Using Latent Process Decomposition to Classify Prostate and Colorectal Cancers

Christopher Ellis

This thesis is submitted for the degree of

Doctor of Philosophy

University of East Anglia

School of Computing Sciences

April 2021

I would like to dedicate this thesis to my loving family . . .

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this thesis are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This thesis is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Christopher Ellis

April 2021

Acknowledgements

I am grateful to my supervisory team: Prof Vincent Moulton, Professor in Computational Biology, UEA; Prof Colin Cooper, Professor of Cancer Genetics, UEA and Dr Dan Brewer, Senior Bioinformatician, UEA. It is thanks to their vision and guidance that I have been able to work on such an exciting project.

I am indebted to Dr Bogdan Luca, UEA graduate, for his continued advice and assistance. The success of his own project was one of the main driving forcing for initiating my own project. I am therefore grateful for the time he has spent explaining his own work to me.

I am thankful to Dr Jeremy Clark, Senior Research Associate, UEA and Dr Rachel Hurst, Senior Research Associate, UEA and all other members of the Cancer Research group for their insights and encouragement during our regular meetings and social trips.

Finally I would like to thank my friends and family for their love and continued support. You have made my time at UEA one to remember and enjoy and have always been there for me during the tough times too. I owe you all much and cannot thank you enough.

Nevertheless, thank you.

Abstract

Cancer classification plays an important role in the clinical management of cancer patients. It enables clinicians to predict how individual cancers will behave and directs the best course of treatment. However, the classification of heterogeneous cancers has proven to be challenging.

To address this problem more advanced classification techniques should be used. In this thesis we focus on the unsupervised Bayesian algorithm Latent Process Decomposition (LPD). This technique has previously been used to classify breast cancer and was recently used to produce a novel classification of prostate cancer. We therefore aim to leverage LPD's ability to classify heterogeneous diseases.

We begin by performing a study on the prostate cancer subtype DESNT, introduced by Luca et al. (2017). By creating and applying a new type of LPD algorithm (OAS-LPD) to the DESNT classification, we establish a DESNT risk score that is an independent predictor of progression alongside existing diagnostic variables (PSA level and Gleason score). DESNT's expression profile is also demonstrated to be detectable in prostate cancer biopsies. Combined, these findings present the possibility for a new clinical test to reduce the over treatment of prostate cancer patients.

In the second part of this thesis we apply LPD to six transcriptome datasets obtained from colorectal cancer (CRC) biopsies. We identify and characterise four new CRC subtypes present across the datasets, including one subtype (designated Pericol) associated with a statistically significant poorer prognosis. Many of the Pericol signature genes are shown

to overlap with other published signatures and the Pericol risk score is identified as an independent predictor of disease recurrence.

Our results demonstrate the existence of poor prognosis categories of human cancers that can be used to assist in the targeting of treatment. They also emphasise the importance of employing biologically appropriate techniques to classify heterogeneous diseases.

Table of contents

List of figures xiv

List of tables xxiv

Abbreviations/Acronyms xxvii

1 Introduction **1**

1.1 Thesis Aims 3

1.2 Chapter Summaries 4

2 Biomedical Background **6**

2.1 Summary 6

2.2 Transcription / Translation 6

2.3 Cancer 8

2.3.1 Mutations 10

2.3.2 Chromosomal Abnormalities 11

2.3.3 DNA Methylation 13

2.4 Tissue Samples and Cell Cultures 14

2.5 Microarrays 16

2.5.1 Exon Microarrays 18

2.6 Discussion 19

3 Computational Background 20

3.1 Summary 20

3.2 Data Normalisation 20

3.2.1 Quantile Normalisation 21

3.2.2 Robust Multiarray Analysis (RMA) 22

3.2.3 ComBat 23

3.2.3.1 Step 1: Data Standardisation 24

3.2.3.2 Step 2: Empirical Bayes Batch Effect Parameter Estimation 25

3.2.3.3 Step 3: Data Adjustment for Batch Effects 25

3.3 Clustering Methods 25

3.3.1 K-means Clustering 26

3.3.2 Topic Models 27

3.3.3 Latent Process Decomposition (LPD) 28

3.3.3.1 Parameter Estimation 31

3.3.3.2 MLE 32

3.3.3.3 MAP 34

3.3.4 One Added Sample LPD (OAS-LPD) 35

3.4 Survival Analysis 35

3.4.1 Kaplan-Meier (KM) Survival Curves 36

3.4.2 Log-rank Test 39

3.4.3 Cox Proportional Hazard (PH) Model 40

3.5 Pathway Analysis 42

3.6 Discussion 43

4 The Prostate and Prostate Cancer 44

4.1 Summary 44

4.2 The Prostate 44

4.3 Prostate Cancer 45
4.3.1 Risk Factors 45
4.3.2 Screening and Early Detection:
The Problems with Prostate Specific Antigen Testing 47
4.3.3 Diagnosis 48
4.3.4 Classification criteria 49
4.3.4.1 Gleason Score 49
4.3.4.2 Tumour Node Metastasis 50
4.3.4.3 ICGC risk stratification 51
4.3.5 Localised Disease and Treatment 52
4.3.5.1 Active Surveillance 53
4.3.5.2 Brachytherapy 53
4.3.5.3 Radiotherapy 53
4.3.5.4 Prostatectomy 54
4.3.5.5 Biochemical Reoccurrence (BCR) 54
4.3.5.6 Androgen Deprivation Therapy (ADT) 55
4.3.6 Metastatic Disease and Castration Resistant Prostate Cancer (CRPC) 56
4.4 Discussion 56

5 Towards the Analysis of DESNT in Prostate Cancer Patient Samples 57
5.1 Summary 57
5.2 Materials 58
5.3 Producing the DESNT LPD Analysis 59
5.3.1 Choosing LPD Parameters 60
5.3.2 LPD Classification 62
5.3.3 Survival Analyses 62
5.3.3.1 Univariate Survival Analysis 64

5.3.3.2 Multivariate Survival Analysis 65
5.3.4 Differentially Expressed Genes 68
5.3.5 Pathway Analysis 69
5.4 DESNT as a Continuous Variable 71
5.5 Biopsy DESNT Analyses 75
5.5.1 Biopsy Samples 75
5.5.2 Applying the DESNT OAS-LPD model 76
5.6 Discussion 79

6 Colorectal Cancer (CRC) 81
6.1 Summary 81
6.2 The Colon 81
6.3 Colorectal Cancer 83
6.3.1 Risk Factors 83
6.3.2 Screening and Early Detection 84
6.3.3 Diagnosis 85
6.3.4 Classification criteria 85
6.3.4.1 Tumour Node Metastasis 86
6.3.4.2 Dukes' Staging 89
6.3.5 Localised and Regional Disease Treatment 90
6.3.5.1 Surgical Resection 90
6.3.5.2 Radiotherapy and Chemotherapy 91
6.3.6 Metastatic Disease Treatment 92
6.3.7 Genetic Alterations and Biomarkers 93
6.3.7.1 Hereditary CRC 95
6.3.7.2 Sporadic CRC 96
6.4 Discussion 97

7 Deriving Molecular Subtypes in Colorectal Cancer 100

7.1 Summary 100

7.2 Materials 101

7.2.1 Datasets 101

7.2.2 Clinical Data 102

7.2.3 Dataset Pre-processing 104

7.3 Creating the LPD Models 106

7.3.1 Choosing LPD Parameters 106

7.3.2 Representative LPD Classification 108

7.4 Analysing the LPD Models 110

7.4.1 Comparing LPD Process Survival 111

7.4.2 Identifying Conserved Processes 113

7.5 Developing a Consensus OAS-LPD Model 115

7.6 Analyses of the CRC Subtypes 118

7.6.1 Novel CRC Subtypes' Clinical Associations 118

7.6.2 Novel CRC Subtypes' Differentially Expressed Genes 120

7.6.3 Novel CRC Subtype Pathways 123

7.6.4 Intersection of Pericol Genes and Published Signatures 124

7.6.5 Methylation of Genes in Pericol 126

7.7 Pericol, a Continuous Predictor of Recurrence? 127

7.8 Discussion 129

8 Conclusions and Future Work 132

8.1 Summary 132

8.2 Prostate Cancer - DESNT 132

8.2.1 Biochemical Risk Assessment 132

8.2.2 OAS-LPD Classification of Biopsy Samples 134

8.3 LPD and Consensus OAS-LPD Classification of Colorectal Cancer 135
8.4 Consensus Molecular Subtypes 137
8.5 Development of Clinical Tests 138
8.6 Improved Versions of LPD 140
8.7 Conclusion 142

References **143**

Appendix A Appendix A **167**

A.1 Discrete Multivariate Cox PH Models 167
A.2 Gene Expression Levels 169
A.3 DESNT Over-Represented Pathways 169
A.4 Discretised Proportional DESNT Assignment Kaplan-Meier Curves 171

Appendix B Appendix B **173**

B.1 LPD Models Normalised Without TCGA Samples 174
B.2 Colorectal Cancer LPD Densities 177
B.3 Colorectal Cancer Representative LPD Models: Gamma Barplots 180
B.4 Colorectal Cancer Representative LPD Models: Kaplain Meier Plots 184
B.5 Colorectal Cancer Differentially Expressed Genes 186
B.5.1 LPD A DEGs 186
B.5.2 LPD B DEGs 187
B.5.3 LPD C DEGs 190
B.5.4 Pericol DEGs 190
B.6 Colorectal Cancer Subtype Pathways 194
B.6.1 LPD A Enriched Pathways 194
B.6.2 LPD B Enriched Pathways 197
B.6.3 Pericol Enriched Pathways 204

B.7 CRC Differential Methylation 222
B.8 CRC Over-Represented Hyper/Hypo-Methylated Pathways 258

List of figures

1.1 The number of patients diagnosed with the fifteen most prevalent cancer groups in England in 2017. Adapted from ONS [1]. 2

2.1 A simple depiction of RNA splicing. 7

2.2 A simple depiction of mRNA being translated into a sequence of amino acids. 8

2.3 Chromosomal Abnormalities: a) reciprocal translocation, b) inversion, c) insertion, d) chromoplexy, e) duplication, f) deletion. 12

2.4 Processing FFPE tissue samples. **a)** Tissue sample is serial sectioned. **b)** Sectioned tissue is placed in a plastic cassette to be processed. The tissue is fixed and embedded in paraffin. **c)** After processing, blocks are formed. Each block consists of tissue embedded in paraffin that is attached to the bottom of the cassette. **d)** A microtome is used to slice the block into slices (typically 4 microns thick) that can be mounted and analysed. Adapted from Lester (2010) [2]. 15

3.1 A graphical representation of an LPD model adapted from Rogers et al. [3]. Each circle corresponds to a variable, the dark circles represent hidden variables, while the empty circles show observed variables. The arrows represent conditional dependencies between variables. 29

3.2 A KM plot calculated for the data in Table 3.3. The thin crosses represent observed events that have been censored. 37

4.1 A diagram of the location of a prostate and the four prostate zones. Adapted from AcademLib [4]. 45

5.1 The log-likelihood against the number of processes using the MLE solution (red curve) and the MAP solution (blue curve) for the MSKCC dataset. The points represent the mean log-likelihood from 100 LPD restarts. Error bars for each point are also provided to demonstrate the distribution of log-likelihoods across the LPD restarts. 61

5.2 A bar chart showing the output from an LPD run, using the MSKCC dataset. Each bar within a process (row) represents the proportion for which that sample was associated with that process (σ). The colour of each bar represents the ICGC category assigned to each sample within the associated clinical data. 63

5.3 Kaplan-Meier survival curves for the eight LPD groups created from the MSKCC dataset, using BCR failure as the event. The number of cancer samples in each group is indicated at the bottom right corner, alongside the number of BCR failures in parentheses. 64

5.4 Kaplan-Meier survival curves comparing DESNT and non-DESNT groups for the MSKCC dataset, using BCR failure as the event. The number of cancer samples in each group is indicated at the bottom right corner, alongside the number of BCR failures in parentheses. 65

5.5 Results from the multivariate Cox PH models, using the MSKCC (A), CancerMap (B), CamCap (C), Stephenson (D) datasets and a combination of the previous four datasets (E). The blue markers denote the hazard ratio for each covariate and the extended bars denote the 95% confidence interval. The log-rank p-value for each covariates' hazard ratio is listed on the right side of the figure. PSA level was split on $\leq / > 10$, Gleason score was split on $\leq / > 7$ and DESNT γ was treated as a continuous variable between 0 to 1. 67

5.6 **A)** A venn diagram of the number of differentially expressed genes in the DESNT group compared to the non-DESNT groups, across the MSKCC, CancerMap, Stephenson and Klein datasets, that were present in at least 80% of the LPD restarts. **B)** The differentially expressed genes in the DESNT group compared to the non-DESNT groups, across the MSKCC, CancerMap, Stephenson and Klein datasets, that were present in at least 20%, 40%, 60% and 80% of the LPD restarts. 69

5.7 **A)** Bar chart showing the variable DESNT γ associations from the representative LPD run for the MSKCC dataset. **B)** Pie charts showing how varied the γ associations are for the range of samples highlighted in Figure 5.7-a that are not DESNT dominant. **C)** Pie charts showing how varied the γ associations are for the range of samples highlighted in Figure 5.7-a that are DESNT dominant. Published in Luca et al. (2020) [5] 72

5.8 **A)** An ordered barchart showing the DESNT γ of every sample used in the accompanying Kaplan-Meier survival plot. **B)** A Kaplan-Meier survival plot using all unique samples from the MSKCC, CancerMap, CamCap and Stephenson datasets, split using the four proportional assignment groups. Published in Luca et al. (2020) [5]. 73

5.9 Cox PH models for the combined prostate cancer dataset, formed from the unique patients in the MSKCC, CancerMap, CamCap and Stephenson datasets, where duplicate patients were removed randomly. The blue markers indicate the hazard ratio for each covariate and the extended bars represent the 95% confidence interval. The log-rank p-value for each covariate is displayed on the right side of the figure. **A)** The covariates were all discretised. The base case for each of the Group variables was $\gamma < 0.001$. Samples were assigned to Group 2, 3 and 4 in the range $0.001 \leq \gamma < 0.3$, $0.3 \leq \gamma < 0.6$ and $0.6 \leq \gamma$ respectively. PSA was split on $(\leq / > 10)$ and Gleason was split on $(\leq / > 7)$. **B)** DESNT represents the continuous range of DESNT γ from 0 - 1. PSA was split on $(\leq / > 10)$ and Gleason was split on $(\leq / > 7)$. . . 74
5.10 Boxplots showing the mean and 95% confidence intervals for the 20 normalised biopsy samples and a random selection of 60 normalised samples from the MSKCC dataset, due to the limited space on the page. 77
5.11 Bar plots showing the LPD γ values for the association between each biopsy sample and OAS-LPD process. The OAS-LPD process primarily associated with each biopsy sample has also been highlighted, in addition to the Gleason Grade of each given biopsy. 77
5.12 Scatter-plot comparing OAS-LPD DESNT γ and Gleason grade for 20 prostate cancer biopsy samples. The blue line denotes the Pearson's correlation and the shaded region the 95% confidence region. 78

6.1 A diagram detailing the sections of a colon. Adapted from Mayo Clinic [6]. 82
6.2 A diagram detailing the sporadic CRC molecular event pathways. Adapted from Szylberg et al. (2015) [7]. 98

7.1 Boxplots depicting 20 random normalised samples from each of the GSE14333plus, GSE39582, GSE41258 and GSE81653 datasets. 105

7.2 Figure depicting the log-likelihoods of each parameter combination for the GSE14333plus dataset. **a)** The log-likelihood plateau. **b-d)** The three groups of input parameters selected for further analysis. 107

7.3 Figure depicting the identification of a representative LPD run, based on the density of *p*-values from a set of 100 log-rank tests, each performed on an individual LPD model using the GSE39582 dataset. The model with the shortest *p*-value distance to the modal density was selected as the representative LPD run. 109

7.4 Figure depicting the identification of a representative LPD run, based on the density of *p*-values from a set of 100 log-rank tests, each performed on an individual LPD model using the TCGA-COAD dataset. The model with the shortest *p*-value distance to the modal density was selected as the representative LPD run. 110

7.5 Figure depicting the LPD γ values (association between a sample and a process) for each LPD process in the GSE39582 representative run. Samples have been grouped by their process with the greatest γ value for ease of viewing. 111

7.6 Kaplan-Meier survival curves showing the disease-free survival of the the six LPD processes from the representative run for the GSE39582 dataset. The total number of samples (with DFS information) for each process is shown in the bottom right, with the number of DFS events displayed in brackets. . 113

7.7 a) Correlation map where each line represents a statistically strong positive correlation between two processes from independent representative models. b) An example of the correlations between all four microarray and TCGA-COAD based models for the Pericol colorectal subtype. 114

7.8 Kaplan-Meier survival plot showing the disease-free survival of the four common processes from the representative LPD runs for the GSE14333plus, GSE39582 and GSE41258 datasets. The total number of samples (with DFS information) for each process is shown in the bottom right, with the number of DFS events displayed in brackets. 115

7.9 Kaplan-Meier survival plot showing the disease-free survival of the four common processes from the consensus OAS-LPD models. The total number of samples (with DFS information) for each process is shown in the bottom right, with the number of DFS events displayed in brackets. 117

7.10 CRC subtype clinical associations. The green up arrows highlight the over-representation of a given clinical factor, while the red down arrows highlight the factors as under-represented. 119

7.11 Venn diagrams showing the intersection of differentially expressed genes across each representative model in our four CRC subtypes. a) LPD A. b) LPD B. c) LPD C. d) Pericol 122

7.12 Circos plot highlighting the genes shared between any two signatures. Genes shared with Pericol are highlighted red. 126

7.13 a) Mean Pericol γ taken from all four OAS LPD models for each sample in the GSE14333plus, GSE39582 and GSE41258 datasets, coloured according to the four discrete Pericol γ groups. b) Kaplan-Meier survival curves for the discretised Pericol γ groups. 128

A.1 Results from the multivariate Cox PH models, using the MSKCC (A), CancerMap (B), CamCap (C), Stephenson (D) datasets and a combination of the previous four datasets (E). The blue markers denote the hazard ratio for each covariate and the extended bars denote the 95% confidence interval. The log-rank p-value for each covariates' hazard ratio is listed on the right side of the figure. PSA level was split on $\leq / > 10$, Gleason score was split on $\leq / > 7$ and DESNT was split on non-DESNT/DESNT membership. 168

A.2 A heatmap depicting the gene expression levels of the 500 genes, used in the LPD classification process, for the CancerMap, Stephenson, Klein and MSKCC datasets. 169

A.3 Kaplan-Meier survival curves comparing the discretised DESNT γ groups for the MSKCC (A), CancerMap (B), CamCap (C), Stephenson (D) and merged dataset (E), using BCR failure as the event. The number of cancer samples in each group is indicated at the bottom right corner, alongside the number of BCR failures in parentheses. 172

B.1 Figure depicting the LPD γ values (association between a sample and a process) for each LPD process in the GSE14333plus representative run when normalising the data without TCGA samples. Samples have been coloured by their Dukes' stage assignment. 174

B.2 Figure depicting the LPD γ values (association between a sample and a process) for each LPD process in the GSE39582 representative run when normalising the data without TCGA samples. Samples have been coloured by their Dukes' stage assignment. 175

B.3 Figure depicting the LPD γ values (association between a sample and a process) for each LPD process in the GSE41258 representative run when normalising the data without TCGA samples. Samples have been coloured by their Dukes' stage assignment. 176
B.4 Figure depicting the LPD γ values (association between a sample and a process) for each LPD process in the GSE81653 representative run when normalising the data without TCGA samples. 177
B.5 Figure depicting the identification of a representative LPD run, based on the density of p-values from a set of 100 log-rank tests, each performed on an individual LPD model using the GSE14333plus dataset. The model with the shortest p-value distance to the modal density was selected as the representative LPD run. 178
B.6 Figure depicting the identification of a representative LPD run, based on the density of p-values from a set of 100 log-rank tests, each performed on an individual LPD model using the GSE41258 dataset. The model with the shortest p-value distance to the modal density was selected as the representative LPD run. 178
B.7 Figure depicting the identification of a representative LPD run, based on the density of p-values from a set of 100 log-rank tests, each performed on an individual LPD model using the GSE81653 dataset. The model with the shortest p-value distance to the modal density was selected as the representative LPD run. 179
B.8 Barplot showing the γ values (Bayesian association) for each sample with each LPD process, in the LPD model built using the GSE14333plus dataset. 180
B.9 Barplot showing the γ values (Bayesian association) for each sample with each LPD process, in the LPD model built using the GSE39582 dataset. . . 181

