M.H., De Jong, J.H., De Boer, O.J., Van Leersum, R., Bijlsma, M.F. and Others (2013). Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nature medicine*, vol. 19, no. 5, p. 614.

Fink, S.P., Myeroff, L.L., Kariv, R., Platzer, P., Xin, B., Mikkola, D., Lawrence, E., Morris, N., Nosrati, A., Willson, J.K.V. and Others (2015). Induction of KIAA1199/CEMIP is associated with colon cancer phenotype and poor patient survival. *Oncotarget*, vol. 6, no. 31, p. 305.

Fischer, H., Stenling, R., Rubio, C. and Lindblom, A. (2001). Colorectal carcinogenesis is associated with stromal expression of COL11A1 and COL5A2. *Carcinogenesis*, vol. 22, no. 6, pp. 875–878.

Fu, X., Fu, N., Guo, S., Yan, Z., Xu, Y., Hu, H., Menzel, C., Chen, W., Li, Y., Zeng, R. and Others (2009). Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC genomics*, vol. 10, no. 1, p. 161.

Gaedcke, J., Grade, M., Jung, K., Camps, J., Jo, P., Emons, G., Gehoff, A., Sax, U., Schirmer, M., Becker, H. and Others (2010). Mutated KRAS results in overexpression of DUSP4, a MAP-kinase phosphatase, and SMYD3, a histone methyltransferase, in rectal carcinomas. *Genes, chromosomes and cancer*, vol. 49, no. 11, pp. 1024–1034.

Galván, J.A., Garcia-Martinez, J., Vázquez-Villa, F., Garc\'\ia-Ocaña, M., Garc\'\ia-Pravia, C., Menéndez-Rodr\'\iguez, P., González-del Rey, C., Barneo-Serra, L. and de los Toyos, J.R. (2014). Validation of COL11A1/procollagen 11A1 expression in TGF-B1-activated immortalised human mesenchymal cells and in stromal cells of human colon adenocarcinoma. *BMC cancer*, vol. 14, no. 1, p. 867.

Ganepola, G.A.P., Nizin, J., Rutledge, J.R. and Chang, D.H. (2014). Use of blood-based biomarkers for early diagnosis and surveillance of colorectal cancer. *World journal of gastrointestinal oncology*, vol. 6, no. 4, p. 83.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J. and Others (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, vol. 5, no. 10, p. R80.

Gil-Bernabé, A.M., Lucotti, S. and Muschel, R.J. (2013). Coagulation and metastasis: what does the experimental literature tell us? *British journal of haematology*, vol. 162, no. 4, pp. 433–441.

Gingras, I., Salgado, R. and Ignatiadis, M. (2015). Liquid biopsy: will it be the 'magic tool' for monitoring response of solid tumors to anticancer therapies? *Current opinion in oncology*, vol. 27, no. 6, pp. 560–567.

Gonzalez, F.J. (2005). Role of cytochromes P450 in chemical toxicity and oxidative stress: studies with CYP2E1. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 569, no. 1-2, pp. 101–110.

Goswami, C.P. and Nakshatri, H. (2014). PROGgeneV2: enhancements on the existing database. *BMC cancer*, vol. 14, no. 1, pp. 1–6.

Grady, W.M. and Carethers, J.M. (2008). Genomic and epigenetic instability in colorectal cancer pathogenesis. *Gastroenterology*, vol. 135, no. 4, pp. 1079–1099.

Gu, L., Saha, S.T., Thomas, J. and Kaur, M. (2019). Targeting cellular cholesterol for anticancer therapy. *The FEBS journal*, vol. 286, no. 21, pp. 4192–4208.

Guinney, J., Dienstmann, R., Wang, X., De Reyniès, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P. and Others (2015).

The consensus molecular subtypes of colorectal cancer. *Nature medicine*, vol. 21, no. 11, pp. 1350–1356.

Haggar, F.A. and Boushey, R.P. (2009). Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. *Clinics in colon and rectal surgery*, vol. 22, no. 04, pp. 191–197.

Harbison, C.T., Horam, C.E. and Khambata-Ford, S. (2011). Developing predictive biomarkers in oncology. *Personalized Medicine*, vol. 8, no. 2, pp. 149–159.

Harris, L., Fritsche, H., Mennel, R., Norton, L., Ravdin, P., Taube, S., Somerfield, M.R., Hayes, D.F. and Bast Jr, R.C. (2007). American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. *Journal of clinical oncology*, vol. 25, no. 33, pp. 5287–5312.

Henry, N.L. and Hayes, D.F. (2012). Cancer biomarkers. *Molecular oncology*, vol. 6, no. 2, pp. 140–146.

Huang, D., Sun, W., Zhou, Y., Li, P., Chen, F., Chen, H., Xia, D., Xu, E., Lai, M., Wu, Y. and Others (2018*a*). Mutations of key driver genes in colorectal cancer progression and metastasis. *Cancer and Metastasis Reviews*, vol. 37, no. 1, pp. 173–187.

Huang, H., Han, Y., Gao, J., Feng, J., Zhu, L., Qu, L., Shen, L. and Shou, C. (2013). High level of serum AMBP is associated with poor response to paclitaxel–capecitabine chemotherapy in advanced gastric cancer patients. *Medical Oncology*, vol. 30, no. 4, p. 748.

Huang, H., Li, T., Ye, G., Zhao, L., Zhang, Z., Mo, D., Wang, Y., Zhang, C., Deng, H., Li, G. and Others (2018*b*). High expression of COL10A1 is associated with poor prognosis in colorectal cancer. *OncoTargets and therapy*, vol. 11, p. 1571.

Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B.S., Bravo, H.C., Davis, S., Gatto, L., Girke, T. and Others (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature methods*, vol. 12, no. 2, pp. 115–121.

Jacobs, R.J., Voorneveld, P.W., Kodach, L.L. and Hardwick, J.C.H. (2012). Cholesterol metabolism and colorectal cancers. *Current opinion in pharmacology*, vol. 12, no. 6, pp. 690–695.

Jaiswal, A.S. and Narayan, S. (2008). A novel function of adenomatous polyposis coli (APC) in regulating DNA repair. *Cancer letters*, vol. 271, no. 2, pp. 272–280.

Jass, J.R., Young, J. and Leggett, B.A. (2002). Evolution of colorectal cancer: change of pace and change of direction. *Journal of gastroenterology and hepatology*, vol. 17, no. 1, pp. 17–26.

Jing, X., Tian, Z.-B., Gao, P.-J., Han, N.-J., Xu, Y.-H., Zhang, H., Ding, X.-L., Wang, X.-W., Man, X. and Zhang, C.P. (2015). Lipopolysaccharide Enhances Beta2-Glycoprotein I Activation of Nuclear Factor kB in Liver Cancer Cells. *Clinical laboratory*, vol. 61, no. 9, pp. 1239–1245.

Jonker, D.J., O'Callaghan, C.J., Karapetis, C.S., Zalcberg, J.R., Tu, D., Au, H.-J., Berry, S.R., Krahn, M., Price, T., Simes, R.J. and Others (2007). Cetuximab for the treatment of colorectal cancer. *New England Journal of Medicine*, vol. 357, no. 20, pp. 2040–2048.

Kaiser, S., Park, Y.-K., Franklin, J.L., Halberg, R.B., Yu, M., Jessen, W.J., Freudenberg, J., Chen, X., Haigis, K., Jegga, A.G. and Others (2007). Transcriptional recapitulation and subversion of embryonic colon development by mouse colon tumor models and human colon cancer. *Genome biology*, vol. 8, no. 7, p. R131.

Kalluri, R. and Weinberg, R.A. (2009). The basics of epithelial-mesenchymal transition. *Journal of Clinical Investigation*, vol. 119, no. 6, pp. 1420–1428.

Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, vol. 28, no. 1, pp. 27–30.

Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, vol. 53, no. 282, pp. 457–481.

Karabulut, M., Alis, H., Bas, K., Karabulut, S., Afsar, C.U., Oguz, H., Gunaldi, M., Akarsu, C., Kones, O. and Aykan, N.F. (2015). Clinical significance of serum claudin-1 and claudin-7 levels in patients with colorectal cancer. *Molecular and clinical oncology*, vol. 3, no. 6, pp. 1255–1267.

Kaur, M., MacPherson, C.R., Schmeier, S., Narasimhan, K., Choolani, M. and Bajic, V.B. (2011). In Silico discovery of transcription factors as potential diagnostic biomarkers of ovarian cancer. *BMC systems biology*, vol. 5, no. 1, p. 144.

Kerr, M.K., Martin, M. and Churchill, G.A. (2000). Analysis of variance for gene expression microarray data. *Journal of computational biology*, vol. 7, no. 6, pp. 819–837.

Ki, D.H., Jeung, H.-C., Park, C.H., Kang, S.H., Lee, G.Y., Lee, W.S., Kim, N.K., Chung, H.C. and Rha, S.Y. (2007). Whole genome analysis for liver metastasis

gene signatures in colorectal cancer. *International journal of cancer*, vol. 121, no. 9, pp. 2005–2012.

Kim, C.Y., Jung, W.Y., Lee, H.J., Kim, H.K., Kim, A. and Shin, B.K. (2012). Proteomic analysis reveals overexpression of moesin and cytokeratin 17 proteins in colorectal carcinoma. *Oncology reports*, vol. 27, no. 3, pp. 608–620.

Kim, D., Langmead, B. and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature methods*, vol. 12, no. 4, pp. 357–360.