B.10 Barplot showing the γ values (Bayesian association) for each sample with each LPD process, in the LPD model built using the GSE41258 dataset. . . 182
B.11 Barplot showing the γ values (Bayesian association) for each sample with each LPD process, in the LPD model built using the GSE81653 dataset. . . 183
B.12 Kaplan-Meier survival curves showing the disease-free survival of the the six LPD processes from the representative run for the GSE14333plus dataset. The total number of samples (with DFS information) for each process is shown in the bottom right, with the number of DFS events displayed in brackets. 184
B.13 Kaplan-Meier survival curves showing the disease-free survival of the the six LPD processes from the representative run for the GSE39582 dataset. The total number of samples (with DFS information) for each process is shown in the bottom right, with the number of DFS events displayed in brackets. . 185
B.14 Kaplan-Meier survival curves showing the disease-free survival of the the six LPD processes from the representative run for the GSE41258 dataset. The total number of samples (with DFS information) for each process is shown in the bottom right, with the number of DFS events displayed in brackets. LPD2 and LPD 3 are not shown on the figure as they only contain normal samples. 186
B.15 Barplot of the top 20 (ordered by *p*-value) GO pathways over-represented in LPD A. 196
B.16 Barplot of the KEGG pathways over-represented in LPD A. 197
B.17 Barplot of the Reactome pathways over-represented in LPD A. 197
B.18 Barplot of the top 20 (ordered by *p*-value) GO pathways over-represented in LPD B. 203
B.19 Barplot of the KEGG pathways over-represented in LPD B. 204

B.20 Barplot of the Reactome pathways over-represented in LPD B. 204

B.21 Barplot of the top 20 (ordered by *p*-value) GO pathways over-represented in Pericol. 221

B.22 Barplot of the top 20 (ordered by *p*-value) KEGG pathways over-represented in Pericol. 221

B.23 Barplot of the top 20 (ordered by *p*-value) Reactome pathways over-represented in Pericol. 222

List of tables

3.1 Table showing an example of the different Gaussian means of G genes in K=3 processes. 30
3.2 Table showing an example of the sample θ vectors containing the proportional assignment of each sample to K=3 processes. 30
3.3 An example of survival data, including the estimated survival probabilities. $\mathbf{t}_{(j)}$ is the survival time (in months) and $\mathbf{n}_{(j)}$ is the number of participants remaining at each survival time. While $\mathbf{m}_{(j)}$ and $\mathbf{q}_{(j)}$ represent the number of failures and censored events respectively, at each survival time. 38
3.4 An example of the steps involved in the log-rank statistic. 40

4.1 Tumour Node Metastasis (TNM) classification system. 51
4.2 ICGC risk categorisation of prostate cancer patients that have received radical prostatectomy. 52
4.3 D'Amico risk stratification for men with localised prostate cancer. 52

5.1 A summary of the prostate cancer datasets used in the LPD analysis. 59
5.2 A summary of the suitable parameters identified for each prostate cancer dataset. 62
5.3 A summary of the new Gleason grade groups. Epstein et al. (2016) [8]. . . 76
5.4 A summary of the prostate biopsy samples summarised into grade groups. . 76

6.1 Tumour Node Metastasis (TNM) classification system for CRC. 87
6.2 AJCC / TNM staging for CRC. 89
6.3 AJCC stages grouped by Dukes' stage for CRC. 89
6.4 Table showing the percentages of cases for each group of CIMP and MSI combinations. CIMP-H, CIMP-L and CIMP-0 refer to high, low and very low methylation levels respectively. MSI-H, MSI-L and MSS refer to high, low and no microsatellite instability respectively. Adapted from Ogino 2008 [9]. 94
6.5 Table highlighting the main CIMP-MSI group associations. Adapted from Ogino 2008 [9]. 94

7.1 Table summarising the unique samples from the datasets used in this chapter. 101
7.2 Table summarising the clinical data associated with the samples used in this chapter. 104
7.3 Table summarising the final model parameters for each dataset. 108
7.4 A summary of the OAS-LPD consensus assignments. 116
7.5 Table summarising the intersection of differentially expressed genes. 121

A.1 Top 20 GO pathways over-represented in the DESNT signature. 169
A.2 KEGG pathways over-represented in the DESNT signature. 170
A.3 Reactome pathways over-represented in the DESNT signature. 171

B.1 Differentially expressed genes within the colorectal cancer LPD A subtype . 186
B.2 Differentially expressed genes within the colorectal cancer LPD B subtype . 187
B.3 Differentially expressed genes within the colorectal cancer LPD C subtype . 190
B.4 Differentially expressed genes within the colorectal cancer Pericol subtype . 190
B.5 GO pathways over-represented in LPD A. 194
B.6 KEGG pathways over-represented in LPD A. 195

B.7 Reactome pathways over-represented in LPD A. 196
B.8 GO pathways over-represented in LPD B. 197
B.9 KEGG pathways over-represented in LPD B. 202
B.10 Reactome pathways over-represented in LPD B. 202
B.11 GO pathways over-represented in Pericol. 204
B.12 KEGG pathways over-represented in Pericol. 217
B.13 Reactome pathways over-represented in Pericol. 218
B.14 A differential methylation analysis performed on the TCGA-COAD methylation dataset using Limma [65] and methylGSA [251]. Results restricted to only include genes where the majority of CpG sites were significantly (adj p-value ≤ 0.05) hyper-methylated and the genes were under expressed, or genes where the majority of CpG sites were significantly (adj p-value ≤ 0.05) hypo-methylated and the genes were over expressed. 222
B.15 GO pathways over-represented in the TCGA-COAD methylation data. . . . 258

Abbreviations/Acronyms

Roman Symbols

ADT	Androgen deprivation therapy
AFAP	Attenuated FAP
AJCC	American Joiny Committee on Cancer
BCR	Biochemical recurrence
BPH	Benign prostatic hyperplasia
CAPOX	Cepecitabine and Oxaliplatin
cDNA	Complementary DNA
CIMP	CpG island methylator phenotype
CIN	Chromosomal instability
CMS	CIMP and MSI status
CNV	Copy-number variant
CpG	Cytosine is followed by guanine
CRC	Colorectal cancer

CRPC	Castration resistant prostate cancer
CRUK	Cancer Research UK
CT	Computed tomography
DAVID	Database for Annotation, Visualisation and Intergrated Discovery
DEG	Differentially expressed gene
DFS	Disease-free survival
DHT	Dihydrotestosterone
DNA	Deoxyribonucleic acid
DRE	Digital rectal examination
EAU	European Association of Urology
EB	Empirical Bayes
EGFR	Epidermal growth factor receptor
ESD	Endoscopic submucosal dissection
FAP	Familial adenomatous polyposis
FF	Fresh frozen
FFPE	Formalin-fixed-paraffin-embedded
FIT	Faecal immunochemical test
FOBT	Faecal occult blood test
FOLFOX	Oxaliplatin and Folinic acid

GO	Gene ontology
HIPEC	Hyperthermic intraperitoneal chemotherapy
HNPCC	Hereditary nonpolyposis colorectal cancer
HR	Hazard ratio
ICGC	International Cancer Genome Consortium
KEGG	Kyoto Encyclopedia of Genes and Genomes
KM	Kaplan-Meier
LDA	Latent Dirichlet allocation
LPD	Latent Process Decomposition
MAB	Maximum androgen blockade
MAP	Maximum a posteriori
MCL	Markov cluster algorithm
MLE	Maximum likelihood estimation
MM	Mismatch
MMR	Mismatch repair
MoAbs	Monocolonal antibodies
MP-MRI	Multi-parametric magnetic resonance imaging
mRNA	Messenger RNA
MSI	Microsatellite instability

MVB	Marginalised VB
NICE	National Institute for Health and Care Excellence
OAS-LPD	One-added-sample LPD
OMIM	Online Mendelian Inheritance in Man
PARP	Poly ADP-ribose polymerase
PCR	Polymerase chain reaction
PH	Proportional hazard
PLSA	Probabilistic latent semantic analysis
PM	Perfect-match
pre-mRNA	Precursor mRNA
PSA	Prostate specific antigen
qPCR	Quantitive PCR
RFS	Relapse-free surivival
RMA	Robust multiarray analysis
RNA	Ribonucleic acid
RP	Radical prostatectomy
RT-qPCR	Reverse transcription qPCR
SACT	Systemic anti-cancer therapy
SBRT	Stereotactic body radiation therapy

SIRT	Selective radiation therapy
SNV	Single-nucleotide variant
TAE	Transanal excision
TCGA	The Cancer Genome Atlas
TME	Total mesorectal excision
TNM	Tumour node metastasis
tRNA	Transfer RNA
TRUS	Transrectal ultrasound
TURP	Transurethral resection of the prostate
UICC	International Union for Cancer Control
VB	Variational Bayes
WT	Wild-type

Chapter 1

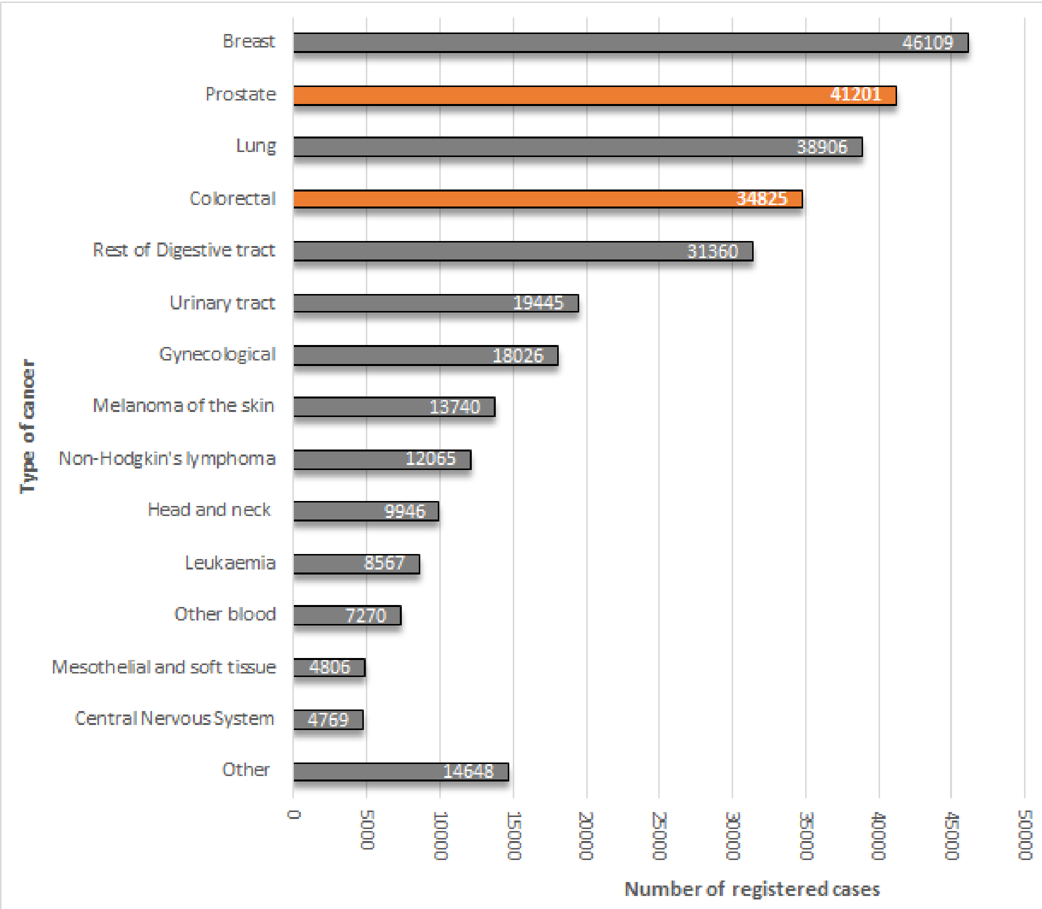
Introduction

The term cancer describes a group of diseases in which abnormal cells divide without control, invade nearby tissues and eventually spread to distant parts of the body [10]. Cancer continues to be one of the leading causes of death worldwide, accounting for 9.6 million deaths in 2018 alone [11]. Among the many types of cancer, colorectal and prostate cancers accounted for 24.9% of all new cancer cases within the UK in 2017 [1] (Figure 1.1).

The name of a cancer is typically derived from the location within the body in which it first develops. However, cancer is not a simple set of diseases and each type of cancer may contain many subtypes, formed through distinct molecular pathways with independent clinical outcomes. The prevalence of these diseases has resulted in many attempts to find cancer subcategories to facilitate the development of targeted diagnosis and treatment options. The identification of specific cancer subtypes could also highlight potential targets for the development of new drugs.

The classification of cancer therefore plays an important role in the diagnosis and treatment of cancer patients. It can prevent unnecessary radical treatments in low risk patients and ensure high risk patients receive the necessary treatment to improve their prognosis. This avoids the complications and side-effects associated with such treatments for low risk patients, while providing radical treatments to patients with the most to gain. The identification of high risk cancer classifications can focus research efforts into the areas with the greatest potential benefits for patients. These research efforts can also target the characteristics of each identified subtype to provide specialised treatment options for specific groups of patients [12].

Fig. 1.1 The number of patients diagnosed with the fifteen most prevalent cancer groups in England in 2017. Adapted from ONS [1].



An important example of the benefits of cancer classification is provided by the identification of five distinct molecular subtypes of breast cancer (normal breast-like, basal, luminal A, luminal B, and ERBB2+) [13]. Together the many breast cancer studies have determined epidemiological, histoclinical, molecular, prognostic and therapeutic features associated with each of these five subtypes [14]. One such key observation within the aggressive basal subtype is an association with germline BRCA1 mutations [15]. However, these mutations are not always present within the basal subtype and are less frequently observed in other subtypes [15]. The identification of BRCA1 mutations is as key step during diagnosis and treatment as these mutations confer a susceptibility to PARP inhibitors that can be used to improve the overall patient survival times [16].

While many molecular subtypes have been successfully identified within cancers such as breast cancer, this has proven to be far more challenging within prostate and colorectal cancers. However, a recent study in 2017 by the University of East Anglia managed to produce a novel set of classifications for prostate cancer, with clinically useful associations [17]. The main reason for this successful unsupervised classification has been attributed to their use of Latent Process Decomposition (LPD), which accounted for the heterogeneity within prostate cancer, optimised the number of clusters and avoided over-fitting the model to noise within the data. This unsupervised Bayesian method has also been applied to breast cancer, where it was previously able to identify four subtypes closely related to the molecular subtypes discussed above [18].

The application of LPD to other cancers may yield similarly promising results and provide new opportunities to tailor treatment options to specific cancer subtypes. Colorectal cancer is one such highly heterogeneous disease whose patients may benefit from the application of LPD to identify common molecular subtypes. While a number of colorectal classifications currently exist within the literature, including the clinically approved Oncotype DX test [19], they have been shown to be independently distinct from one another [20]. This discordance suggests the current classifications are limited by a lack of robustness and that further work is required to reliably classify the disease at a molecular level.

1.1 Thesis Aims

The aim of this thesis is to develop novel classifications of heterogeneous diseases, focusing on prostate and colorectal cancers, that can be used to stratify patients into high and low risk groups. To accomplish this we make use of the unsupervised Bayesian classifier called Latent Process Decomposition. We split the thesis into two separate pieces of work that each focus on the application of LPD within one of these two types of cancer.

In the first part of this thesis we shall focus on prostate cancer. In particular, we extend upon the work by Luca et al. [17] to explore the potential uses for the poor prognosis subtype known as DESNT. We modify the LPD algorithm to classify new samples into the LPD processes of an existing model (such as DESNT) and show the feasibility of applying this transcriptomic model to prostate biopsy samples.

In the second part of this thesis we will present a set of novel classifications of colorectal cancers, containing both good and poor prognosis groups based on their molecular subtypes. We will characterise each of the subtypes using their transcriptomic signatures and accompanying clinical data. Clinically relevant correlations to these subtypes will be shown in additional to an analysis of the genetic pathways critical to the development of these colorectal cancer subtypes.

1.2 Chapter Summaries

We now summarise the contents of the rest of this thesis, including a summary of my contributions:

- In **Chapter 2** we introduce the biological principles key to this thesis. This will include the defining characteristics of cancer and the technologies used to quantify transcriptomes.

- In **Chapter 3** we introduce the computational approaches used to classify cancer. This chapter will predominately focus on Latent Process Decomposition, as it has shown a strong ability to classify heterogeneous cancers. We will also describe the algorithms used to analyse the survival times of patients.

- In **Chapter 4** we discuss the defining features of prostate cancer and current methods to quantify patient risk. We conclude this chapter with a discussion focused on the prostate cancer treatment options and the limitations of current tests.

- In **Chapter 5** we apply a Latent Process Decomposition algorithm to five prostate cancer datasets, as described in Luca et al. (2017) [17]. We then extend this work to further analyse the importance of the DESNT prostate cancer subtype in relation to determining the risk of recurrence in patients. We end the chapter by performing a preliminary study aiming to detect DESNT in prostate cancer biopsies using a novel type of LPD. My contribution to new analyses and results within this chapter include the analysis of DESNT as a continuous predictor of biochemical recurrence and the classification of prostate biopsy samples using OAS-LPD.

- In **Chapter 6** we discuss the defining features of colorectal cancer, highlighting the differences between hereditary and sporadic forms of the cancer. We also consider the range of transcriptomic factors known to influence the progression of colorectal cancer and explore the available treatment options.

- In **Chapter 7** we produce a new classification framework for colorectal cancer by applying LPD to six transcriptome datasets. We then use these LPD models to construct a consensus OAS-LPD model and examine the clinical differences between each of the consensus subtypes. We identify a poor prognosis subtype capable of predicting the risk of disease relapse and show LPD's ability to derive subtypes similar to those found in the current literature. My contribution to new analyses in this chapter extends to all work presented.

- In **Chapter 8** we conclude the findings of this thesis with a discussion on some of the possible future directions for this research.

Chapter 2

Biomedical Background

2.1 Summary

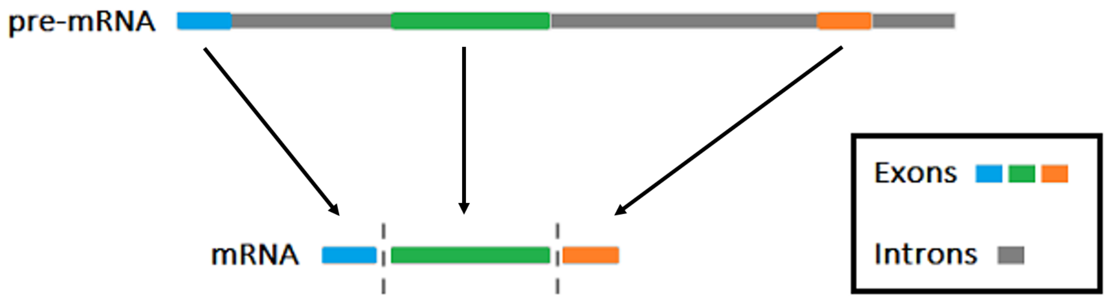
In this chapter we present the main biological concepts related to cancer. In later chapters we build upon this information to describe the medical approaches related to the clinical management of prostate and colorectal cancers. We begin by describing the basic molecular biology surrounding the formation of cancerous tissue. We then introduce the technological approach (microarrays) used to produce the datasets analysed in later chapters of this thesis.