Kim, H.C., Kim, Y.S., Oh, H.-W., Kim, K., Oh, S.-S., Kim, J.-T., Kim, B.Y., Lee, S.-J., Choe, Y.-K., Kim, D.H. and Others (2014*a*). Collagen triple helix repeat containing 1 (CTHRC1) acts via ERK-dependent induction of MMP9 to promote invasion of colorectal cancer cells. *Oncotarget*, vol. 5, no. 2, p. 519.

Kim, S.-K., Kim, S.-Y., Kim, J.-H., Roh, S.A., Cho, D.-H., Kim, Y.S. and Kim, J.C. (2014*b*). A nineteen gene-based risk score classifier predicts prognosis of colorectal cancer patients. *Molecular oncology*, vol. 8, no. 8, pp. 1653–1666.

Kleinbaum, D.G. and Klein, M. (2010). *Survival analysis*. Springer.

Kumari, S. and Malla, R. (2015). New insight on the role of plasminogen receptor in cancer progression. *Cancer growth and metastasis*, vol. 8, pp. CGM—-S27335.

Langfelder, P. and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, vol. 9, no. 1, p. 559.

Le Marchand, L., Wang, H., Rinaldi, S., Kaaks, R., Vogt, T.M., Yokochi, L. and Decker, R. (2010). Associations of plasma C-peptide and IGFBP-1 levels with risk of colorectal adenoma in a multiethnic population. *Cancer Epidemiology and Prevention Biomarkers*, vol. 19, no. 6, pp. 1471–1477.

Ledergerber, C. and Dessimoz, C. (2011). Base-calling for next-generation sequencing platforms. *Briefings in bioinformatics*, vol. 12, no. 5, pp. 489–497.

Lee, J.T. (2012). Epigenetic regulation by long noncoding RNAs. *Science*, vol. 338, no. 6113, pp. 1435–1439.

Leinonen, R., Sugawara, H., Shumway, M. and Collaboration, I.N.S.D. (2010). The sequence read archive. *Nucleic acids research*, vol. 39, no. suppl_1, pp. D19—-D21.

Levin, T.R. and Corley, D.A. (2013). Colorectal-cancer screening—coming of age.

Li, B., Feng, W., Luo, O., Xu, T., Cao, Y., Wu, H., Yu, D. and Ding, Y. (2017). Development and validation of a three-gene prognostic signature for patients with hepatocellular carcinoma. *Scientific reports*, vol. 7, no. 1, pp. 1–13.

Li, T., Huang, H., Shi, G., Zhao, L., Li, T., Zhang, Z., Liu, R., Hu, Y., Liu, H., Yu, J. and Others (2018). TGF-B1-SOX9 axis-inducible COL10A1 promotes invasion and metastasis in gastric cancer via epithelial-to-mesenchymal transition. *Cell death & disease*, vol. 9, no. 9, pp. 1–18.

Li, X., Yu, W., Liang, C., Xu, Y., Zhang, M., Ding, X. and Cai, X. (2020). INHBA is a prognostic predictor for patients with colon adenocarcinoma. *BMC cancer*, vol. 20, pp. 1–10.

Liang, Y., Zhang, C., Ma, M.-H. and Dai, D.-Q. (2018). Identification and prediction of novel non-coding and coding RNA-associated competing endogenous RNA networks in colorectal cancer. *World journal of gastroenterology*, vol. 24, no. 46, p. 5259.

Liao, Y., Smyth, G.K. and Shi, W. (2013). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, vol. 30, no. 7, pp. 923–930.

Liao, Y., Wang, J., Jaehnig, E.J., Shi, Z. and Zhang, B. (2019). WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic acids research*, vol. 47, no. W1, pp. W199—-W205.

Lin, K., Lipsitz, R., Miller, T. and Janakiraman, S. (2008). Benefits and harms of prostate-specific antigen screening for prostate cancer: an evidence update for the US Preventive Services Task Force. *Annals of internal medicine*, vol. 149, no. 3, pp. 192–199.

Liu, D., Duan, W., Guo, H., Xu, X. and Bai, Y. (2011). Meta-analysis of associations between polymorphisms in the promoter regions of matrix metalloproteinases and the risk of colorectal cancer. *International journal of colorectal disease*, vol. 26, no. 9, p. 1099.

Liu, Y., Colby, J.K., Zuo, X., Jaoude, J., Wei, D. and Shureiqi, I. (2018). The role of PPAR in metabolism, inflammation, and cancer: Many characters of a critical transcription factor. *International journal of molecular sciences*, vol. 19, no. 11, p. 3339.

Locker, G.Y., Hamilton, S., Harris, J., Jessup, J.M., Kemeny, N., Macdonald, J.S., Somerfield, M.R., Hayes, D.F. and Bast Jr, R.C. (2006). ASCO 2006 update of recommendations for the use of tumor markers in gastrointestinal cancer. *Journal of clinical oncology*, vol. 24, no. 33, pp. 5313–5327.

Logan, C.Y. and Nusse, R. (2004). The Wnt signaling pathway in development and disease. *Annu. Rev. Cell Dev. Biol.*, vol. 20, pp. 781–810.