2.2 Transcription / Translation

Organisms store genetic information in DNA (deoxyribonucleic acid), a double stranded, helical structure, composed of chains of nucleotides. These nucleotides contain a phosphate group, a sugar group and one of four nitrogen bases (Adenine, Cytosine, Guanine and Thymine). The central dogma of molecular biology states that DNA is *transcribed* into RNA (ribonucleic acid), which is then *translated* into proteins [21]. These proteins form the structural and functional elements of an organism's cells.

A *gene* is a segment, or multiple segments, of DNA that codes for a given protein. The process of synthesising a protein from its coding gene is referred to as *gene expression*. More specifically these genes are transcribed into *pre-mRNA* (precursor messenger RNA). The pre-mRNA is then *spliced*, a process where some portions (introns) are removed and the remaining portions (exons) are ligated together to form *mRNA* (messenger RNA), as shown in Figure 2.1. The selection of exons can be changed to produce alternative mRNA transcripts.

Fig. 2.1 A simple depiction of RNA splicing.

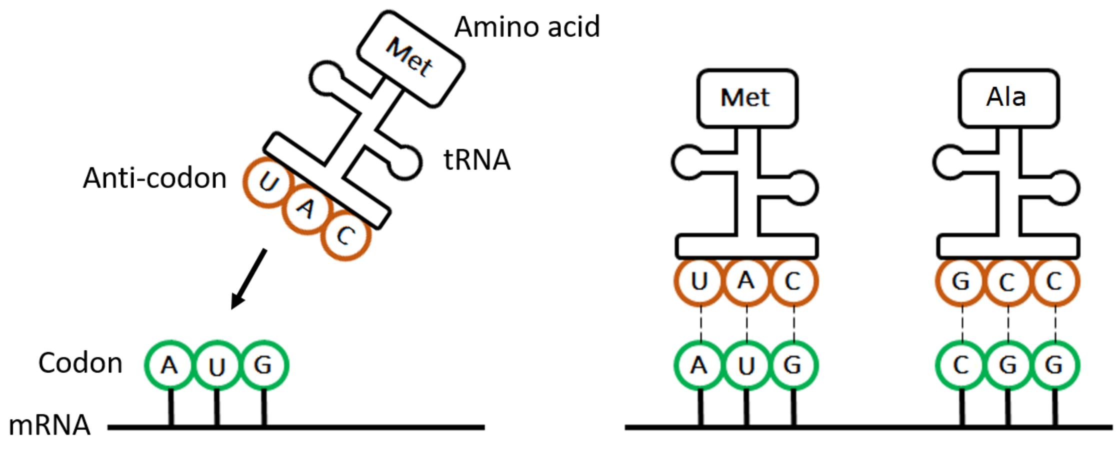


The mRNA nucleotide sequence is then translated into a sequence of amino acids within a cellular structure called a *ribosome*. The strand of mRNA is split into sequences of three nucleotides (codons). Amino acids, attached to tRNA (transfer RNA), are then arranged into the order corresponding to the order of codons by matching each mRNA codon with the anti-codon on the tRNA molecules (Figure 2.2).

The codon responsible for initialising translation is AUG. Upstream of the initialisation codon is the 5' untranslated region (5'UTR) of mRNA responsible for regulating translation. Translation is terminated by one of three stop codons (UAG, UAA, UGA), which precede the 3' untranslated region (3'UTR). Within the 3'UTR exist regions that commonly influence the subsequent gene expression.

The quantity of each protein produced by translation is thought to be determined by the amount of RNA. In practice there is a poor correlation between levels of mRNA and proteins [22]. The quantity of RNA can vary over time, based on many external stimuli and

Fig. 2.2 A simple depiction of mRNA being translated into a sequence of amino acids.



internal needs. The amount of mRNA transcribed is referred to as the *gene expression level*. We can use gene expression levels as a proxy of what mechanisms are functionally altered within a cell. While there has been a great deal of progress in understanding the mechanisms that control the amount of DNA transcribed, such as the need for transcription factors to bind to *promoters* (a region of DNA upstream of the transcription sequence) and *distal enhancers* (regions within the noncoding DNA that stimulate transcription), it is still not fully understood [23]. For the purposes of this thesis we will be investigating the gene expression levels in cancer samples and the potential epigenetic effects (non-genetic influences) that could explain these changes in gene expression levels.

2.3 Cancer

Cancer is a disease of the genome, consisting of many individual diseases that are characterised by uncontrolled cell division. The progression from a normal cell to a cancerous cell is a multi-stage process known as tumorigenesis. Normal cells contain numerous safe guards against uncontrolled division, which must be disabled or bypassed to become a cancer cell. Over time cells can accumulate numerous mutations, inheriting the mutations from previous cell generations and developing new mutations alike. By accumulating a sufficient number of mutations to disable and bypass the various safe guards, a cell may begin to divide uncontrollably.

Hanahan and Weinberg (2000) [24] described six essential capabilities that cancer cells must acquire to multiply and spread: self-sufficiency in growth signals, insensitivity to growth-inhibitory signals, evasion of programmed cell death (apoptosis), limitless replicative potential, sustained angiogenesis and tissue invasion and metastasis.

The first four capabilities are essential to begin the formation of cancers, without them the cells would neither divide uncontrollably or survive once they did begin to multiply. Cancers that develop within solid tissue can create tumours. As these tumours begin to grow in size they require a steadily increasing supply of oxygen, nutrients and waste removal [25]. To accomplish this the tumours usually hijack the mechanisms responsible for angiogenesis, to spread new blood vessels from existing ones [26].

As a tumour progresses into the latter stages it begins to invade the surrounding tissue and ultimately spreads to new distant sites, developing metastatic tumours. A metastatic tumour is typically the result of cancerous cells spreading into the blood, which transports them to a distant site. This process is known as metastasis and is the main cause of cancer-related death as a result of multiple organ failure.

More recently Hanahan and Weinberg (2011) [27] presented two additional emerging-hallmarks of cancer: reprogramming of energy metabolism and evading immune destruction. They first highlight the observation that many cancer cells limit their energy metabolism to glycolysis (the anaerobic breakdown of glucose into pyruvic and lactic acids) [28], irrespective of the presence of oxygen. Initially this appears to be counter-intuitive given the 18-fold lower efficiency of ATP production by glycolysis, compared to mitochondrial oxidative phosphorylation [29]. However, the lactate produced via the glycolytic pathway can be utilised by the surrounding cancer cells as part of the citric acid cycle to provide another source of energy [30]. Additionally, the glycolytic intermediates can be used in multiple biosynthetic pathways including the production of amino acids [31]. The products of these biosynthetic pathways can in turn be utilised in the assembly of new cells to support cell proliferation.

The second emerging-hallmark, evading immune destruction, highlights the ability for some cancer cells to avoid detection by the immune system. Better prognosis has been observed in several forms of human cancer where immune cells have heavily infiltrated the solid tumours, such as colorectal and ovarian tumours [32, 33], while increased incidence has been observed in immunocompromised patients [34]. However, cancer cells may also prevent the immune system from killing these cells in individuals with normal immune systems by utilising immunosuppressive factors, such as TGF-β, to suppress the actions of infiltrating immune cells.

2.3.1 Mutations

A genetic mutation is the alteration to a sequence of nucleotides within a portion of DNA. When a single nucleotide is substituted the genetic mutation is referred to as a point mutation. These mutations are the result of external factors such as radiation, ultraviolet light or chemicals, or endogenetic factors such as errors in DNA repair. When mutations take place within a protein coding gene, the protein produced by this mutated gene may also change. The changes to the protein may in turn result in detrimental effects to its ability to perform its normal function(s).

There are a number of mutations that play an important role in the development of cancers. Mutations to the tumour suppressor gene *TP53* occur in up to 50% of tumours depending on the type of cancer, making it one of the most common mutations within cancer [35]. Mutations to this gene commonly result in its inability to initiate apoptosis, or to activate DNA repair proteins, preventing it from stopping the formation of cancers.

While TP53 is commonly mutated in many different types of cancer, other genes have been found to be mutated in a vast number of tumours attributed to one or more specific types of cancer. Examples of this are the *APC* tumour suppressor gene that has been strongly associated with colorectal cancers [36] and *BRCA1* / *BRCA2* that have been associated with an increased risk of both Prostate and Breast cancers [37, 38].

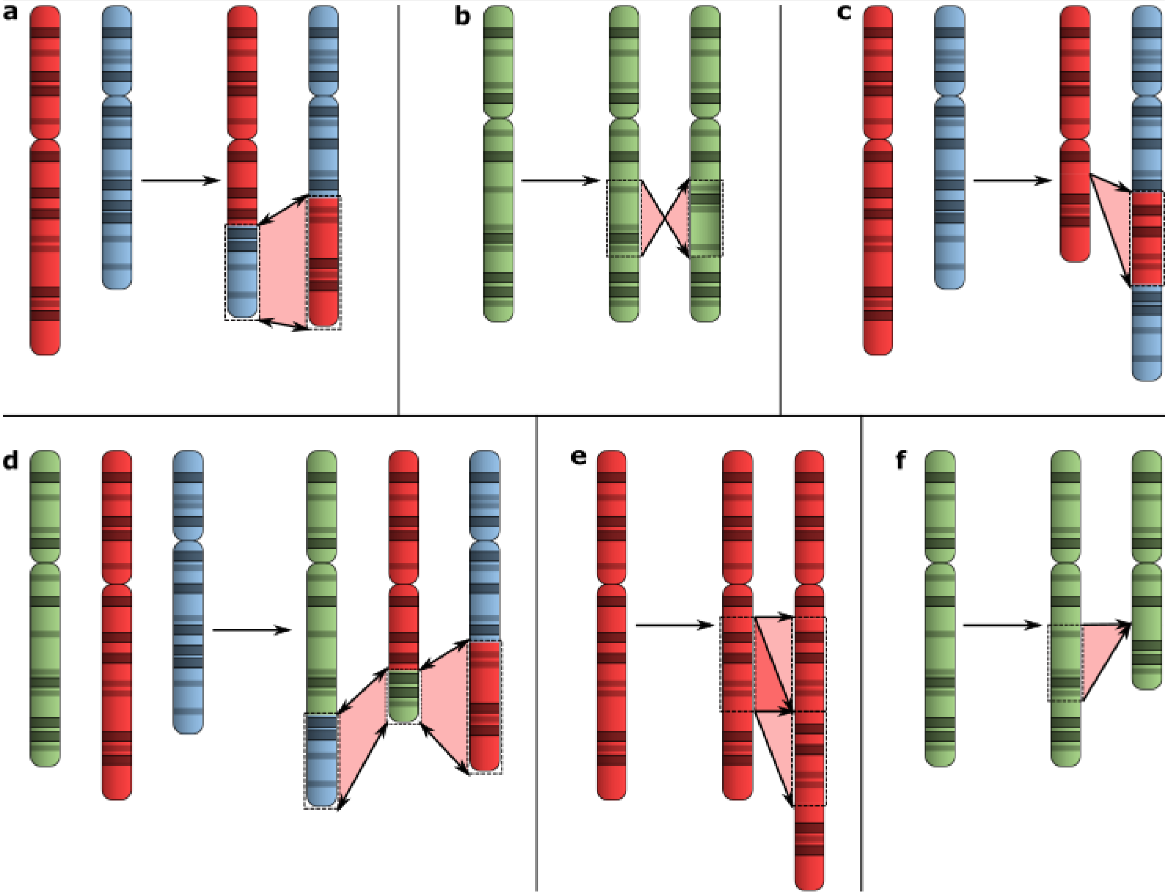
2.3.2 Chromosomal Abnormalities

Chromosomal abnormalities are a class of genetic alteration that can result in either an increase or decrease to the number of chromosomes in a patient. Alternatively they can result in a change to the structure of a patient's chromosomes. These abnormalities can be found in almost all major tumour types and are split into two main subclasses: *balanced chromosomal rearrangements* and *chromosomal imbalances*. [39]

Chromosomal rearrangements change the structure of a chromosome without affecting the number of copies of a gene. They consist of *reciprocal translocations* (two chromosomes exchanging portions), *inversions* (a portion of a chromosome is inverted) and *insertions* (a portion of a chromosome is inserted into another) [39] as shown in figure 2.3a-c. A more complex form of chromosomal rearrangement coined *chromoplexy* also exists where multiple inter translocations occur simultaneously between multiple chromosomes (Figure 2.3d) [40, 41].

The breakpoints of chromosomal rearrangements often occur within a gene transcript or within proximity to the promoter region. In these cases the rearrangement may result in a *gene fusion*, where parts from two distinct genes form a new chimeric gene with new or altered functionality [39]. These chimeric genes may no longer respond to regular control mechanisms, or produce proteins that are non-responsive. This behaviour can be seen in almost all cases of chronic myeloid leukaemia where the *BCR* gene located on chromosome

Fig. 2.3 Chromosomal Abnormalities: a) reciprocal translocation, b) inversion, c) insertion, d) chromoplexy, e) duplication, f) deletion.



22 and *ABL1* gene located on chromosome 9 form a chimeric gene following reciprocal translocation [42].

Chromosomal imbalance refers to abnormalities that arise through the loss or gain of genetic material. This can occur to a portion within a chromosome or to the entire chromosome. These duplications or deletions (Figure 2.3e,f) directly affect the production of proteins associated with the genes and regulation of the pathways they are involved in. The loss of the *PTEN* tumour suppressor gene through deletion is a prime example, as it results in the deregulation of the *PIK3/Akt* pathway [43]. This pathway plays an important role in the maintaining a balanced cell proliferation, cell growth and in apoptosis. The deletion of *PTEN* is therefore a major contributor to the development of many cancers [44].

The duplication or deletion of genomic sequences containing 50 or more base-pairs are commonly referred to as copy-number variants (CNVs) [45]. The frequency of CNVs can vary greatly among a population and are considered as important as single nucleotide polymorphisms in defining genetic diversity [46]. The percentage of a genome affected by CNVs is known as the CNV burden. Within the PAM50 breast cancer subtype CNV burden has been shown to be significantly associated with disease survival, suggesting the potential use of CNV burden as a prognostic biomarker [47].

2.3.3 DNA Methylation

There are many other ways to alter a gene's normal activity without changing the DNA sequence. One such example is a type of epigenetic mechanism called DNA methylation; the addition of a methyl group (CH_3) to a CpG site (a location where a cytosine nucleotide is followed by a guanine nucleotide) to suppress the expression of a target gene. The addition of methyl groups is facilitated by the family of enzymes called DNA methyltransferases (DNMTs). These enzymes allow a methyl group to bind to the fifth carbon of cytosine bases to form 5-methylcytosine.

CpG sites that are significantly denser than the surrounding DNA sequence are known as CpG islands. Unlike sparse CpG sites that are commonly methylated, CpG islands are typically unmethylated, but can cause gene silencing if they become methylated [48]. Gardiner-Garden and Frommer [49] provided the first formal definition of CpG islands as regions that meet the following three requirements:

- A region greater than 200 base-pairs in length.
- A G+C content greater than 50%.
- An observed versed expected ratio for the occurrence of CpGs of more than 0.6.

These CpG islands frequently occur in transcription start sites and in promoter regions. Methylation of CpG islands plays an essential role in mammalian development and maintaining genetic stability [50]. However, if this methylation occurs improperly it can lead to the promotion of diseases. In the context of cancer development this improper methylation may occur in mismatch-repair genes and tumour suppressor genes, resulting in the transcriptional silencing of these genes and increasing the risk of tumour development [51]. One such example of methylation promoting the development of cancer is the hyper-methylation of MLH1 mismatch-repair genes in colorectal cancer (discussed further in Chapter 6.3.7).

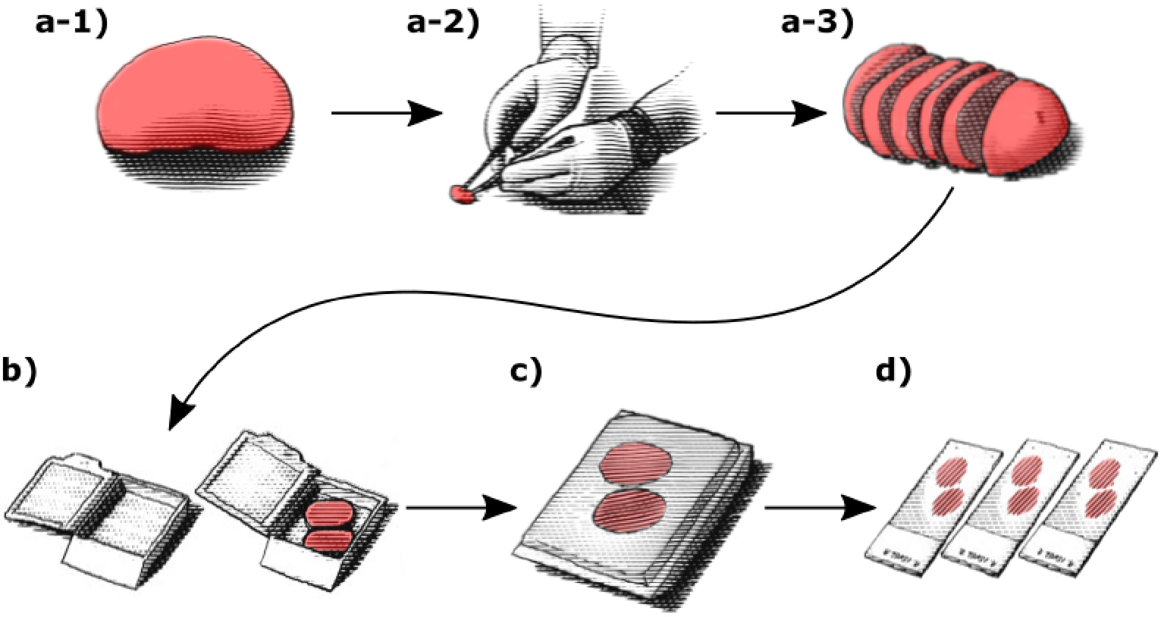
2.4 Tissue Samples and Cell Cultures

There are two main approaches to analysing cancer using tissue samples. The first approach is to extract a sample from a clinical biopsy, store it and later analyse it using a variety of available techniques, such as microarrays or methylation arrays. The second is to grow the cancer cell cultures *in vitro* or *in vivo* under controlled conditions and analyse their development, or use the mature cells in a variety of -omics based experiments [52].

The two approaches bring their own advantages and disadvantages. Clinical samples accurately reflect how cancers develop naturally, providing a snapshot of the diseases' progression in individual patients. The information that can be extracted from clinical samples regarding how the cells evolved is however restricted to indirect information as the conditions cannot be tightly controlled. Acquiring the samples can also be unpleasant for the patient and result in undesired side effects. The collection process also carries the risk of providing a limit source of material. Cell cultures on the other hand provide a dynamic view of the tumour cell proliferation and a readily available source of additional material. The conditions of the culture can be closely controlled and their effects measured, but they may not accurately reflect what happens in vivo.

Clinical samples can be obtained from normal, primary tumour or metastatic tissue. To prevent the samples from degrading they must be *fresh frozen (FF)* or *formalin-fixed-paraffin-embedded (FFPE)*. FF samples are created by submerging the fresh tissue sample in liquid nitrogen. The FF samples can later be analysed by slowly thawing them in solution [53] or quickly grounding them to preserve RNA integrity. FFPE samples are created in a two step process. They are first treated with formalin to preserve the tissue, before being embedded in paraffin to support the tissue. FFPE samples can be analysed later by cutting them into slices or microscopic sections (Figure 2.4).

Fig. 2.4 Processing FFPE tissue samples. **a)** Tissue sample is serial sectioned. **b)** Sectioned tissue is placed in a plastic cassette to be processed. The tissue is fixed and embedded in paraffin. **c)** After processing, blocks are formed. Each block consists of tissue embedded in paraffin that is attached to the bottom of the cassette. **d)** A microtome is used to slice the block into slices (typically 4 microns thick) that can be mounted and analysed. Adapted from Lester (2010) [2].



One of the main advantages of using FF samples is that they usually contain better quality RNA [54]. However, FF samples must be stored in dedicated freezer storage making them far less cost-efficient than FFPE samples, which can be stored at room temperature [55]. Due to their cost efficiency and ability to be processed faster, FFPE archives are far more abundant.

The relative abundance of FFPE samples has made them an appealing source of material for retrospective studies. Researchers have since compared FFPE and FF paired samples to test whether the difference in RNA quality significantly affects study results [56] and concluded that FF is the ideal source of material, but FFPE is suitable for gene expression and single-nucleotide variant (SNV) detection [55].

As previously discussed, cell cultures are an alternative way to study cancer development. To establish a cell culture, cells must be isolated from a tissue sample. These cells are referred to as the primary cells and are the best representation of the in vivo state [57]. Once the cell culture has undergone multiple sub-cultures it produces a *cell line*. These cell lines can be grown and injected into immunodeficient organisms, typically mice, to obtain an efficient in vivo model.

While primary cell cultures offer the best representation of the original in vivo state, they typically have a finite lifespan and are not well characterised. Established immortalised cell lines, derived from tumours that are capable of reproducing infinitely, provide an alternative solution to these problems. These cell lines are both well characterised and available for a wide range of cancer subtypes [58].

2.5 Microarrays

Microarrays are genomic tools that can simultaneously measure the expression level of thousands of genes, or other transcripts. The data output from a microarray is a matrix of real positive values representing the expression levels, with normal or log-normal expression level distributions [59]. One of the first attempts at producing these tools, albeit on a much smaller scale, appeared in 1975 [60]. However the modern description of a microarray did not appear until the mid-1990s [61].

The three main types of microarrays are two-colour arrays, bead arrays and Affymetrix arrays [62]. Here we focus on Affymetrix arrays which were primarily used throughout this thesis. Modern microarrays are small silicon or glass slides that can contain millions of *probes* (spots). Millions of copies of the same single-stranded DNA sequence are attached to the microarray surface at each spot. These DNA sequences each correspond to a region of interest within a genome. The probes can be grouped into two main types, *perfect match* (PM) probes and *mismatch* (MM) probes that together form a *probe pair*. The middle base pair of the PM probe is typically changed to form the corresponding MM probe, which is intended to measure non-specific binding.

To measure the gene expression level of a sample, RNA is first extracted from the sample cells. The RNA is then amplified and converted to complementary DNA (cDNA) in a reaction called reverse transcription. The cDNA is then labelled using a fluorescent dye and injected onto the microarray. Depending on the expression level of each gene a greater or lower number of complementary sequences hybridise to each probe. A laser then scans the microarray measuring the luminosity of each spot [63]. The luminosity of each spot is then converted to a set of numeric values, which must be normalised to take into account background noise, slide position and other non-biological effects.

To avoid background noise and the position of an interrogation probe within the microarray from affecting the results, a microarray usually has several probes measuring the expression of the same biological sequence (exon in this case). These probes typically map to different locations with the same genomic region and are placed in different positions on the chip to prevent localised biases. The group of probes that interrogate the same region are known as a probeset. During the normalisation of the microarray data, the expression level estimates from each probe are adjusted and summarised to obtain a single estimate for each probeset. The main limitation of microarrays is their inability to measure the expression of every known gene, due to the limited number of probesets.

One of the many uses of microarray technologies is to identify differentially expressed genes between two groups. This could be between cancerous and non-cancerous tissue;

patients with different clinical outcomes; or samples before and after a given treatment. Statistical methods, such as *t*-tests, adaptive ranking and two-way clustering can be used to analyse the output from microarrays to identify differentially expressed genes between samples [64]. One of the most commonly used R packages to identify differentially expressed genes is the Limma package [65].

Microarrays can also be used to derive gene signatures that can be used as biomarkers. These signatures can be used to discriminate between multiple conditions and classify new samples into distinct clinical outcomes [66]. The extensive repositories of microarray data further lends itself to this form of research.

2.5.1 Exon Microarrays

The analyses performed later in this report predominately use the highest resolution microarrays currently available. These high resolution microarrays are called *Affymetrix GeneChip Human Exon 1.0 ST Arrays* and will be referred to in this report as *exon microarrays*.

Exon microarrays contain over 5.5 million probes to interrogate over 1 million known or predicted exons. They contain on average 4 probes per exon and 40 probes per gene [67]. This comprehensive coverage enables analyses to be performed at both the exon and gene levels. For the purposes of this work we will be focusing on gene level analysis.

In a standard Affymetrix microarray there is usually a single mismatch probe for every perfect match probe. These mismatch probes have the same sequence as their perfect probe counter part, however one nucleotide in the middle of the sequence is altered, giving rise the name mismatch probes. The reason for having these probes comes into effect when correcting for background noise. They allow the non-specific hybridisation levels to be estimated and so help with the data normalisation [68].

Exon microarrays in contrast do not contain mismatch probes. Due to this several standard normalisation algorithms that rely on mismatch probes cannot be used. To compensate for the lack of mismatch probes other algorithms, such as RMA, PLIER and SCAN can be used [69]. These other algorithms use two alternative types of probe that can be found in exon microarrays. These probes are:

- **Genomic background probes** - Probes from regions of the genome that are unlikely to be transcribed.
- **Anti-genomic background probes** - Probes that are not found in the genome.

2.6 Discussion

In this chapter we have introduced the main biological concepts and technologies relevant to this thesis. We have presented the central dogma of molecular biology and explored the main data sources that will be used in our analyses. In the next chapter we will explore the bioinformatics methods used to analyse our data and the machine learning algorithm (latent process decomposition) used to produce our classifications of prostate and colorectal cancer. The next chapter will also explain that latent process decomposition assumes a normal distribution of gene expression level. It is important to highlight that microarrays have been selected as the main source of data due to their abundance and because they fit this distributional assumption.

Chapter 3

Computational Background

3.1 Summary

In this chapter we start by introducing the data normalisation techniques used to remove the batch effects from our gene expression data. After, we discuss the clustering technique used in Chapters 5 and 7 to define groups of patients based on their gene expression profiles. We then introduce the survival analysis models used to analyse the clinical risks associated with these classifications, which can be used to inform patient prognosis and treatment. Finally we discuss the application of pathway analysis in understanding the mechanisms driving the development and progression of our novel subtypes.

3.2 Data Normalisation

Before gene expression data can be analysed it must be thoroughly normalised to remove the batch effects and biases that can occur during the sample extraction and processing [70]. Batch effects must be accounted for across each sample within a given dataset as well as across each dataset within a study. To account for all of these factors we will now discuss three normalisation techniques that have been used in the work later in this thesis: *Robust multiarray analysis*, *ComBat* and *Quantile normalisation*.

3.2.1 Quantile Normalisation

Quantile normalisation is a technique used in statistics to make two distributions identical in terms of their statistical properties [71]. Originally called *quantile standardisation*, quantile normalisation was taken from statistics and applied to microarray data to tackle the inter- and intra-chip gene expression variability [72] [73]. It was motivated by the idea that the distribution of two data vectors are the same if they can be plotted as a straight diagonal line on a quantile-quantile plot. By projecting data points onto this line in the n^{th} dimension we can therefore transform the data into the same distribution as one another.

To achieve this projection we first let $\boldsymbol{q}_i = (q_{i1}, ..., q_{in})$ for $i = 1, ..., p$ be the vector of the i^{th} quantiles for all n arrays and $\boldsymbol{d} = (\frac{1}{\sqrt{n}}, ..., \frac{1}{\sqrt{n}})$ be the unit diagonal. We can then transform the quantiles of $\boldsymbol{q}$ to lie along the diagonal $\boldsymbol{d}$ [72]

$$\text{proj}_{\boldsymbol{d}}\boldsymbol{q_i} = \left(\frac{1}{n} \sum_{j=1}^{n} qij, ..., \frac{1}{n} \sum_{j=1}^{n} qij \right). \tag{3.1}$$

Transforming the data into the same distribution requires substituting the original data with the mean expression quantile across all arrays. Given a matrix $\boldsymbol{A}$, containing $\boldsymbol{n}$ arrays as the columns and $\boldsymbol{g}$ genes as the rows, we can transform the distribution through a five stage process:

1. Create a new matrix $\boldsymbol{B}$, of size $\boldsymbol{n} \times \boldsymbol{g}$, containing the numerically ascending ranks of the columns from $\boldsymbol{A}$.
2. Reorder each column of $\boldsymbol{A}$ into ascending order.
3. Calculate the mean of each row of $\boldsymbol{A}$ and store in vector $\boldsymbol{V}$.

4. Rank the values of $\boldsymbol{V}$ into numerically ascending order.

5. Substitute the ranked values of $\boldsymbol{V}$ into the corresponding ranks of $\boldsymbol{B}$ to ensure all arrays contain the same distribution.

3.2.2 Robust Multiarray Analysis (RMA)

The RMA algorithm by Irizarry et al. [74] is one of the most commonly used techniques for the normalisation of exon microarrays. In their work they identified that the MM probes within microarrays (Chapter 2.5) capture both the background noise and the transcript signal similar to that of the PM probes. The consequence of these findings suggestions that the difference between MM probes and PM probes would not be enough to remove the background noise and non-specific binding. To overcome these limitations Irizarry et al. proposed the RMA algorithm to better measure gene expression using log-transformed PM values following background correction and quantile normalisation.

The RMA algorithm begins with background correction to account for unwanted non-specific binding. The model assumes that the observed PM probe intensities are the combined result of the true signal (S) and some background noise (B). This can be represented by:

$$\boldsymbol{PM} = \boldsymbol{S} + \boldsymbol{B}, \tag{3.2}$$

where S is assumed to follow an exponential (positive) distribution (λ) and B is assumed to follow a normal distribution with mean μ and variance σ. These assumptions allow an empirical Bayes approach to be used to estimate λ, μ, σ from the data. Once these values have been estimated we can predict and correct for **B** by minimising the mean squared error. Following background correction, quantile normalisation is employed to transform the probe intensities of each microarray to the same distribution.

The final step of the RMA algorithm is to obtain a single value per probeset, through the summation of all probe intensities within a given probeset. Li and Wong [75] highlight the challenge of this summation as the variation of probe intensities from any given probeset may be very large, due to probe-specific effects. RMA overcomes this challenge by taking advantage of the reproducibility of these probe-specific effects and uses the following linear additive model:

$$\boldsymbol{Y}_{ijn} = \mu_{in} + \alpha_{jn} + \varepsilon_{ijn}, \quad \boldsymbol{i} = 1,...,\boldsymbol{I}, \quad \boldsymbol{j} = 1,...,\boldsymbol{J}, \quad \boldsymbol{n} = 1,...,\boldsymbol{n}, \tag{3.3}$$

where $\boldsymbol{i}$ is the index of a microarray, $\boldsymbol{n}$ is the probeset index of the microarray, $\boldsymbol{j}$ is the probe index of the probeset, $\boldsymbol{Y}_{ijn}$ represents the $\log_2$ background-adjusted and quantile normalised expression level of probe j in probeset n from microarray i, μ_{in} is the $\log_2$ expression level of probeset n in microarray i, α_{jn} represents the probe affinity effect of probe j from probeset n and ε_{ijn} is an identically distributed independent error term with a mean of 0 [74]. The above model is robust against outliers by employing a median polish algorithm [76] to estimate model parameters. The output of RMA is an estimate of μ_i as the log scale measure of expression.

3.2.3 ComBat

ComBat normalisation was proposed by Johnson et al. (2007) [77] as an extension to model-based location/scale adjustment using empirical Bayes (EB) to account for outliers in small sample sizes. It makes use of systematic batch biases that are common across many genes to shrink the batch effect parameter estimates. The method contains a three stage process: standardisation of the data, EB batch effect parameter estimation and data adjustment for batch effects.

Johnson et al. (2007) [77] initially assume a location/scale adjustment model as follows:

$$\boldsymbol{Y}_{ijg} = \alpha_g + \boldsymbol{X}\beta_g + \gamma_{ig} + \delta_{ig}\varepsilon_{ijg}, \tag{3.4}$$

where $\boldsymbol{Y}_{ijg}$ represents the expression value for gene $\boldsymbol{g}$ in sample $\boldsymbol{j}$ from batch $\boldsymbol{i}$, α_g is the overall gene expression, $\boldsymbol{X}$ is a design matrix for sample conditions and $\boldsymbol{X}\beta_g$ is the vector of regression coefficients corresponding to $\boldsymbol{X}$. The values γ_{ig} and δ_{ig} represent the additive and multiplicative batch effects respectively of batch $\boldsymbol{i}$ for gene $\boldsymbol{g}$. The error term ε_{ijg} is assumed to follow a Gaussian distribution with an expected value of zero and a variance of σ_g^2. The batch-adjusted data is therefore given as:

$$\boldsymbol{Y}_{ijg} = \frac{\boldsymbol{Y}_{ijg} - \widehat{\alpha}_g - \boldsymbol{X}\hat{\beta}_g - \widehat{\gamma}_{ig}}{\hat{\delta}_{ig}} + \widehat{\alpha}_g + \boldsymbol{X}\widehat{\beta}_g, \tag{3.5}$$

where $\widehat{\alpha}_g$, $\widehat{\beta}_g$, $\widehat{\gamma}_{ig}$ and $\widehat{\delta}_{ig}$ are estimates for the previous model parameters α_g, β_g, γ_{ig} and δ_{ig} respectively.

3.2.3.1 Step 1: Data Standardisation

The magnitude of gene expression could vary across genes due to probe sensitivity. This must be accounted for to prevent bias being introduced to the EB estimates. To avoid this bias Johnson et al. (2007) [77] begin by standardising the data gene-wise to produce a similar mean and variance for each gene. They employ a gene-wise ordinary least-squares approach and constrain $\sum_i n_i \widehat{\gamma}_{ig} = 0$ for all genes ($g = 1, \ldots, G$). The variance is then estimated as $\widehat{\sigma}_g^2 = \frac{1}{N}\sum_{ij}(\boldsymbol{Y}_{ijg} - \widehat{\alpha}_g - \boldsymbol{X}\widehat{\beta}_g - \widehat{\gamma}_{ig})^2$. From this the standardised data can be calculated as:

$$\boldsymbol{Z}_{ijg} = \frac{\boldsymbol{Y}_{ijg} - \widehat{\alpha}_g - \boldsymbol{X}\widehat{\beta}_g}{\widehat{\sigma}_g}. \tag{3.6}$$

3.2.3.2 Step 2: Empirical Bayes Batch Effect Parameter Estimation

Assuming the standardised data, $\boldsymbol{Z}_{ijg}$, follows a normal distribution with mean γ_{ig} and variance δ_{ig}^2, we can also assume that prior distributions of the batch effect parameters are approximately

$$\gamma_{ig} \sim \boldsymbol{N}(\boldsymbol{Y}_i, {\tau_i}^2) \quad \text{and} \quad {\delta_{ig}}^2 \sim \text{Inv-Gamma}(\lambda_i, \theta_i), \tag{3.7}$$

where the above hyperparameters are estimated empirically using the method of moments. Using these distributional assumptions, the EB batch effect parameter estimates are given by the following conditional posterior means

$$\gamma_{ig} = \frac{\boldsymbol{n}_i \bar{\tau}_i^2 \widehat{\gamma}_{ig} + \delta_{ig}^{2*} \bar{\gamma}_i}{\boldsymbol{n}_i \bar{\tau}_i^2 + \delta_{ig}^{2*}} \quad \text{and} \quad \delta_{ig}^{2*} = \frac{\bar{\theta}_i + \frac{1}{2}\sum_j (\boldsymbol{Z}_{ijg} - \gamma_{ig}^*)^2}{\frac{\boldsymbol{n}_j}{2} + \bar{\boldsymbol{\lambda}}_i - 1}. \tag{3.8}$$

3.2.3.3 Step 3: Data Adjustment for Batch Effects

The data can now be adjusted for batch effects using the estimators, γ_{ig}^* and δ_{ig}^{2*}. These effects can result from either human or technical errors and biases that occur throughout the processing and analysis of each sample. Using the EB estimated batch effects the batch-adjusted data, γ_{ijg}^*, can be calculated as

$$\gamma_{ijg}^* = \frac{\widehat{\sigma}_g}{\widehat{\delta}_{ig}^*}(\boldsymbol{Z}_{ijg} - \widehat{\gamma}_{ig}^*) + \widehat{\alpha}_g + \boldsymbol{X}\widehat{\beta}_g. \tag{3.9}$$

3.3 Clustering Methods

Machine learning is a branch of artificial intelligence that uses unlabelled data, or past experiences, to determine the parameters of a mathematical or statistical model, which can then be used to categorise new data. The widespread application of machine learning techniques vary from spam filtering [78] and fraud detection [79] to the classification of cancer data [80]. There are two main approaches that machine learning techniques use: supervised methods and unsupervised methods [81].

The main difference between the two approaches is due to the type of data they each use in the training phase. Supervised methods classify objects based on models that are trained using sets of objects for which the objects' classes are already known. These objects with predefined classes are referred to as *labelled objects*, while *unlabelled objects* are objects for which the class is unknown. Unsupervised methods only use unlabelled objects and must derive the objects' classes by grouping (clustering) objects with similar characteristics [81].

Many different clustering methods exist, each with a different definition for how they assign samples to their clusters. These methods can broadly be separated into two types: *hard clustering* where each object does or does not belong to a cluster and *soft clustering* (also known as *fuzzy clustering*) where each object partially belongs to multiple clusters. These definitions can be extended to form a variety of clustering types, such as *exclusive clustering*, *hierarchical clustering* and *probabilistic clustering* [82].

Exclusive clustering refers to classifying objects into non-overlapping clusters. Hierarchical clustering requires objects to belong to a given cluster and to also belong to any parent clusters associated with the given cluster. Probabilistic clustering is a form of soft clustering, where a sample is assigned to each cluster with a certain probability. The probability of a sample belonging to each cluster is between zero and one, with the sum of probabilities for a given sample equalling one.

3.3.1 *K*-means Clustering

K-means clustering [83–85] is a technique that belongs to the unsupervised exclusive clustering class of machine learning algorithms, where samples are partitioned into distinct categories. This algorithm is a special case of the Expectation Maximization (EM) algorithm (a natural generalisation of maximum likelihood estimation), where the covariances are zero and the mixture weights are equal [86]. The EM algorithm has been shown to converge within a finite number of steps, a property inherited and demonstrated by the K-means clustering algorithm [87]. The simplistic nature, finite runtime and ability to fine tune the K-means algorithm has resulted in its wide-spread application [88].

The K-means algorithm works by initialising $\boldsymbol{K}$ centroids (central cluster points) and assigns each sample to the nearest centroid's cluster. Within each cluster the centroid is then redefined as the point closest to the centre of the cluster. This is commonly calculated using the mean Euclidean distance between all points within a cluster, however other distance measures can be applied. When the centroids are redefined some samples may become closer to the centroids of other clusters. The membership of each clusters' samples is then updated. The previous steps are repeated until the process converges to a state where no samples change clusters and the centroids remain stable. This algorithm is presented schematically in Algorithm 1.

Algorithm 1 K-means Algorithm

```
Initialise the K centroids randomly.
repeat
    Assign each sample to the closest centroid.
    Update each centroid's position to the centre of its cluster.
until The centroids do no change (converge).
```

3.3.2 Topic Models

Topic models are a type of unsupervised probabilistic classifier that aim to discover abstract topics within a collection of documents through statistical modelling. These topics are derived from clustering similar words together, which is achieved by analysing the frequency that individual words occur with others.

Historically topic models were used to find patterns in documents containing natural language, but in recent years this has been extended to many fields of research including bioinformatics [89]. The origins of topic modelling can be found in the latent semantic indexing work by Deerwester et al. [90], however one of the first true topic modellers is the later work by Hofmann [91] called probabilistic latent semantic analysis (PLSA). Latent Dirichlet allocation (LDA) was proposed in 2003 by Blei et al. [92] as an extension to PLSA and has since been used as the basis of many new forms of topic modeller.