Love, M.I., Huber, W. and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, vol. 15, no. 12, p. 550.

Lu, D.Y., Ding, J., Lu, T.R., Yarla, N.S., Wu, H.Y. and Others (2018). Cancer Bioinformatics in Cancer Therapy. *Adv Proteomics Bioin-form: APBI-111. DOI*, vol. 10.

Lu, L., Sun, Y., Li, Y. and Wan, P. (2015). The polymorphism MMP1- 1607 (1G> 2G) is associated with a significantly increased risk of cancers from a meta-analysis. *Tumor Biology*, vol. 36, no. 3, pp. 1685–1693.

Mackenzie, R.J. (2018). RNA-seq: Basics, Applications and Protocol. Available at: https://www.technologynetworks.com/genomics/articles/rna-seq-basics-applications-and-protocol-299461

Madhunapantula, S.V., Mosca, P. and Robertson, G.P. (2010). Steroid hormones drive cancer development. *Cancer biology & therapy*, vol. 10, no. 8, pp. 765–766.

Mantione, K.J., Kream, R.M., Kuzelova, H., Ptacek, R., Raboch, J., Samuel, J.M. and Stefano, G.B. (2014). Comparing bioinformatic gene expression profiling methods: microarray and RNA-Seq. *Medical science monitor basic research*, vol. 20, p. 138.

Marley, A.R. and Nan, H. (2016). Epidemiology of colorectal cancer. *International journal of molecular epidemiology and genetics*, vol. 7, no. 3, p. 105.

Marshall, K.W., Mohr, S., Khettabi, F.E., Nossova, N., Chao, S., Bao, W., Ma, J., Li, X.-J. and Liew, C.-C. (2010). A blood-based biomarker panel for stratifying current risk for colorectal cancer. *International journal of cancer*, vol. 126, no. 5, pp. 1177–1186.

McMahon, B. and Kwaan, H.C. (2007). The plasminogen activator system and cancer. *Pathophysiology of haemostasis and thrombosis*, vol. 36, no. 3-4, pp. 184–194.

Moloney, J.N. and Cotter, T.G. (2018). ROS signalling in the biology of cancer. In: *Seminars in cell & developmental biology*, vol. 80, pp. 50–64. Elsevier.

Morán, A., Iniesta, P., de Juan, C., González-Quevedo, R., Sánchez-Pernaute, A., D\'\iaz-Rubio, E., y Cajal, S.R., Torres, A., Balibrea, J.L. and Benito, M. (2002). Stromelysin-1 promoter mutations impair gelatinase B activation in high microsatellite instability sporadic colorectal tumors. *Cancer research*, vol. 62, no. 13, pp. 3855–3860.

Moss, A.C., Lawlor, G., Murray, D., Tighe, D., Madden, S.F., Mulligan, A.-M., Keane, C.O., Brady, H.R., Doran, P.P. and MacMathuna, P. (2006). ETV4 and Myeov knockdown impairs colon cancer cell line proliferation and invasion. *Biochemical and biophysical research communications*, vol. 345, no. 1, pp. 216–221.

Motsuku, L.M., Chen, W.C., Muchengeti, M.M., Mac Quene, T., Kellett, P., Mohlala, M.I., Chu, K.M., Singh, E. and Naidoo, M. (2020). Colorectal Cancer Incidence and Mortality Trends by Sex and Population Group in South Africa: 2002-2014.

Mundade, R., Imperiale, T.F., Prabhu, L., Loehrer, P.J. and Lu, T. (2014). Genetic pathways, prevention, and treatment of sporadic colorectal cancer. *Oncoscience*, vol. 1, no. 6, p. 400.

Murray, G.I., Duncan, M.E., O'Neil, P., Melvin, W.T. and Fothergill, J.E. (1996). Matrix metalloproteinase–1 is associated with poor prognosis in colorectal cancer. *Nature medicine*, vol. 2, no. 4, pp. 461–462.

Nakagawa, S., Miyoshi, N., Ishii, H., Mimori, K., Tanaka, F., Sekimoto, M., Doki, Y. and Mori, M. (2011). Expression of CLDN1 in colorectal cancer: a novel marker for prognosis. *International journal of oncology*, vol. 39, no. 4, pp. 791–796.

Nakanishi, Y., Diaz-Meco, M.T. and Moscat, J. (2019). Serrated Colorectal Cancer: The Road Less Travelled? *Trends in cancer*.

Network, C.G.A. and Others (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, vol. 487, no. 7407, p. 330.

Nosho, K., Yoshida, M., Yamamoto, H., Taniguchi, H., Adachi, Y., Mikami, M., Hinoda, Y. and Imai, K. (2005). Association of Ets-related transcriptional factor E1AF expression with overexpression of matrix metalloproteinases, COX-2 and iNOS in the early stage of colorectal carcinogenesis. *Carcinogenesis*, vol. 26, no. 5, pp. 892–899.

Okano, M., Yamamoto, H., Ohkuma, H., Kano, Y., Kim, H., Nishikawa, S., Konno, M., Kawamoto, K., Haraguchi, N., Takemasa, I. and Others (2013). Significance of INHBA expression in human colorectal cancer. *Oncology reports*, vol. 30, no. 6, pp. 2903–2908.

Okoniewski, M. and Miller, C.J. (2006). Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC bioinformatics*, vol. 7, no. 1, p. 276.

Ontario, H.Q. and Others (2009). Fecal occult blood test for colorectal cancer screening: an evidence-based analysis. *Ont Health Technol Assess Ser*, vol. 9, no. 10, pp. 1–40.

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F.L., Walker, M.G., Watson, D., Park, T. and Others (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine*, vol. 351, no. 27, pp. 2817–2826.

Parikh, R., Mathai, A., Parikh, S., Sekhar, G.C. and Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian journal of ophthalmology*, vol. 56, no. 1, p. 45.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A. and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, vol. 14, no. 4, p. 417.

Peltier, J., Roperch, J.-P., Audebert, S., Borg, J.-P. and Camoin, L. (2016). Quantitative proteomic analysis exploring progression of colorectal cancer: Modulation of the serpin family. *Journal of proteomics*, vol. 148, pp. 139–148.

Peng, L., Bian, X.W., Xu, C., Wang, G.M., Xia, Q.Y., Xiong, Q. and Others (2015). Large-scale RNA-Seq transcriptome analysis of 4043 cancers and 548 normal tissue controls across 12 TCGA cancer types. *Scientific reports*, vol. 5, p. 13413.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
Available at: https://www.r-project.org/

Rhodes, D.R., Kalyana-Sundaram, S., Mahavisno, V., Varambally, R., Yu, J., Briggs, B.B., Barrette, T.R., Anstet, M.J., Kincead-Beal, C., Kulkarni, P. and Others (2007). Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia (New York, NY)*, vol. 9, no. 2, p. 166.

Rhodes, D.R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A. and Chinnaiyan, A.M. (2004). ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia (New York, NY)*, vol. 6, no. 1, p. 1.

Romiguier, J., Ranwez, V., Douzery, E.J.P. and Galtier, N. (2010). Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome research*, vol. 20, no. 8, pp. 1001–1009.

Rook, A.H., Kehrl, J.H., Wakefield, L.M., Roberts, A.B., Sporn, M.B., Burlington, D.B., Lane, H.C. and Fauci, A.S. (1986). Effects of transforming growth factor beta

on the functions of natural killer cells: depressed cytolytic activity and blunting of interferon responsiveness. *The Journal of Immunology*, vol. 136, no. 10, pp. 3916–3920.

RStudio Team (2015). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.
Available at: http://www.rstudio.com/

Sameer, A.S., Nissar, S., Qadri, Q., Alam, S., Baba, S.M. and Siddiqi, M.A. (2011). Role of CYP2E1 genotypes in susceptibility to colorectal cancer in the Kashmiri population. *Human genomics*, vol. 5, no. 6, p. 530.

Sarma, J.V. and Ward, P.A. (2011). The complement system. *Cell and tissue research*, vol. 343, no. 1, pp. 227–235.

Sartore-Bianchi, A., Martini, M., Molinari, F., Veronese, S., Nichelatti, M., Artale, S., Di Nicolantonio, F., Saletti, P., De Dosso, S., Mazzucchelli, L. and Others (2009). PIK3CA mutations in colorectal cancer are associated with clinical resistance to EGFR-targeted monoclonal antibodies. *Cancer research*, vol. 69, no. 5, pp. 1851–1857.

Schell, M.J., Yang, M., Teer, J.K., Lo, F.Y., Madan, A., Coppola, D., Monteiro, A.N.A., Nebozhyn, M.V., Yue, B., Loboda, A. and Others (2016). A multigene mutation classification of 468 colorectal cancers reveals a prognostic role for APC. *Nature communications*, vol. 7, no. 1, pp. 1–12.

Schneider, V.A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P.A., Murphy, T.D., Pruitt, K.D., Thibaud-Nissen, F., Albracht, D. and Others (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome research*, vol. 27, no. 5, pp. 849–864.

Shaukat, A., Mongin, S.J., Geisser, M.S., Lederle, F.A., Bond, J.H., Mandel, J.S. and Church, T.R. (2013). Long-term mortality after screening for colorectal cancer. *New England Journal of Medicine*, vol. 369, no. 12, pp. 1106–1114.

Shen, J., Stass, S.A. and Jiang, F. (2013). MicroRNAs as potential biomarkers in human solid tumors. *Cancer letters*, vol. 329, no. 2, pp. 125–136.

Shibue, T. and Weinberg, R.A. (2017). EMT, CSCs, and drug resistance: the mechanistic link and clinical implications. *Nature reviews Clinical oncology*, vol. 14, no. 10, p. 611.