3.3.3 Latent Process Decomposition (LPD)

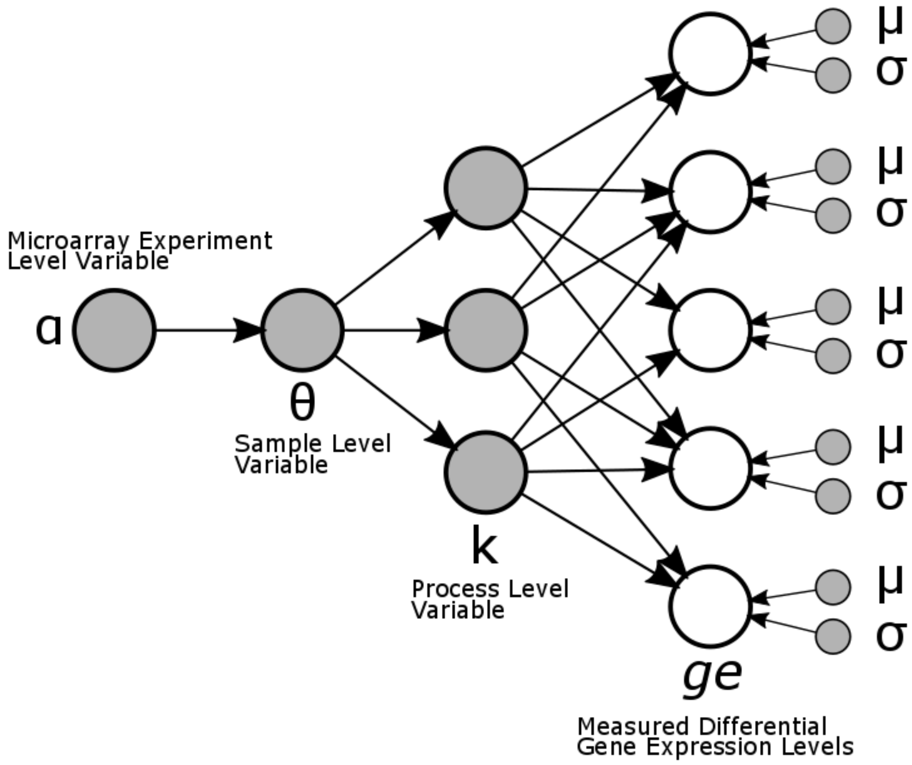
In this section we introduce Latent Process Decomposition (LPD), which forms a large part of the analysis presented in Chapters 5 and 7. LPD is a hierarchical Bayesian technique developed by Rogers et al. [3] as an extension of the Latent Dirichlet Allocation (LDA) approach [92]. Data objects within an LPD model are therefore allowed to have a partial membership to multiple clusters (processes). This simulates the potential for a given object to contain some, or all, of the defining characteristics of multiple clusters.

In the context of this project we assume that each LPD process represents a biological process within cancers, each with a distinct gene expression pattern. Prostate cancer is highly heterogeneous [93], with the potential for a mixture of tumour foci and foci subclones to be present in each sample [94]. It is consequently possible that a combination of several biological processes will be displayed within the expression profile for each sample. LPD is therefore useful as it can represent each sample as a percentage of each process.

To do this the LPD method first determines an expression profile for each process, consisting of the expected expression level of each gene in the process. The model can then estimate how well the expression profile of each process matches the gene expression levels of a given sample.

The LPD method is described by Rogers el al. [3] as follows: In an LPD model containing $\boldsymbol{K}$ processes (known in advanced), formed from a given dataset $\boldsymbol{D}$, LPD considers that each gene $\boldsymbol{g}$ in a set of genes $\boldsymbol{G}$ has a distribution specific to each process. The distribution of each gene $\boldsymbol{g}$ in process $\boldsymbol{k}$ is assumed to follow a Gaussian distribution with mean μ_{gk} and variance σ_{gk}. The distribution of processes, $\boldsymbol{\theta}$, that contribute to the observed expression profile for a given sample α from $\boldsymbol{D}$ is represented as a K-dimensional vector. Each element of this vector, $\boldsymbol{\theta}_k$, contains a value between 0 and 1, where the sum of all the elements equals 1. The distribution $\boldsymbol{\theta}$ is assumed to come from a Dirichlet distribution specific to the given dataset $\boldsymbol{D}$. A graphical representation of LPD is presented in Figure 3.1 and supported by Table 3.1 and Table 3.2.

Fig. 3.1 A graphical representation of an LPD model adapted from Rogers et al. [3]. Each circle corresponds to a variable, the dark circles represent hidden variables, while the empty circles show observed variables. The arrows represent conditional dependencies between variables.



Gene	Process		
	A	**B**	**C**
1	6.90457	7.29901	8.01408

2	6.31902	5.96126	7.86269
3	4.73213	4.34013	5.84578
4	7.00041	8.13431	4.24155
5	6.08666	5.83762	5.48995
...	...	...	...
G-1	7.81492	6.73804	6.27111
G	8.32382	6.23671	6.15953

Table 3.1 Table showing an example of the different Gaussian means of G genes in K=3 processes.

Sample	**Process**		
	A	**B**	**C**
1	0.0185	0.0117	0.9698
2	0.0006	0.0051	0.9944
3	0.2091	0.0005	0.7905
4	0.5134	0.3595	0.1272
5	0.6135	0.3399	0.0467
6	0.3503	0.0225	0.6272
7	0.3931	0.0573	0.5495
8	0.5507	0.0005	0.4488
...	...	...	...
α**-1**	0.1994	0.0005	0.8002
α	0.9764	0.0005	0.0232

Table 3.2 Table showing an example of the sample θ vectors containing the proportional assignment of each sample to K=3 processes.

3.3.3.1 Parameter Estimation

Bayesian models, such as LPD, use a dataset of observed data $\boldsymbol{D}$ and a set of unknown parameters $\boldsymbol{H}$. In this model the unknown parameters that need to be estimated are $\boldsymbol{H} = \{\alpha, \mu, \sigma, \boldsymbol{\theta}\}$. When fitting a model with parameters $\boldsymbol{H}$ to dataset $\boldsymbol{D}$, we are interested in estimating the values for $\boldsymbol{H}$ for which the posterior probability $p(\boldsymbol{H} \mid \boldsymbol{D})$ is maximised. This is more commonly referred to as the maximum posteriori (MAP).

Bayes' rule can be employed to estimate the MAP as:

$$p(\boldsymbol{H}|\boldsymbol{D}) = \frac{p(\boldsymbol{D}|\boldsymbol{H})p(\boldsymbol{H})}{p(\boldsymbol{D})}. \tag{3.10}$$

The factor $p(\boldsymbol{D} \mid \boldsymbol{H})$ is known as the likelihood, while $p(\boldsymbol{H})$ is the prior. We are interested in finding the value of $\boldsymbol{H}$ for which the posterior probability is maximised, therefore the denominator of the above equation can be ignored as it does not depend on $\boldsymbol{H}$. This leads to the following equation:

$$p(\boldsymbol{H}|\boldsymbol{D}) \propto p(\boldsymbol{D}|\boldsymbol{H})p(\boldsymbol{H}), \tag{3.11}$$

Where the MAP is proportional to the product of the likelihood and the prior. When the prior is uniform (uninformative) the probability $p(\boldsymbol{H})$ is constant across $\boldsymbol{H}$, resulting in the MAP solution becoming the maximum likelihood solution (MLE). In this situation we are interested in finding the values of $\boldsymbol{H}$ for which the likelihood $p(\boldsymbol{H} \mid \boldsymbol{D})$ is maximised.

One of the main problems associated with MLE is its tendency to over-fit the model to the dataset used to train the model [95]. This results in a model with a poor ability to accurately predict the cluster membership of an object from an independent dataset. To overcome this shortcoming appropriate non-uniform (informative) priors should be defined. This additional information results in the use of the MAP solution instead of the MLE solution. LPD provides implementations for both MLE and MAP solutions, however we employed the MAP solution to avoid over-fitting the model.

3.3.3.2 MLE

For the MLE solution the likelihood of a set of $\boldsymbol{T}$ training samples is:

$$p(\boldsymbol{D}|\mu,\sigma,\alpha). \tag{3.12}$$

The *log* function is a monotonous increasing function, making the search for the maximum likelihood equivalent to finding the maximum log-likelihood, defined as $log\ p(\boldsymbol{D} \mid \boldsymbol{H})$. It is usually easier to estimate the maximum log-likelihood. Using the log-likelihood instead of the likelihood and factorising over individual samples, the previous equation can be updated as:

$$log p(\boldsymbol{D}|\mu,\sigma,\alpha) = \sum_{t=1}^{T} log p(t|\mu,\sigma,\alpha). \tag{3.13}$$

Marginalising over the latent variable θ allows the expression to be expanded as follows:

$$log p(\boldsymbol{D}|\mu,\sigma,\alpha) = \sum_{t=1}^{T} log \int_{\theta} p(t|\mu,\sigma,\boldsymbol{\theta})p(\boldsymbol{\theta}|\alpha)d\boldsymbol{\theta}, \tag{3.14}$$

where the probability of a sample can be expressed in terms of its individual components, giving the following:

$$log p(t|\mu,\sigma,\alpha) = log \int_{\theta} \left\{ \prod_{g=1}^{G} \sum_{k=1}^{K} \mathcal{N}(\boldsymbol{e}_{gt}|k,\mu_{gk},\sigma_{gk}),\theta_k \right\} p(\boldsymbol{\theta}|\alpha)d\boldsymbol{\theta}, \tag{3.15}$$

where $\mathcal{N}$ denotes the normal distribution.

The summation over $\boldsymbol{k}$ inside the above equation makes the log-likelihood intractable. To solve this problem we will use the technique provided by Rogers et al. (2005) [3], known as the Bayesian variational inference framework, to estimate the parameters.

A lower bound on equation 3.15 can be inferred [96] by introducing two sets of variational parameters, $\boldsymbol{Q}_{kgt}$ and $\boldsymbol{\gamma}_{tk}$. The lower bound is guaranteed to be lower than, or equal to, the log-likelihood at any given point. It is useful to introduce the lower bound as it can be maximised more easily and the maximums usually result in good approximations for the model parameters. The two variational parameters are defined as:

$$\boldsymbol{Q}_{kgt} = \frac{\mathcal{N}(\boldsymbol{e}_{gt}|\boldsymbol{k}, \mu_{gk}, \sigma_{gk})\, exp\{\psi(\gamma_{tk})\}}{\sum_{k=1}^{K} \mathcal{N}(\boldsymbol{e}_{gt}|\boldsymbol{k}, \mu_{gk}, \sigma_{gk})\, exp\{\psi(\gamma_{tk})\}}, \tag{3.16}$$

where $\psi(\boldsymbol{x})$ is the digamma function and:

$$\boldsymbol{\gamma}_{tk} = \boldsymbol{\alpha}_k + \sum_{g=1}^{G} \boldsymbol{Q}_{kgt}. \tag{3.17}$$

Using these variational parameters for a given $\boldsymbol{\alpha}_k$, the model parameters are obtaining from the following equations:

$$\boldsymbol{\mu}_{gk} = \frac{\sum_{t=1}^{T} \boldsymbol{Q}_{kgt} \boldsymbol{e}_{gt}}{\sum_{t'=1}^{T} \boldsymbol{Q}_{kgt'}}, \tag{3.18}$$

$$\boldsymbol{\sigma}_{gk}^2 = \frac{\sum_{t=1}^{T} \boldsymbol{Q}_{kgt} (\boldsymbol{e}_{gt} - \boldsymbol{\mu}_{gk})^2}{\sum_{t'=1}^{T} \boldsymbol{Q}_{kgt'}}, \tag{3.19}$$

and:

$$\boldsymbol{\alpha}_{new} = \boldsymbol{\alpha}_{old} - \mathbf{H}(\boldsymbol{\alpha}_{old}^{-1})\mathbf{g}(\boldsymbol{\alpha}_{old}), \tag{3.20}$$

where $\mathbf{H}(\alpha)$ is the Hessian matrix and $\mathbf{g}(\alpha)$ is the gradient. The updated Dirichlet parameter $\boldsymbol{\alpha}$ is represented by $\boldsymbol{\alpha}_{new}$ and the previous iteration of $\boldsymbol{\alpha}$ is represented by $\boldsymbol{\alpha}_{old}$.

3.3.3.3 MAP

Informative priors can now be introduced when calculating the model parameters, to avoid the over-fitting associated with MLE. A logical expectation, regarding a given dataset, would be that the majority of genes would not be differentially expressed in a given process; while a smaller number of genes would have a process-specific distribution. This prior belief can be represented on the mean parameters μ_{gk} using a normal distribution with a mean of zero.

To avoid over-fitting the model, caused by the collapse of the Gaussian function to a single point, we must ensure that the variances will never be equal to zero. Equally we can assume the variance parameters μ_{gk}^2 will tend to be close to one.

The parameter priors are therefore set to:

$$p(\mu_{gk}) \propto \mathscr{N}(0, \sigma_\mu), \tag{3.21}$$

and

$$p(\sigma_{gk}^2) \propto exp\left\{-\frac{s}{\sigma_{gk}^2}\right\}. \tag{3.22}$$

The values σ_μ and s are called hyper-parameters. In full Bayesian models the hyper-parameters are estimated alongside the other model parameters. LPD however, is a type of Empirical Bayes model in which the hyper-parameters are estimated independently.

The estimation of the parameters, μ_{gk} and σ_{gk}^2, must be updated for the MAP estimation to incorporate the informative priors as follows:

$$\mu_{gk} = \frac{\sigma_\mu^2 \sum_{t=1}^T \boldsymbol{Q}_{kgt} e_{gt}}{\sigma_{gk}^2 + \sigma_\mu^2 \sum_{t=1}^T \boldsymbol{Q}_{kgt}}, \tag{3.23}$$

$$\sigma_{gk}^2 = \frac{\sum_{t=1}^T \boldsymbol{Q}_{kgt}(e_{gt} - \mu_{gk})^2 + 2s}{\sum_{t=1}^T \boldsymbol{Q}_{kgt}}. \tag{3.24}$$

3.3.4 One Added Sample LPD (OAS-LPD)

Attempting to decompose new samples using LPD, without needing to retrain an existing model was not previously possible. To overcome this problem we proposed a modified version of LPD called one added sample LPD (OAS-LPD) [5].

In OAS-LPD the model parameters, μ_{gk}, σ^2_{gk} and α, from Rogers et al. [3] are taken from an existing LPD model and frozen. The remaining variational parameters, $\boldsymbol{Q}_{kgt}$ and γ_{tk}, relating to the new samples are iteratively updated until they converge as described in EQ. 3.16 and 3.17. The new samples are therefore decomposed into the subtypes derived from the original model.

Due to the lack of retraining, OAS-LPD confers the benefit of decomposing samples magnitudes of time faster than it would take to compute a new LPD model. This additional benefit, in conjunction with the lack of retraining, lends itself to the classification of clinical samples where regular LPD would struggle.

3.4 Survival Analysis

The information presented in this section and associated subsections is predominately based on the book by Kleinbaum and Klein [97].

Survival analysis is a collection of statistical techniques for the analysis of data for which the variable of interest is the *time* until an *event* occurs. The event is not restricted to only the death of an individual, it can alternatively be the recovery time, relapse, or any other experience of interest. Time is typically measured in years, months, weeks, or days. However it may alternatively refer to the age of an individual during an event. For simplicity, the time until an event occurs is commonly referred to as the *survival time*, irrespective of the type of event. The occurrence of an event is referred to as a *failure*.

Typically in a medical study, the participants are monitored for a period of time following an initial event, such as the diagnosis of a disease. During this period of observation some participants will experience failure, however other participants may not experience the event of interest during the entirety of the study. We do not know the survival time of the patients that fall into the latter group, as such they are said to be *censored*.

There are generally three reasons why censoring may occur:

1. A participant does not experience the event of interest before the study ends.

2. A participant is lost to follow-up during the study period. This could either be due to the participant withdrawing from the study, or the participant no longer communicating with study representatives.

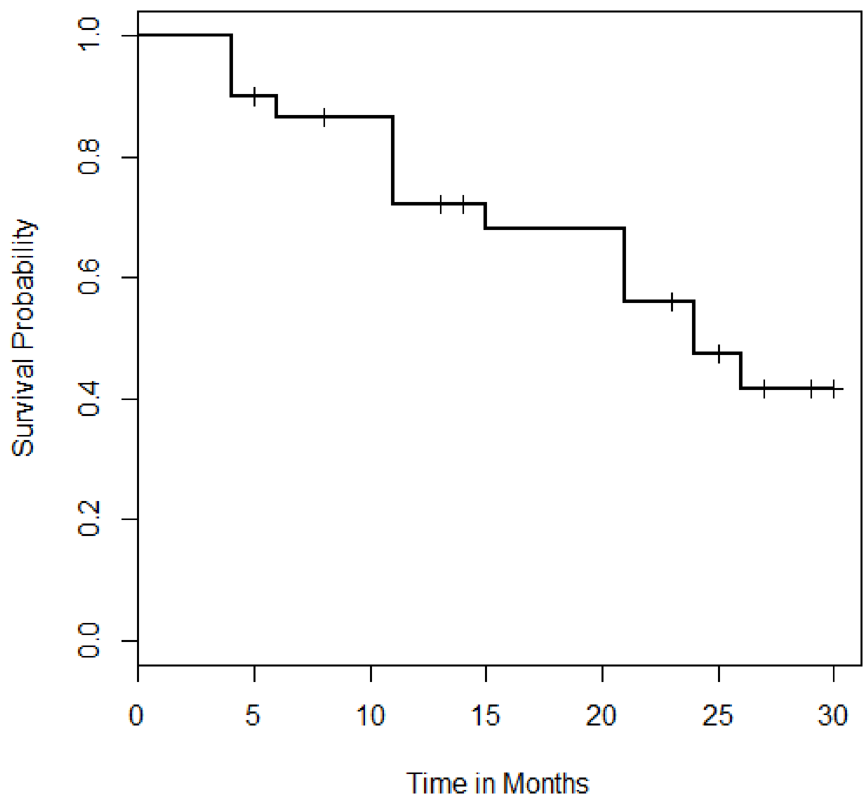
3. A participant dies during the study period, where the event of interest is not death, or the death is attributable to a reason outside the scope of the study.

3.4.1 Kaplan-Meier (KM) Survival Curves

One way of modelling survival data is through Kaplan-Meier (KM) survival curves, where the survival probability is represented as a function of time. The survival probability, $S(t)$, represents the probability that a given participant survives past a point in time $\boldsymbol{t}$. Due to the finite number of participants in a study, the estimated survival curve function, $\hat{S}(t_{(j)})$, is a step-function rather than a smooth curve, as shown in Figure 3.2.

In Table 3.3 we demonstrate how to estimate the step-function using a fictitious dataset. The first column, $\mathbf{t}_{(j)}$, contains distinct time points where failures occurred, sorted into ascending order. It is important to note that the survival time of all participants in the study must be measured in a consistent unit of time. For the purposes of this demonstration, the survival time was measured in months after the initial event for each patient. The first row of the table always begins with $\mathbf{t}_{(j)} = 0$, even though there are no failures at this point in time.

Fig. 3.2 A KM plot calculated for the data in Table 3.3. The thin crosses represent observed events that have been censored.



This is to take into account the potential for censored events to take place before the first failure.

The second column, $\mathbf{n}_{(j)}$, contains the number of participants still in the study at $\mathbf{t}_{(j)}$, including the participant(s) that failed at that point in time. The participants that were censored from $\mathbf{t}_{(j)}$ up until $\mathbf{t}_{(j+1)}$ are also included in the number of participants at $\mathbf{t}_{(j)}$. The participants in $\mathbf{n}_{(j)}$ are known as the *risk set*. The third column, $\mathbf{m}_{(j)}$, represents the number of participants that failed at time $\mathbf{t}_{(j)}$. The fourth column, $\mathbf{q}_{(j)}$, represents the number of participants that were censored in the risk set between $\mathbf{t}_{(j)}$ and $\mathbf{t}_{(j+1)}$.