Shiozawa, J., Ito, M., Nakayama, T., Nakashima, M., Kohno, S. and Sekine, I. (2000). Expression of matrix metalloproteinase-1 in human colorectal carcinoma. *Modern Pathology*, vol. 13, no. 9, pp. 925–933.

Siegel, R.L., Miller, K.D., Fedewa, S.A., Ahnen, D.J., Meester, R.G.S., Barzi, A. and Jemal, A. (2017). Colorectal cancer statistics, 2017. *CA: a cancer journal for clinicians*, vol. 67, no. 3, pp. 177–193.

Siegel, R.L., Miller, K.D., Goding Sauer, A., Fedewa, S.A., Butterly, L.F., Anderson, J.C., Cercek, A., Smith, R.A. and Jemal, A. (2020*a*). Colorectal cancer statistics, 2020. *CA: a cancer journal for clinicians*.

Siegel, R.L., Miller, K.D. and Jemal, A. (2019). Cancer statistics, 2019. *CA: A Cancer Journal for Clinicians*, vol. 69, no. 1, pp. 7–34.

Siegel, R.L., Miller, K.D. and Jemal, A. (2020*b*). Cancer statistics, 2020. *CA: a cancer journal for clinicians*, vol. 70, no. 1, pp. 7–30.

Singh, A.B., Sharma, A. and Dhawan, P. (2012). Claudin-1 expression confers resistance to anoikis in colon cancer cells in a Src-dependent manner. *Carcinogenesis*, vol. 33, no. 12, pp. 2538–2547.

Sinicropi, D., Qu, K., Collin, F., Crager, M., Liu, M.-L., Pelham, R.J., Pho, M., Dei Rossi, A., Jeong, J., Scott, A. and Others (2012). Whole transcriptome RNA-Seq analysis of breast cancer recurrence risk using formalin-fixed paraffin-embedded tumor tissue. *PloS one*, vol. 7, no. 7, p. e40092.

Sipos, F., Germann, T.M., Wichmann, B., Galamb, O., Spisák, S., Krenács, T., Tulassay, Z., Molnár, B. and Muzes, G. (2014). MMP3 and CXCL1 are potent stromal protein markers of dysplasia–carcinoma transition in sporadic colorectal cancer. *European journal of cancer prevention*, vol. 23, no. 5, pp. 336–343.

Siravegna, G. and Bardelli, A. (2016). Blood circulating tumor DNA for non-invasive genotyping of colon cancer patients. *Molecular oncology*, vol. 10, no. 3, pp. 475–480.

Siveen, K.S., Raza, A., Ahmed, E.I., Khan, A.Q., Prabhu, K.S., Kuttikrishnan, S., Mateo, J.M., Zayed, H., Rasul, K., Azizi, F. and Others (2019). The role of extracellular vesicles as modulators of the tumor microenvironment, metastasis and drug resistance in colorectal cancer. *Cancers*, vol. 11, no. 6, p. 746.

Skrzypczak, M., Goryca, K., Rubel, T., Paziewska, A., Mikula, M., Jarosz, D., Pachlewski, J., Oledzki, J. and Ostrowsk, J. (2010). Modeling oncogenic signaling in colon tumors by multidirectional analyses of microarray data directed for maximization of analytical reliability. *PloS one*, vol. 5, no. 10, p. e13091.

Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, vol. 3, no. 1.

Sourvinou, I.S., Markou, A. and Lianidou, E.S. (2013). Quantification of circulating miRNAs in plasma: effect of preanalytical and analytical parameters on their isolation and stability. *The Journal of Molecular Diagnostics*, vol. 15, no. 6, pp. 827–834.

Stark, R., Grzelak, M. and Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature Reviews Genetics*, vol. 20, no. 11, pp. 631–656.

Stefanius, K., Ylitalo, L., Tuomisto, A., Kuivila, R., Kantola, T., Sirniö, P., Karttunen, T.J. and Mäkinen, M.J. (2011). Frequent mutations of KRAS in addition to BRAF in colorectal serrated adenocarcinoma. *Histopathology*, vol. 58, no. 5, pp. 679–692.

Stracke, M.L., Murata, J., Aznavoorian, S. and Liotta, L.A. (1994). The role of the extracellular matrix in tumor cell metastasis. *In vivo (Athens, Greece)*, vol. 8, no. 1, pp. 49–58.

Su, C., Zhao, J., Hong, X., Yang, S., Jiang, Y. and Hou, J. (2019). Microarray-based analysis of COL11A1 and TWIST1 as important differentially-expressed pathogenic genes between left and right-sided colon cancer. *Molecular medicine reports*, vol. 20, no. 5, pp. 4202–4214.

Sun, L., Pan, J., Peng, L., Fang, L., Zhao, X., Sun, L., Yang, Z. and Ran, Y. (2012). Combination of haptoglobin and osteopontin could predict colorectal cancer hepatic metastasis. *Annals of surgical oncology*, vol. 19, no. 7, pp. 2411–2419.