$\mathbf{t}_{(j)}$	$\mathbf{n}_{(j)}$	$\mathbf{m}_{(j)}$	$\mathbf{q}_{(j)}$	$\hat{\boldsymbol{S}}(\mathbf{t}_{(j)})$
0	30	0	0	1
4	30	3	1	$1.0 \times 27/30 = 0.9$
6	26	1	1	$0.9 \times 25/26 = 0.8654$
11	24	4	2	$0.8654 \times 20/24 = 0.7212$
15	18	1	0	$0.7212 \times 17/18 = 0.6811$
21	17	3	1	$0.6811 \times 14/17 = 0.5610$
24	13	2	3	$0.5610 \times 11/13 = 0.4747$
26	8	1	2	$0.4747 \times 7/8 = 0.4154$

Table 3.3 An example of survival data, including the estimated survival probabilities. $\mathbf{t}_{(j)}$ is the survival time (in months) and $\mathbf{n}_{(j)}$ is the number of participants remaining at each survival time. While $\mathbf{m}_{(j)}$ and $\mathbf{q}_{(j)}$ represent the number of failures and censored events respectively, at each survival time.

The final column, $\hat{\boldsymbol{S}}(\mathbf{t}_{(j)})$, demonstrates the estimation of the survival probability at each given time point. The general formula for this function is the product of two factors:

$$\hat{\boldsymbol{S}}(\mathbf{t}_{(j)}) = \hat{\boldsymbol{S}}(\mathbf{t}_{(j-1)}) \times \boldsymbol{p}(\boldsymbol{T} > \mathbf{t}_{(j)} | \boldsymbol{T} \geq \mathbf{t}_{(j)}), \tag{3.25}$$

where first factor, $\hat{\boldsymbol{S}}(\mathbf{t}_{(j-1)})$, represents the probability of surviving past the previous failure time $\mathbf{t}_{(j-1)}$ and the second factor $\boldsymbol{p}(\boldsymbol{T} > \mathbf{t}_{(j)} | \boldsymbol{T} \geq \mathbf{t}_{(j)})$ represents the probability of surviving past the time $\mathbf{t}_{(j)}$, given the participant survived until at least $\mathbf{t}_{(j)}$. As demonstrated in Table 3.3, the survival probability at $\mathbf{t}_{(j)}$ requires the product of all of the previous terms. It is therefore often referred to as a *product-limit* formula.

3.4.2 Log-rank Test

One of the main objectives of survival analysis is to determine whether there is a statistically significant difference between two or more groups within a study. A commonly used technique to achieve this goal is to perform a log-rank test using multiple KM survival curves. An example of where this would be useful is during the testing of a new drug, compared with a placebo and/or existing drugs, to test whether the drug improves patient survival rates.

The log-rank test is a form of $\boldsymbol{\chi}^2$ test that compares estimates of the hazard functions of each group at every ordered observable event time. In Table 3.4 we present an example of the log-rank test by Kleinbaum and Klein (2005) [97], using 42 leukaemia patients split into two groups. The first group contains 21 patients using a placebo, while the second group contains 21 patients undergoing treatment. The data has been sorted into ascending order based on the failure time $\boldsymbol{t}_{(j)}$. The columns $\boldsymbol{n}_{(gj)}$ represent the risk set size for group $\boldsymbol{g}$ at each time $\boldsymbol{t}_{(j)}$. Likewise, columns $\boldsymbol{m}_{(gj)}$ represent the number of patients in group $\boldsymbol{g}$ that experienced failure at time $\boldsymbol{t}_{(j)}$. Columns $\boldsymbol{e}_{(gj)}$ represent the number of patients in group $\boldsymbol{g}$ that were expected to experience failure at time $\boldsymbol{t}_{(j)}$. The expected cell counts for column $\boldsymbol{e}_{(gj)}$ are calculated as:

$$\boldsymbol{e}_{(gj)} = \frac{\boldsymbol{n}_{(gj)}}{\sum_g \boldsymbol{n}_{(gj)}} \times \sum_g \boldsymbol{m}_{(gj)}, \tag{3.26}$$

which is the proportion of participants in group $\boldsymbol{g}$ at time $\boldsymbol{t}$, multiplied by the total number of failures at time $\boldsymbol{t}$.

The log-rank statistic uses the sum of the observed failures, minus the expected failures, as shown in the last two columns of Table 3.4. The log-rank statistic for the two groups is calculated as:

$$\text{Log-rank statistic} = \frac{(\boldsymbol{O}_g - \boldsymbol{E}_g)^2}{Var(\boldsymbol{O}_g - \boldsymbol{E}_g)}, \tag{3.27}$$

		Risk Set		**O**		**E**		**O – E**	
j	$\mathbf{t}_{(j)}$	$\mathbf{n}_{(1j)}$	$\mathbf{n}_{(2j)}$	$\mathbf{m}_{(1j)}$	$\mathbf{m}_{(2j)}$	$\mathbf{e}_{(1j)}$	$\mathbf{e}_{(2j)}$	$\mathbf{m}_{(1j)} - \mathbf{e}_{(1j)}$	$\mathbf{m}_{(2j)} - \mathbf{e}_{(2j)}$
1	1	21	21	0	2	(21/42) × 2	(21/42) × 2	-1.00	1.00
2	2	21	19	0	2	(21/40) × 2	(19/40) × 2	-1.05	1.05
3	3	21	17	0	1	(21/38) × 1	(17/38) × 1	-0.55	0.55
4	4	21	16	0	2	(21/37) × 2	(16/37) × 2	-1.14	1.14
5	5	21	14	0	2	(21/35) × 2	(14/35) × 2	-1.20	1.20
6	6	21	12	3	0	(21/33) × 3	(12/33) × 3	1.09	-1.09
7	7	17	12	1	0	(17/29) × 1	(12/29) × 1	0.41	-0.41
8	8	16	12	0	4	(16/28) × 4	(12/28) × 4	-2.29	2.29
9	10	15	8	1	0	(15/23) × 1	(8/23) × 1	0.35	-0.35
10	11	13	8	0	2	(13/21) × 2	(8/21) × 2	-1.24	1.24
11	12	12	6	0	2	(12/18) × 2	(6/18)× 2	-1.33	1.33
12	13	12	4	1	0	(12/16) × 1	(4/16) × 1	0.25	-0.25
13	15	11	4	0	1	(11/15) × 1	(4/15) × 1	-0.73	0.73
14	16	11	3	1	0	(11/14) × 1	(3/14) × 1	0.21	-0.21
15	17	10	3	0	1	(10/13) × 1	(3/13) × 1	-0.77	0.77
16	22	7	2	1	1	(7/9) × 2	(2/9) × 2	-0.56	0.56
17	23	6	1	1	1	(6/7) × 2	(1/7) × 2	-0.71	0.70
Total		0	0	9	21	19.26	10.74	-10.26	10.26

Table 3.4 An example of the steps involved in the log-rank statistic.

where ***g*** represents either of the two groups, as they each result in the same final value. This calculation can be generalised to include 3 or more groups, however this report will not go into further detail on this.

Once the log-rank statistic has been obtained, a p-value can be derived, as the log-rank statistic is approximately equal to a χ^2 test with ***G – 1*** degrees of freedom, where ***G*** is the total number of groups.

3.4.3 Cox Proportional Hazard (PH) Model

The Cox proportional hazard (PH) model [98] is one of the most popular statistical models for performing multivariate survival analysis. Its popularity and widespread use can be attributed to the model being robust, such that the results from a Cox PH model will be a close approximation to the true parametric model.

The Cox PH model has three main purposes:

1. To test if a variable is a statistically significant factor on survival probability, taking into account the effects of other covariates.

2. To provide a point estimate (hazard ratio) that describes the impact on survival probability when a variable's value changes.

3. To provide a confidence interval for the hazard ratio.

The function central to the Cox PH model is known as the *hazard function*. This function is defined as:

$$h(t,\mathbf{X}) = h_0(t)exp\left\{\sum_{i=1}^{\mathbf{n}} \beta_i x_i\right\}, \tag{3.28}$$

where $\mathbf{X} = \{x_1, x_2, .., x_n\}$ is a set of $\boldsymbol{n}$ explanatory variables and $\{\beta_1, \beta_2, .., \beta_n\}$ are a set of $\boldsymbol{n}$ coefficients corresponding to them.

The hazard function described in Equation 3.28 models the hazard rate of an individual with a given set of explanatory variables, as a function of time formed from two factors. The first factor, $h_0(t)$, is called the *baseline hazard function* and is only a function of time. Its purpose is to explain how the hazard changes over time, before considering the explanatory variables. The second factor, $exp\left\{\sum_{i=1}^{\mathbf{n}} \beta_i x_i\right\}$, is only a function of explanatory variables, which does not consider time. The parameters $\{\beta_1, \beta_2, .., \beta_n\}$ can be estimated using a partial maximum-likelihood approach and will henceforth be denoted as $\{\hat{\beta}_1, \hat{\beta}_2, .., \hat{\beta}_n\}$.

The hazard rate of one individual compared with a second individual is called the hazard ratio. The hazard ratio can be calculated for two individuals with sets of instances of the explanatory variables, $\mathbf{X}$ and $\mathbf{X}'$, as:

$$\widehat{HR} = \frac{\hat{h}(t,X)}{\hat{h}(t,X')} = exp\left\{\sum_{i=1}^{\mathbf{n}} \hat{\beta}_i (x_i - x_i')\right\}, \tag{3.29}$$

where the final exponential has been simplified, due to the factor, $h_0(t)$, cancelling out in the numerator and denominator of the two hazard functions.

Informally, the hazard ratio describes the odds of experiencing faster failure with every one unit increase in the value x_i, after adjusting for the effects of other coefficients. When the variable is categorical, the hazard ratio describes the odds of experiencing failure faster for the individuals in one category, compared to a baseline category after adjusting for other covariates.

It is important to note that the set of explanatory variables **X** are time-independent variables. This is a requirement to fulfil the PH assumption that the baseline hazard is exclusively a function of time. In spite of that, it is possible to use time-dependent variables through the use of the extended Cox model [99], however the PH assumption is no longer fulfilled and such a model will not be discussed further in this thesis. Additionally the model assumes that each variable is independent and contributes a linear relationship to Cox model.

3.5 Pathway Analysis

Biological pathways describe how multiple genes can interact to promote or suppress the production of molecules within a cell [100]. Many curated databases of biology pathways exist. Three of the largest and most widely cited databases are Gene ontology (GO) [101], Reactome [102] and the manual curated database called Kyoto encyclopaedia of genes and genomes (KEGG) [103]. The most common use of pathway analysis is to identify biological functions that correlate with an under-represented or over-represented set of genes. These gene sets are typically determined by testing for a difference in expression between two conditions, such as patient ethnicity, disease status, or other known factors.

Each pathway within a database has a known background frequency, while the set of conditional genes provides the sample frequency. The background frequency describes the number of genes annotated for a given pathway relative to the number of genes in the entire database, where the database could be all the genes on a microarray, or within a genome. A variety of statistical tests can be used to determine whether a pathway is under/over-represented in the set of conditional genes, with a given measurement of confidence. The hypergeometric test provides one way to measure the probability of a pathway being under/over-represented [104]. The associated p-value for such a test is a measure of the probability of the observed gene set containing greater or fewer genes related to a specific pathway than the expected frequency from the background set. However, as the tests are performed individually on each pathway the resulting p-values should be adjusted for multiple comparisons.

3.6 Discussion

In this chapter we have presented the main computational and statistical techniques and approaches used in this thesis. In Chapters 5 and 7 we will explore how LPD can be used to produce new classifications of prostate and colorectal cancers. We will also demonstrate the clinical associates between these new classifications using the statistical approaches discussed in this chapter. Before doing this, we will expand upon Chapter 2 in Chapter 4 to provide specific background on prostate cancer.

Chapter 4

The Prostate and Prostate Cancer

4.1 Summary

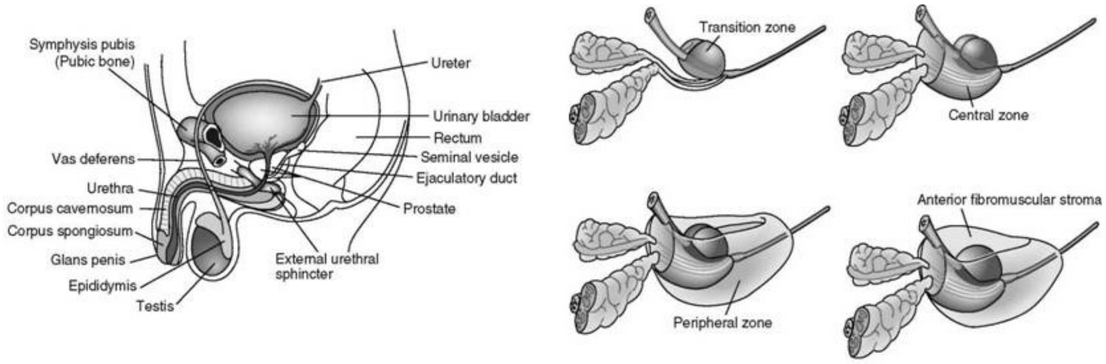
In this chapter we discuss key information pertaining to the prostate and prostate cancer, including risk factors and current disease treatments. This information will serve as the basis for understanding the current limitations of prostate cancer diagnosis and treatment. In Chapter 5 we aim to reduce unnecessary treatment (discussed in this chapter) by developing an approach to identify the risk of disease relapse. Our approach will use a novel technique combined with the risk factors discussed below.

4.2 The Prostate

The prostate is a glandular (70%) and fibromuscular (30%) organ forming part of the male reproductive system. It is surround by a capsule of collagen, elastin and smooth muscle and is located underneath the bladder. The prostate can be divided into four zones: Transitional zone, Central zone, Peripheral zone and Anterior fibromuscular stroma [105] (Figure 4.1)

The transitional zone accounts for around 5-10% of a prostate's glandular tissue and is estimated to contain 20% of the cancers occurring within the prostate [106]. The central zone

Fig. 4.1 A diagram of the location of a prostate and the four prostate zones. Adapted from AcademLib [4].



is structurally different from the rest of the prostate and accounts for 25% of the glandular tissue within a prostate. Only 1-5% of prostate cancers originate within the central zone.

The peripheral zone is the largest contributor to the glandular tissue, containing approximately 70% of all glandular tissue within the prostate. It is also the most common area for cancers to develop, with 75% of prostate cancers originating here [106]. The anterior fibromuscular stroma is rarely associated with prostate cancer development, although it contributes up to a third of the total mass of a prostate.

4.3 Prostate Cancer

4.3.1 Risk Factors

Across the wide range of potential prostate cancer risk factors only a small subset have been consistently reproducible and accepted. The risk factors that have been accepted include age, race/ethnicity and a positive family history of prostate cancer [107–109].

Age is a common risk factor in cancer development, with an increase in time allowing for a greater mutational burden to accumulate. This risk is especially true in prostate cancer with an increase in age corresponding to an increase in prostate cancer prevalence [110]. A study by Zlotta et al. (2013) [111] found that Asian men aged 81-90 years old were almost twice as likely to have prostate cancer as Asian men aged 61-70 years old (58.8% and 30.8% risk respectively).

Race is the second widely accepted risk factor for prostate cancer. In the UK, Black men are more than three times as likely to develop prostate cancer than White men [112]. Conversely Asian men are at less risk of developing prostate cancer than White men [113].

Family history plays an important role in prostate cancer risk. Men with a first degree relative diagnosed with prostate cancer were found to be at significantly greater risk of developing prostate cancer themselves [114]. Bratt et al (2016) [114] observed the risk of developing prostate cancer increasing in the general population from 4.1% to 14.9% in men with 1 brother with diagnosed prostate cancer by the age of 65 years old. This increased risk was found to decrease over time, but was still a significant factor to consider. It should however be noted that while race and family history confer a genetic predisposition to prostate cancer, immigrating to a new country for extended periods of time often results in developing the same risk as the local population [115]. This emphasises the impact of environmental factors in the development of prostate cancer.

Patient diet is one of the main environmental factors proposed as the reason for this migratory change in prostate cancer risk. While there have been conflicting studies regarding the overall effect of a population's diet on prostate cancer risk, some consistently reported findings do exist. One such finding refers to the consumption of cruciferous vegetables and their ability to provide a statistically significant protective effect against the development of prostate cancer [116]. Cohen et al. (2000) [116] hypothesise that cruciferous vegetables provide this protection through the hydrolysis of the glucosinolates found in cruciferous vegetables This supports the earlier work by Lee et al. (1994) [117] and Moskaluk et al. (1997) [118] into glutathione S-transferase activity protecting against prostate cancer.

A broader hypothesis regarding diet and prostate cancer risk is that a Western diet (defined as a diet containing a greater volume of red/processed meats, dairy and other high fat products, fried foods and high sugar content) confers an increased risk compared to that of an Asian diet (broadly defined as a diet containing an increased volume of fruit, vegetables and fish and a reduced volume of the products found in a Western diet). This hypothesis has however yielded conflicting results, with some studies accepting and others rejecting the null hypothesis [119, 120].

4.3.2 Screening and Early Detection: The Problems with Prostate Specific Antigen Testing

The early detection of cancer is an important part in the diagnosis and treatment of the disease. Screening is a form of early detection, whereby a test is performed on patients that are at risk of developing the disease, but are yet to present any symptoms. The dominate screening technique used to detect prostate cancer is to test the level of prostate specific antigen (PSA) in peripheral blood [121]. The same technique is also commonly used to test patients presenting symptoms of prostate cancer.

PSA is a single-chain glycoprotein produced by both normal and malignant cells within the prostate gland [122]. The level of PSA found within a man's blood commonly corresponds to the volume of prostate tissue within his body. Since the prostate gradually enlarges over time and its volume varies between men, testing for the level of PSA provides limited diagnostic value (discussed further in Section 4.3.3).

Although high levels of PSA are associated with the presence of prostate cancer, there is no consistent evidence that this leads to a reduction in cancer-specific mortality rates [123]. PSA screening therefore leads to significant over-diagnosis and over-treatment. Over-diagnosis refers to the detection of a disease that would not have shown any clinical symptoms during the lifespan of the patient. It has been estimated that PSA screening has led to the over-diagnosis rate of up to 44% [124].

Over-diagnosis leads to unnecessary anxiety and the potential of complications to arise from invasive tests, such as taking a biopsy. It also leads to unnecessary treatment and the high risk of a reduced quality of life associated with this treatment. In patients that underwent radical prostectomy between 16-59% of patients reported urinary incontinence [125] and >50% reported erectile problems [126]. The unneeded surgery also carries the risk of infection (reported in 20-25% of patients) and even death (reported in <0.5% of patients) [127].

4.3.3 Diagnosis

Patients that are suspected of having prostate cancer, such as those with lower urinary tract symptoms (incontinence and increased frequency of urination) [128], are first given a PSA test and a digital rectal examination (DRE). However, PSA tests lack both sensitivity and specificity. Many patients with prostate cancer have low levels of PSA [129], whereas high levels of PSA have also been associated with other conditions, such as benign prostatic hyperplasia (BPH). Cancers located in the peripheral zone can be detected by a DRE, provided the tumour is larger than 0.2 mL [130].

The European Association of Urology (EAU) and the National Institute for Health and Care Excellence (NICE) guidelines on prostate cancer recommend that a definitive diagnosis should be made using a needle biopsy [130, 131]. They also recommend that the results from a PSA test and DRE should be combined with knowledge of other risk factors, such as age, race and family history, to determine whether there is sufficient risk to justify a needle biopsy.

The standard biopsy technique is a transrectal ultrasound (TRUS) guided biopsy to collected 10-12 cores [130]. The drawback to this is that a TRUS-guided biopsy misses up to 20-30% of clinically relevant cancers [132]. If the first biopsy is negative, but other factors still indicate significantly high risk of a cancer existing, then a second biopsy may be performed with the aim of collecting 20 or more cores [130].

An alternative to the TRUS-guided biopsy was recently presented by Ahmed et al. [133] as part of the PROMIS trial. They analysed the effectiveness of using multi-parametric magnetic resonance imaging (MP-MRI) instead of the standard biopsy. The advantage to this technique is its non-invasive nature, preventing the risk of unneeded surgical complications. Their findings indicate that up to 25% of prostate biopsies could be avoided by first performing an MP-MRI. They recommend that a biopsy should still be taken in the remaining 75% of cases using the MP-MRI as a guide, due to it's lower specificity.

4.3.4 Classification criteria

Once the prostate cancer has been diagnosed it must be evaluated to determine the most suitable way of managing the disease. This can be done using the PSA and DRE results, in combination with computed tomography (CT) and MP-MRI scans to determine the current progression of the disease. A strategy for managing the disease is then developed by assessing the PSA levels, DRE results, Gleason score, tumour node metastasis (TNM) stage and other pathological features.

4.3.4.1 Gleason Score

Gleason score is a grading system based on the sum of the two most common tumour patterns observed in the biopsy by a histopathologist [134]. The most common pattern is referred to as the primary pattern, while the second most common pattern is referred to as the secondary pattern. A secondary pattern is only assigned upon meeting the condition that it is present in at least 5% of the total patterns. If this condition is not fulfilled then the grade assigned to the primary pattern is doubled.

Each of the two patterns' grades take a value between 1 and 5, creating a range of Gleason sums between 2 and 10. Grade 1 is assigned to well-differentiated tissue, while grade 5 is assigned to poorly-differentiated tissue. A higher Gleason sum is associated with a poorer prognosis. It should be noted that a Gleason sum of 7 created from a score of 4 + 3 (primary grade 4 and secondary grade 3) has significantly worse outcome than the score 3 + 4 [135].

4.3.4.2 Tumour Node Metastasis

Tumour Node Metastasis (TNM) classification is the standard system for staging malignant tumours by the American Joint Committee on Cancer (AJCC) and the International Union for Cancer Control (UICC). It comprises of three parts, described by Brierley et al. (2017) [136] as:

- **T:** describes the primary tumour site.
- **N:** describes the regional lymph node involvement.
- **M:** describes the presence or otherwise of distant metastatic spread

T - Primary Tumour			
TX -	Primary tumour cannot be assessed.		
T0 -	No evidence of primary tumour.		
T1 -	Clinically inapparent tumour that is not palpable.		
	T1a -	Tumour incidental histological finding in 5% or less of tissue resected.	
	T1b -	Tumour incidental histological finding in more than 5% of tissue resected.	
	T1c -	Tumour identified by needle biopsy (e.g., because of elevated PSA).	
T2 -	Tumour that is palpable and confined within prostate.		
	T2a -	Tumour involves one half of one lobe or less.	
	T2b -	Tumour involves more than half of one lobe, but not both lobes.	

	T2c - Tumour involves both lobes.
T3 -	Tumour extends through the prostatic capsule.
	T3a - Extracapsular extension (unilateral or bilateral) including microscopic bladder neck involvement.
	T3b - Tumour invades seminal vesicle(s).
T4 -	Tumour is fixed or invades adjacent structures other than seminal vesicles: external sphincter, rectum, levator muscles, and/or pelvic wall.
N - Regional Lymph Nodes	
NX -	Regional lymph nodes cannot be assessed.
N0 -	No regional lymph node metastasis.
N1 -	Regional lymph node metastasis.
M - Distant Metastasis	
M0 -	No Distant metastasis.
M1 -	Distant metastasis.
	M1a - Non-regional lymph node(s).
	M1b - Bone(s).
	M1c - Other site(s).

Table 4.1 Tumour Node Metastasis (TNM) classification system.

4.3.4.3 ICGC risk stratification

Prostate cancer patients that have undergone prostatectomy can be categorised into three distinct risk groups, based on the UK International Cancer Genome Consortium (ICGC) consensus (Professor Chris Foster, personal communication). The criteria for each of the three groups are presented in Table 4.2.

Risk Level	**Criteria**
Low Risk	**1)** PSA $\leq$ 10ng/ml AND Gleason = 3+3
	2) PSA $\leq$ 10ng/ml AND Gleason = 3+4 AND no extra capsular extension
Medium Risk	**1)** 10ng/ml < PSA $\leq$ 20ng/ml
	2) Gleason = 4+3 AND no extra capsular extension
	3) Gleason = 3+4 AND extra capsular extension
High Risk	**1)** PSA > 20ng/ml
	2) Gleason sum > 7
	3) Gleason = 4+3 AND extra capsular extension
	4) Seminal vesicle invasion

Table 4.2 ICGC risk categorisation of prostate cancer patients that have received radical prostatectomy.

4.3.5 Localised Disease and Treatment

Patients with localised prostate cancer (clinical stage T1/T2) are stratified into risk categories using D'Amico stratification [137], as shown in Table 4.3. Patients with a low level of risk are enrolled onto active surveillance or watchful waiting programmes. Patients with an intermediate or high level of risk are usually referred for radical treatments.

Level of risk	PSA		Gleason		Clinical stage
Low risk	<10 ng/ml	and	$\leq$6	and	T1-T2a
Intermediate risk	10-20 ng/ml	or	7	or	T2b
High risk	>20 ng/ml	or	8-10	or	$\geq$T2c

Table 4.3 D'Amico risk stratification for men with localised prostate cancer.

4.3.5.1 Active Surveillance

The purpose of active surveillance is to avoid or delay the treatment of patients with low risk prostate cancer. This helps to reduce over-treatment without affecting the rate of survival. If their disease progresses then the patients are referred for radical treatments, such as *Brachytherapy*, *Radiotherapy*, or *Prostatectomy*.

Low risk patients receive PSA and DRE screenings at regular intervals, every 3 months for the first 2 years and every 6 months afterwards. Repeat biopsies are also performed 6-12 months after the initial biopsy [138]. Patients are removed from active surveillance and offered radical treatment in the event of PSA double within 3 years; histological progression (Gleason ≥ 7) following repeat biopsy; clinical progression ($\geq$ T3); or at the patient's request.

4.3.5.2 Brachytherapy

Low risk patients with an early stage of localised prostate cancer can receive brachytherapy [139]. This form of treatment involves radioactive seeds being implanted into the prostate. External beam radiotherapy can also be used in combination with brachytherapy to maximise the efficiency of the treatment.

As brachytherapy is minimally invasive it is less invasive than other techniques, such as prostatectomy, this reduces the risks associated with surgery. In addition to this benefit, it has been reported that patients who received brachytherapy had significantly less urinary and sexual problems than patients who received radical prostatectomy [140].

4.3.5.3 Radiotherapy

External beam radiotherapy is another minimally invasive therapy [131]. This involves directing multiple radiation beams from different angles to intersect at the tumour. The side-effects associated with radiotherapy include fatigue, irritation and nausea, however the long-term side-effects, such as incontinence and impotence are less common than in patients who receive radical prostatectomy. Radiotherapy is also commonly combined with hormone therapy (described below) in localised and locally advanced prostate cancer to reduce the overall cancer-specific mortality rate [141].

4.3.5.4 Prostatectomy

Patients with intermediate or high risk prostate cancer are most likely to be recommended for radical prostatectomy (RP). This is an invasive procedure where the prostate gland, seminal vesicles and a portion of the surrounding tissue are surgically removed.

Patients that underwent RP have shown a relatively good outcome. The proportion of patients free from cancer progression at 5, 10 and 15 years after prostatectomy has been estimated at 82%, 77% and 75% respectively [142]. Bianco et al. also found that the cancer-specific survival at 5, 10 and 15 years was 99%, 95% and 89% respectively.

While RP has been shown to improve survival probability more than radiotherapy it has also been associated with a greater rate of urinary and sexual problems. Approximately 60% of patients were free of cancer, continent and potent two years after RP [142].

4.3.5.5 Biochemical Reoccurrence (BCR)

After radical treatment a patient's PSA level is monitored. Typically, if a patient is observed to yield a PSA level above 0.2 ng/mL during two consecutive check-ups, then the patient is considered to have BCR [130]. BCR is often considered an early warning of a cancer's clinical progression into metastasis, before other signs become apparent. However BCR often pre-dates metastasis by several years, as such it is important to avoid using some secondary treatments too early [143].

4.3.5.6 Androgen Deprivation Therapy (ADT)

Androgens are steroid hormones that control the development of the male sexual organs and male characteristics. One of the most commonly known androgens *testosterone* is produced by the testicles, while an even more potent androgen *dihydrotestosterone* (DHT) is produced from testosterone using *5α-reductase* [144].

Androgens are also involved in the development of prostate cancer. Reducing the levels of androgens or preventing them from binding to androgen receptors therefore slows down the development of a tumour and can often result in tumour shrinkage [145]. This is known as androgen deprivation therapy.

Androgen deprivation requires surgical or chemical castration. Chemical castration uses compounds that block the androgen receptors called androgen blockers. A combination of surgical and chemical castration is referred to as maximum androgen blockade (MAB) [146].

In patients that have localised prostate cancer, ADT is usually used along with radiotherapy. It is also commonly used when radical treatments are no longer effective, or if patients are unfit to receive radical treatment [147].

ADT has been shown to significantly improve the disease free survival rate when using alongside radiotherapy compared to using only radiotherapy. Bolla et al. (1997) [148] performed a study on over 400 patients to compare these survival rates. They found that radiotherapy+ADT resulted in a disease free survival rate of 85%, compared to only 48% in patients who only received radiotherapy. Similar results were found in a follow up study that looked into the 5-year disease free survival [149].

While there is a clear benefit to using ADT to help treat localised prostate cancer, it comes with significantly adverse effects. These effects include muscular and bone mass loss, depression, erectile dysfunction, anaemia and both cardiovascular and endocrinological problems [150]. Some patients may therefore decline receiving ADT.

4.3.6 Metastatic Disease and Castration Resistant Prostate Cancer (CRPC)

Patients that progress to metastatic disease can receive ADT to create a period of remission. This period does not last indefinitely and in virtually all patients the disease becomes unresponsive to ADT. This stage is referred to as *castration resistant prostate cancer* (CRPC) and is characterised by a continuous rise in the level of PSA; progression of the existing disease; or the appearance of new metastases [151].

In cases where CRPC metastases begin to develop in a bone, the micro-environment of the bone drives the development of further metastases [152]. The resulting metastases cause both bone fragility and pain to the patient. Radium-223 dichloride can be administered to prolong the survival of men with CRPC bone metastases and to improve/alleviate their symptoms [153]. This radio isotope therapy is currently used as a palliative treatment option, however clinical trials looking into its use as a combined therapy are currently ongoing [154].

4.4 Discussion

In this chapter we have explored the biology of the prostate and the known risks associated with prostate cancer. We have looked at how cases of prostate cancer are diagnosed and the current limitations of using PSA tests. We have also discussed the current clinical management of prostate cancer patients and the variety of treatment options available. In the next chapter we will build upon the molecular subtype of prostate cancer known as DESNT and perform novel analyses to determine the usefulness of this subtype in predicting the risk of recurrence in prostate cancer patients.

Chapter 5

Towards the Analysis of DESNT in Prostate Cancer Patient Samples

5.1 Summary

In this chapter we build upon the work of Luca et al. (2017) [17] with the long term aim of classifying clinical biopsies using the poor prognosis DESNT subtype. To work towards this goal we begin by applying the unsupervised Bayesian technique introduced in Chapter 3, called Latent Process Decomposition (LPD), on five prostatectomy datasets. We apply LPD on these datasets to show that we are able to produce the same DESNT subtype as Luca et al. (2017), which is characterised by the down-regulation of a core set of 45 genes.

We then apply a novel classification technique, called OAS-LPD (introduced in Chapter 3.3.4), to the five prostatectomy datasets and a small cohort of biopsy samples. OAS-LPD is shown to detect DESNT cancers within both prostatectomy and biopsy samples. Finally, we demonstrate that the risk of biochemical recurrence within prostate cancer patients can be determined by analysing the proportion of γ associated with the DESNT subtype.

5.2 Materials

The work in this chapter was performed using five datasets denoted MSKCC, CancerMap, CamCap, Stephenson and Klein, further details can be found in Table 5.1. The first four datasets each contained clinical data, however this information was not available for Klein.

The MSKCC dataset was published by Taylor et al. (2010) [155] and can be downloaded from the GEO repository under GSE21032. We have only used the sub-series GSE21034 in this work, a choice made by Luca et al. [17], to limit the range and quality of the platforms used. This sub-series contains 370 Affymetrix Human Exon 1.0 ST Array experiments and was the only dataset that contained samples from metastatic tissue, cell-lines and xenografts. For consistency with the other datasets, only the samples from primary tumours and normal prostate tissues were used. The resulting dataset contained 320 samples.

The CancerMap dataset was created by combining two Affymetrix Human Exon 1.0 ST array datasets, which will individually be referred to as ICR and Cambridge. Both of these datasets were formed as part of a joint project to collect fresh prostate cancer samples from prostatectomy patients, at the Royal Marsden NHS Foundation Trust, London, UK and Addenbrooke's Hospital, Cambridge, UK. The samples were then prepared at the Institute of Cancer Research (ICR), London, UK and CRUK Institute, Cambridge, UK. The ICR dataset contains 81 patients and the Cambridge dataset contains 73 patients, with up to four samples per patient.

The CamCap dataset was created by combining two Illumina HumanHT-12 V4.0 expression beadchip datasets. These datasets were published by Ross-Adams et al. (2015) [156] and are available from GEO under GSE70768 and GSE70769. The first dataset, GSE70768, contains 186 radical prostectomy samples and 13 TURP (transurethral resection of the prostate) samples. This was the only dataset to contain TURP samples, as such these were removed for consistency with the other datasets. GSE70769 only contains 94 primary tumour samples.

The resulting dataset contained 280 samples. It should be noted however that CamCap and CancerMap have 40 patients in common.

The Stephenson dataset was published by Stephenson et al. (2005) [157]. This dataset contains 89 Affymetrix U133A human gene arrays taken after radical prostatectomy from patients with clinically localised prostate cancer.

The Klein dataset was published by Klein et al. (2015) [158] and is available at GEO under GSE62667. This dataset contains 182 formalin-fixed and paraffin-embedded (FFPE) primary tumour samples analysed with Affymetrix Human Exon 1.0 ST Arrays. No clinical data is provided for this dataset.

	Samples		Primary Tumour		Benign	
	Total	Unique	Total	Unique	Total	Unique
MSKCC	320	160	262	131	58	29
CancerMap	235	154	209	137	24	17
CamCap	280	207	207	207	73	73
Stephenson	89	89	78	78	11	11
Klein	182	182	182	182	0	0

Table 5.1 A summary of the prostate cancer datasets used in the LPD analysis.

5.3 Producing the DESNT LPD Analysis

To see if we could obtain the DESNT classification of prostate cancer, we performed an independent LPD analysis on each of the five microarray datasets, previously named MSKCC, CancerMap, CamCap, Stephenson and Klein. The 500 probesets with the greatest variance across the MSKCC dataset were selected for use in LPD, due to the infeasible computation time required to use all probesets. The use of the 500 most variable probesets was previously shown to be sufficient by Luca et al. [17] and Carrivick et al. [18]. Within our analysis these 500 probesets mapped to 492 genes in the MSKCC dataset. For the other datasets, the probesets mapping to these 492 genes were used in LPD.

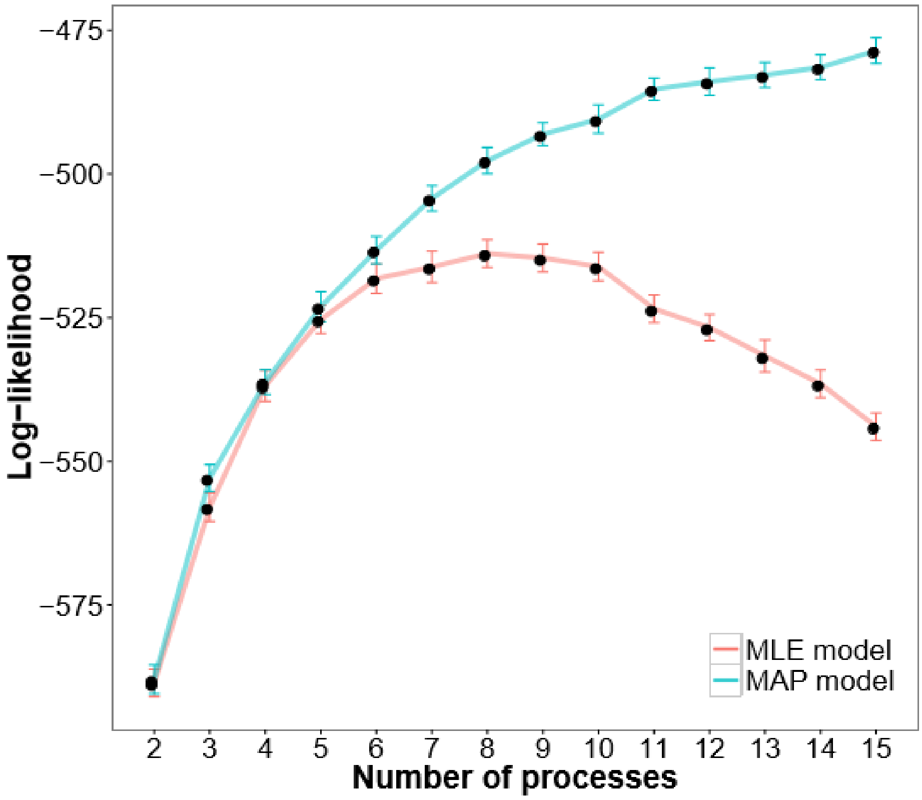
5.3.1 Choosing LPD Parameters

As discussed in Section 3.3.3, there are two forms of LPD, the maximum likelihood (MLE) model and maximum posterior (MAP) model. The MAP variant is more suitable to use as a final model as it helps to prevent over-fitting. However, to use the MAP variant two starting parameters, σ and the number of processes, must be carefully selected. To identify appropriate values for these parameters we used a three step process:

- **Step 1:** The number of processes within the dataset is estimated using the MLE model. The log-likelihood is calculated across a range of values and the value corresponding to the highest log-likelihood is deemed suitable. In our experiments we used a range of 2-15 processes to find a suitable number.
- **Step 2:** A suitable value for sigma is selected using the MAP model. The suitable number of processes from step 1 is used, along with a range of values for sigma. The σ value that produced the highest log-likelihood is deemed suitable to use in the next step. Similar to Rogers et al. (2005) [3], we used a set of small negative values which were: -0.01, -0.05, -0.1, -0.2, -0.3, -0.5, -0.75, -1 and -2.
- **Step 3:** The number of processes from step 1 is validated using the MAP model. The suitable σ value from step 2 is used as a starting parameter in addition to a range of process numbers. The number of processes at which the log-likelihood reaches a plateau is deemed suitable. The same range of values from step 1 were used in step 3 (2-15 processes).

LPD can give slightly different results each time it is run by converging to a different local maxima. This is caused by the randomised nature of some starting parameters within the LPD model and the repeated resampling. In light of this, the LPD alogorithm should be run multiple times to produce robust parameters. To ensure our value of σ and our number of processes were robust, we restarted the LPD algorithm 100 times at each step and used the average across all the runs for each value. An example of the results are shown for the MSKCC dataset in Figure 5.1. A summary of the suitable sigma values and number of processes for each dataset is provided in Table 5.2.

Fig. 5.1 The log-likelihood against the number of processes using the MLE solution (red curve) and the MAP solution (blue curve) for the MSKCC dataset. The points represent the mean log-likelihood from 100 LPD restarts. Error bars for each point are also provided to demonstrate the distribution of log-likelihoods across the LPD restarts.



Dataset	σ	No. of Processes
MSKCC	-0.5	8
CancerMap	-0.5	8
CamCap	-0.05	6
Stephenson	-0.75	3
Klein	-0.3	5

Table 5.2 A summary of the suitable parameters identified for each prostate cancer dataset.