Sunami, E., Tsuno, N., Osada, T., Saito, S., Kitayama, J., Tomozawa, S., Tsuruo, T., Shibata, Y., Muto, T. and Nagawa, H. (2000). MMP-1 is a prognostic marker for hematogenous metastasis of colorectal cancer. *The oncologist*, vol. 5, no. 2, pp. 108–114.

Tan, C.R.C., Zhou, L. and El-Deiry, W.S. (2016). Circulating tumor cells versus circulating tumor DNA in colorectal cancer: pros and cons. *Current colorectal cancer reports*, vol. 12, no. 3, pp. 151–161.

Tang, Z., Kang, B., Li, C., Chen, T. and Zhang, Z. (2019). GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic acids research*, vol. 47, no. W1, pp. W556—-W560.

Tang, Z., Li, C., Kang, B., Gao, G., Li, C. and Zhang, Z. (2017). GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic acids research*, vol. 45, no. W1, pp. W98—-W102.

Therneau, T.M. (2020). A Package for Survival Analysis in R.
Available at: https://cran.r-project.org/package=survival

Thomas, D.A. and Massagué, J. (2005). TGF-beta directly targets cytotoxic T cell functions during tumor evasion of immune surveillance. *Cancer cell*, vol. 8, no. 5, pp. 369–380.

Tsui, N.B.Y., Ng, E.K.O. and Lo, Y.M.D. (2002). Stability of endogenous and added RNA in blood specimens, serum, and plasma. *Clinical chemistry*, vol. 48, no. 10, pp. 1647–1653.

Tuppurainen, K., Mäkinen, J.M., Junttila, O., Liakka, A., Kyllönen, A.P., Tuominen, H., Karttunen, T.J. and Mäkinen, M.J. (2005). Morphology and microsatellite instability in sporadic serrated and non-serrated colorectal cancer. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, vol. 207, no. 3, pp. 285–294.

van Dam, S., Võsa, U., van der Graaf, A., Franke, L. and de Magalhães, J.P. (2018). Gene co-expression analysis for functional classification and gene-disease predictions. *Briefings in bioinformatics*, vol. 19, no. 4, pp. 575–592.

van de Wetering, M., Francies, H.E., Francis, J.M., Bounova, G., Iorio, F., Pronk, A., van Houdt, W., van Gorp, J., Taylor-Weiner, A., Kester, L. and Others (2015). Prospective derivation of a living organoid biobank of colorectal cancer patients. *Cell*, vol. 161, no. 4, pp. 933–945.

Van Dijk, E.L., Auger, H., Jaszczyszyn, Y. and Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in genetics*, vol. 30, no. 9, pp. 418–426.

Van Duijnhoven, F.J.B., Bueno-De-Mesquita, H.B., Calligaro, M., Jenab, M., Pischon, T., Jansen, E.H.J.M., Frohlich, J., Ayyobi, A., Overvad, K., Toft-Petersen, A.P. and Others (2011). Blood lipid and lipoprotein concentrations and colorectal cancer risk in the European Prospective Investigation into Cancer and Nutrition. *Gut*, vol. 60, no. 8, pp. 1094–1102.

Vega, P., Valent\'\in, F. and Cubiella, J. (2015). Colorectal cancer diagnosis: pitfalls and opportunities. *World journal of gastrointestinal oncology*, vol. 7, no. 12, p. 422.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A. and Kinzler, K.W. (2013). Cancer genome landscapes. *Science*, vol. 339, no. 6127, pp. 1546–1558.

Walker, C., Mojares, E. and del R\'\io Hernández, A. (2018). Role of extracellular matrix in development and cancer progression. *International journal of molecular sciences*, vol. 19, no. 10, p. 3028.

Wang, C., Li, P., Xuan, J., Zhu, C., Liu, J., Shan, L., Du, Q., Ren, Y. and Ye, J. (2017*a*). Cholesterol enhances colorectal cancer progression via ROS elevation and MAPK signaling pathway activation. *Cellular Physiology and Biochemistry*, vol. 42, no. 2, pp. 729–742.

Wang, J.J., Li, X.M., He, L., Zhong, S.Z., Peng, Y.X. and Ji, N. (2017*b*). Expression and Function of Long Non-coding RNA CASC19 in Colorectal Cancer. *Zhongguo yi xue ke xue yuan xue bao. Acta Academiae Medicinae Sinicae*, vol. 39, no. 6, p. 756.

Wang, L., Wang, S. and Li, W. (2012*a*). RSeQC: quality control of RNA-seq experiments. *Bioinformatics*, vol. 28, no. 16, pp. 2184–2185.

Wang, X., Wu, F., Huang, R., Xue, F., Liang, G., Tao, M., Cai, P., Huang, Y. and Others (2012*b*). Transcriptome profiling of the cancer, adjacent non-tumor and distant normal tissues from a colorectal cancer patient by deep sequencing. *PloS one*, vol. 7, no. 8, p. e41001.

Wang, X.-D., Lu, J., Lin, Y.-S., Gao, C. and Qi, F. (2019). Functional role of long non-coding RNA CASC19/miR-140-5p/CEMIP axis in colorectal cancer progression in vitro. *World journal of gastroenterology*, vol. 25, no. 14, p. 1697.

Wu, Y. and Xu, Y. (2020). Clinical Significance of COL11A1 and Its Effect on Immune Infiltration in Colorectal Cancer. *Available at SSRN 3514645*.

Wu, Y.-H., Chang, T.H., Huang, Y.-F., Huang, H.D. and Chou, C.-Y. (2014). COL11A1 promotes tumor progression and predicts poor clinical outcome in ovarian cancer. *Oncogene*, vol. 33, no. 26, pp. 3432–3440.

Xu, X., Zhang, Y., Williams, J., Antoniou, E., McCombie, W.R., Wu, S., Zhu, W., Davidson, N.O., Denoya, P. and Li, E. (2013). Parallel comparison of Illumina RNA-Seq and Affymetrix microarray platforms on transcriptomic profiles generated from 5-aza-deoxy-cytidine treated HT-29 colon cancer cells and simulated datasets. *BMC bioinformatics*, vol. 14, no. 9, p. S1.

Yang, S., Sun, Z., Zhou, Q., Wang, W., Wang, G., Song, J., Li, Z., Zhang, Z., Chang, Y., Xia, K. and Others (2018). MicroRNAs, long noncoding RNAs, and circular

RNAs: potential tumor biomarkers and targets for colorectal cancer. *Cancer Management and Research*, vol. 10, p. 2249.

Yip, K.-T., Das, P.K., Suria, D., Lim, C.-R., Ng, G.-H. and Liew, C.-C. (2010). A case-controlled validation study of a blood-based seven-gene biomarker panel for colorectal cancer in Malaysia. *Journal of experimental & clinical cancer research*, vol. 29, no. 1, p. 128.

Yu, G., Wang, L.-G., Han, Y. and He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*, vol. 16, no. 5, pp. 284–287.

Yu, Q. and Stamenkovic, I. (2000 jan). Cell surface-localized matrix metalloproteinase-9 proteolytically activates TGF-beta and promotes tumor invasion and angiogenesis. *Genes & development*, vol. 14, no. 2, pp. 163–176.

Zhang, B. and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, vol. 4, no. 1.

Zhang, R., Liu, Q., Li, T., Liao, Q. and Zhao, Y. (2019*a*). Role of the complement system in the tumor microenvironment. *Cancer Cell International*, vol. 19, no. 1, p. 300.

Zhang, T., Guo, J., Gu, J., Wang, Z., Wang, G., Li, H. and Wang, J. (2019*b*). Identifying the key genes and microRNAs in colorectal cancer liver metastasis by bioinformatics analysis and in vitro experiments. *Oncology reports*, vol. 41, no. 1, pp. 279–291.

Zhang, X., Zhao, X.-W., Liu, D.-B., Han, C.-Z., Du, L.-L., Jing, J.-X. and Wang, Y. (2014). Lipid levels in serum and cancerous tissues of colorectal cancer patients. *World Journal of Gastroenterology: WJG*, vol. 20, no. 26, p. 8646.

Zhao, L., Wang, H., Sun, X. and Ding, Y. (2010). Comparative proteomic analysis identifies proteins associated with the development and progression of colorectal carcinoma. *The FEBS journal*, vol. 277, no. 20, pp. 4195–4204.

Zhu, H., Li, F., Tao, K., Wang, J., Scurlock, C., Zhang, X. and Xu, H. (2020). Comparison of the participation rate between CT colonography and colonoscopy in screening population: a systematic review and meta-analysis of randomized controlled trials. *The British Journal of Radiology*, vol. 93, no. 1105, p. 20190240.

Zoratto, F., Rossi, L., Verrico, M., Papa, A., Basso, E., Zullo, A., Tomao, L., Romiti, A., Russo, G.L. and Tomao, S. (2014). Focus on genetic and epigenetic events

of colorectal cancer pathogenesis: implications for molecular diagnosis. *Tumor Biology*, vol. 35, no. 7, pp. 6195–6206.

Zucker, S. and Vacirca, J. (2004). Role of matrix metalloproteinases (MMPs) in colorectal cancer. *Cancer and Metastasis Reviews*, vol. 23, no. 1-2, pp. 101–117.

Zuo, Z., Hu, H., Xu, Q., Luo, X., Peng, D., Zhu, K., Zhao, Q., Xie, Y. and Ren, J. (2020). BBCancer: an expression atlas of blood-based biomarkers in the early diagnosis of cancers. *Nucleic Acids Research*, vol. 48, no. D1, pp. D789—-D796.