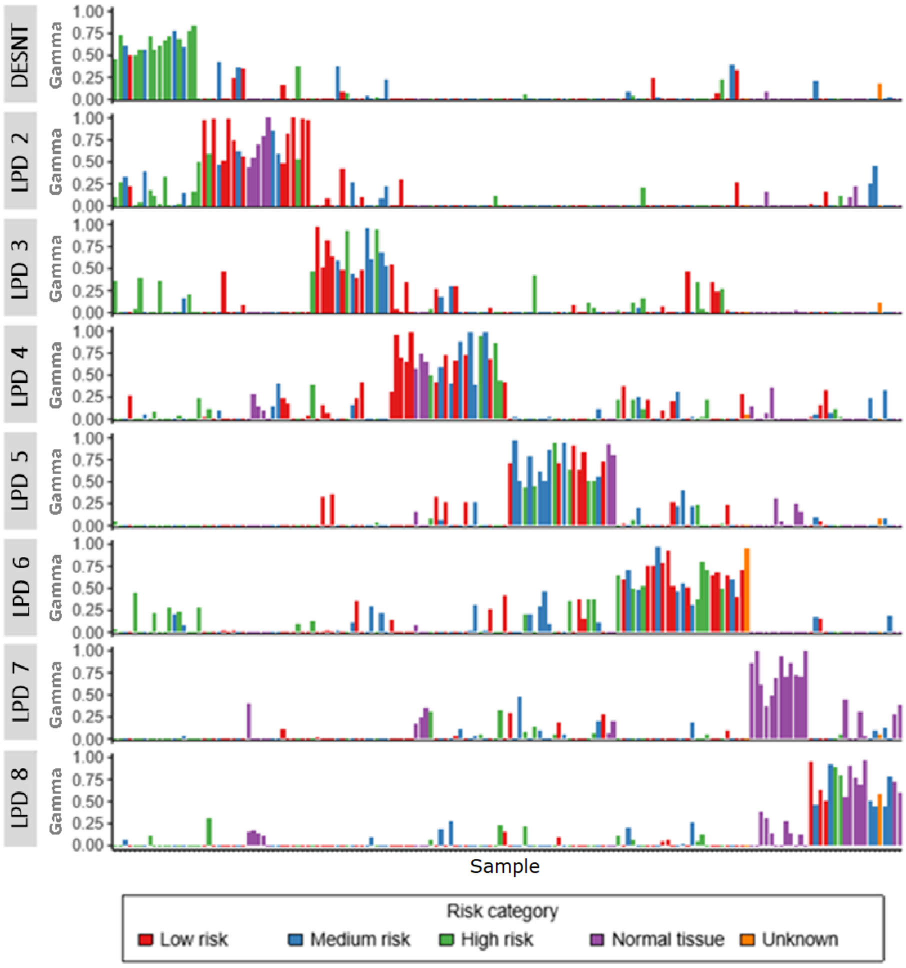
5.3.2 LPD Classification

We applied the LPD algorithm, using the results from Table 5.2, on the samples from the five prostate cancer datasets described in Section 5.2 (MSKCC, CancerMap, CamCap, Stephenson and Klein) to produce an unsupervised classification of prostate cancer. As previously discussed, LPD generates non-identical results. To produce a classification that represented a closer portrayal of the underlying processes within each dataset, we restarted the LPD algorithm 100 times. An example of one of these runs for the MSKCC dataset is shown below in Figure 5.2.

5.3.3 Survival Analyses

In order to perform survival analysis, samples must be exclusively classified into separate groups and have accompanying clinical data for a given event. Since LPD produces a set of probabilistic results, we had to convert these results into a set of exclusive associations. To achieve this, each sample was assigned to the process with the largest contribution to its expression profile. The clinical event used to compare survival times was biochemical recurrence (BCR). As the Klein dataset did not contain clinical data, we were unable to use it during the survival analyses.

Fig. 5.2 A bar chart showing the output from an LPD run, using the MSKCC dataset. Each bar within a process (row) represents the proportion for which that sample was associated with that process (σ). The colour of each bar represents the ICGC category assigned to each sample within the associated clinical data.

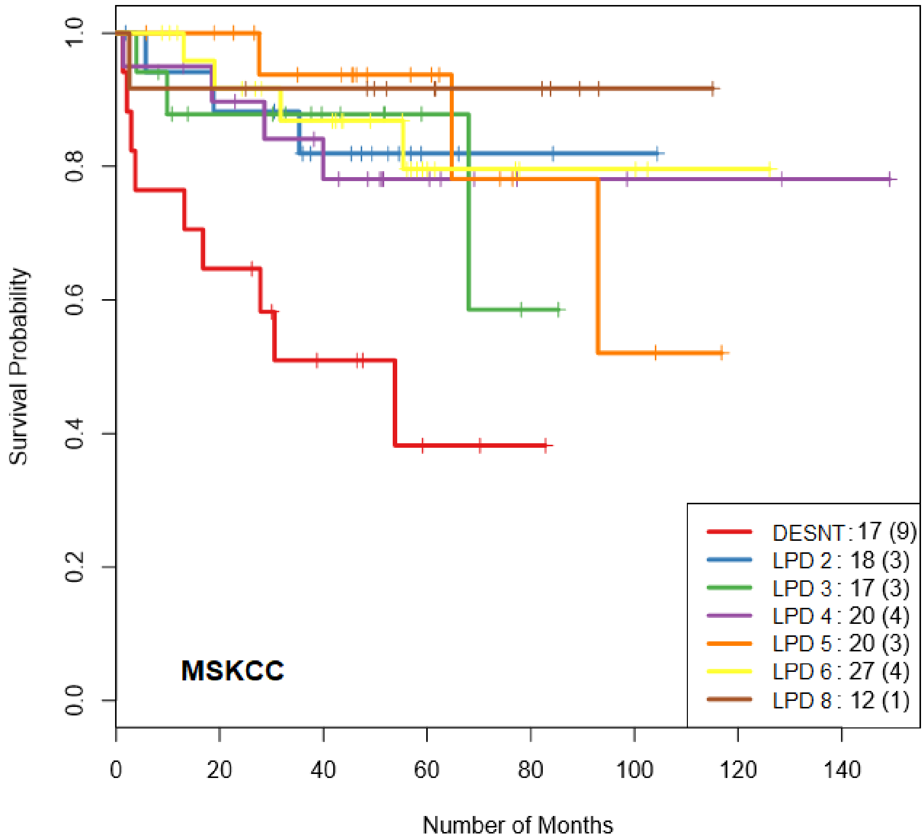


Across the 100 LPD runs the set of randomised variables created the potential for samples to change their main process assignment. To take this variability into account we determined the average run, as this would be a better representation of the underlying processes and allow us to confidently assign samples to their main processes in that given run. To find the average run we plotted the density distribution of log-rank p-values across the 100 LPD runs. The run with a log-rank p-value closest to the mode of the distribution was then selected as the representative run.

5.3.3.1 Univariate Survival Analysis

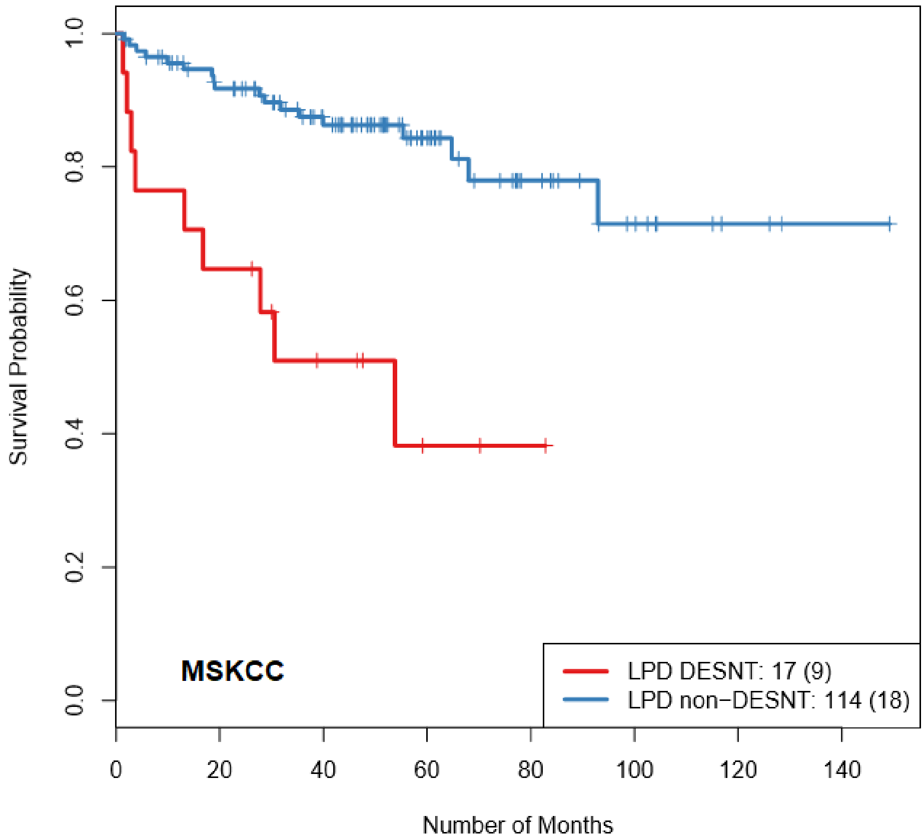
The log-rank *p*-values for each of the representative runs were very low (MSKCC 4.88×10^{-3}, CancerMap 1.57×10^{-5}, CamCap 6.27×10^{-3}, Stephenson 1.75×10^{-4}), suggesting that the LPD groups had statistically different rates of BCR failure. Survival curves were then plotted for each of the LPD groups, as shown in Figure 5.3 for the MSKCC dataset.

Fig. 5.3 Kaplan-Meier survival curves for the eight LPD groups created from the MSKCC dataset, using BCR failure as the event. The number of cancer samples in each group is indicated at the bottom right corner, alongside the number of BCR failures in parentheses.



For all of the datasets, the Kaplan-Meier plots contained a minimum of one curve that showed a lower survival probability over time, compared to the other curves. To test whether this difference in survival probability was significant, we compared the lowest survival curve against all the other curves combined. The resulting Kaplan-Meier survival curves are shown in Figure 5.4, with the low survival curve denoted as DESNT. A log-rank test was performed on each dataset to compare the DESNT and non-DESNT curves. The resulting log-rank p-values from these tests (MSKCC 2.64×10^{-5}, CancerMap 2.98×10^{-8}, CamCap 1.22×10^{-3}, Stephenson 4.28×10^{-5}) showed a statistically significant difference in survival probability, for the DESNT group compared with other groups, across all the datasets.

Fig. 5.4 Kaplan-Meier survival curves comparing DESNT and non-DESNT groups for the MSKCC dataset, using BCR failure as the event. The number of cancer samples in each group is indicated at the bottom right corner, alongside the number of BCR failures in parentheses.



5.3.3.2 Multivariate Survival Analysis

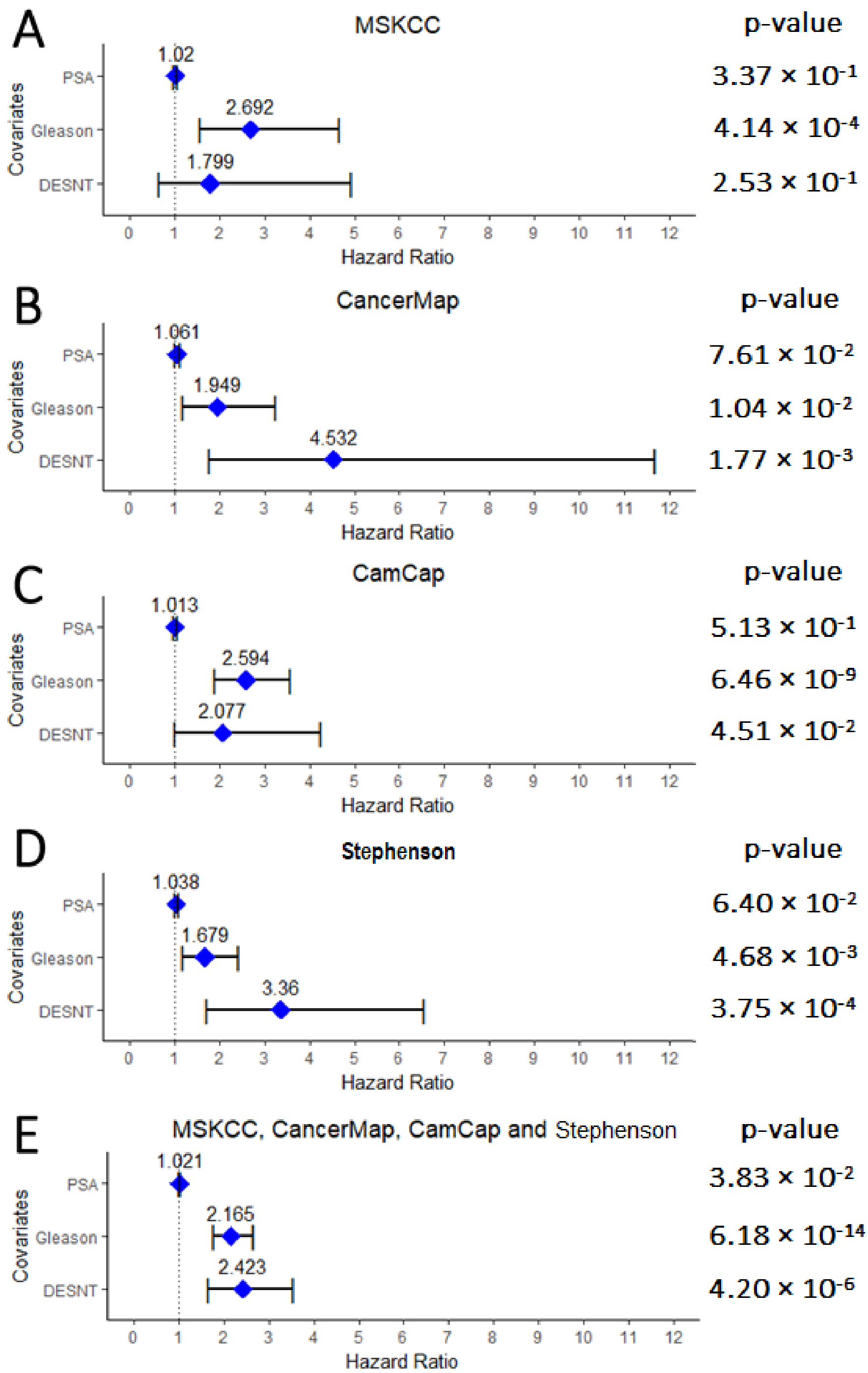
Once we had determined that DESNT was a statistically significant predictor of BCR failure, we began to access whether or not DESNT could be used as an independent predictor of BCR failure. To achieve this we performed a multivariate survival analysis using a Cox PH model. For the purpose of this report an extended Cox PH model was not used. Due to this, pathological stage was not used as it was shown to be a non-independent factor over time by Luca et al. [17]. In contrast to this, PSA level and Gleason score each fulfilled the PH assumption, allowing them to be covariates within the Cox PH model.

A Cox PH model was generated for each dataset, using DESNT membership, PSA level ($\leq$ / $>$ 10) and Gleason score ($\leq$ / $>$ 7) as the covariates. These results are depicted in Figure 5.5 (A-D). DESNT membership was found to be a statistically significant indicator in the majority of the datasets tested, however this result was not found when using the MSKCC dataset. The MSKCC Cox PH model showed that DESNT membership had a hazard ratio (1.799) greater than one, but due to the 95% confidence interval (0.658) extending significantly below one, the p-value (2.53×10^{-1}) associated with this hazard ratio was not statistically significant. This wide confidence interval is likely due to the relatively low number of samples assigned to DESNT within the MSKCC dataset (17/131 samples).

The Cox PH models generated from the CancerMap, CamCap and Glinsky datasets each had a hazard ratio greater than one (4.532, 2.077 and 3.360 respectively), indicating that DESNT membership was a predictor of BCR failure. These hazard ratios were also statistically significant (1.77×10^{-3}, 4.51×10^{-2} and 3.75×10^{-4} respectively).

Since the DESNT group was relatively small and in some cases produced large confidence intervals for a given dataset, we performed further analysis by merging the datasets together. The MSKCC, CancerMap, CamCap and Glinsky datasets were all merged together, with duplicate patients from CancerMap and CamCap removed at random. The results from this model are shown in Figure 5.5 (E) and Appendix A.1 (E). Membership to the DESNT group was found to be the largest significant predictor of BCR failure, after adjusting for the effects of other covariates. This result suggests that DESNT membership is a predictor of BCR failure, independent of PSA level and Gleason score.

Fig. 5.5 Results from the multivariate Cox PH models, using the MSKCC (A), CancerMap (B), CamCap (C), Stephenson (D) datasets and a combination of the previous four datasets (E). The blue markers denote the hazard ratio for each covariate and the extended bars denote the 95% confidence interval. The log-rank p-value for each covariates' hazard ratio is listed on the right side of the figure. PSA level was split on $\leq / > 10$, Gleason score was split on $\leq / > 7$ and DESNT γ was treated as a continuous variable between 0 to 1.



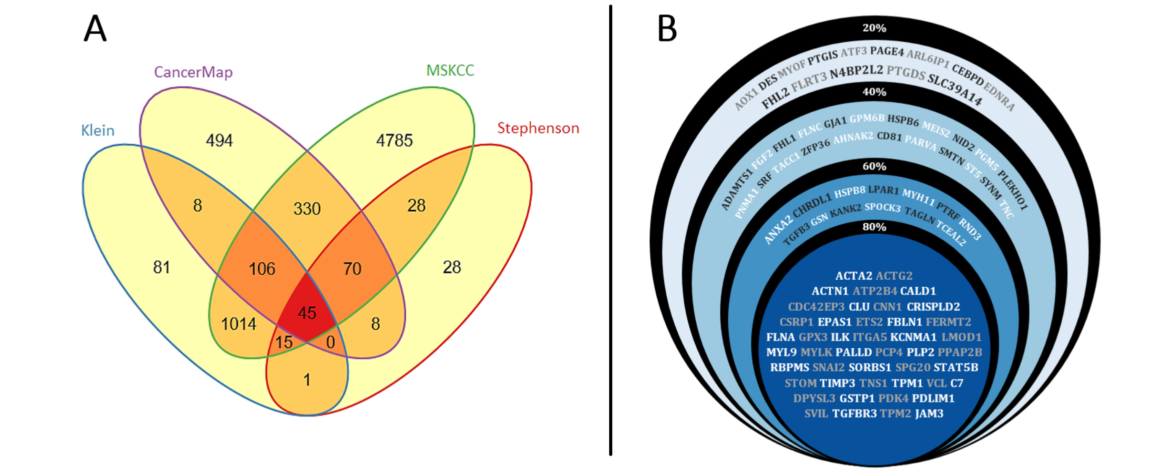
5.3.4 Differentially Expressed Genes

In this section we identify a set of genes that were differentially expressed in DESNT compared to non-DESNT cancers. All of the available probesets were used for this analysis, as it is possible that some genes outside of the 500 listed could still be discriminative for the DESNT group. Similar to our previous analysis of LPD, we used all 100 LPD restarts for a more robust analysis.

We selected the genes that were differentially expressed in each LPD restarts' DESNT group (compared to all other groups combined) and narrowed the list down to the genes that were present in 20%, 40%, 60% and 80% of the restarts. We then determined which of these genes were present across each of the datasets. Figure 5.6 (A) depicts the number of genes that were differentially expressed across multiple datasets in at least 80% of the LPD restarts. Figure 5.6 (B) depicts the genes that were differentially expressed in all four of the datasets.

A total of 45 genes were identified that were differentially expressed in the DESNT group across all four datasets, in at least 80% of the runs. This set of genes matches the list published by Luca et al. (2017) [17]. A heatmap depicting the gene expression levels of the 500 genes used within the LPD classifications has been included in Appendix A.2.

Fig. 5.6 A) A venn diagram of the number of differentially expressed genes in the DESNT group compared to the non-DESNT groups, across the MSKCC, CancerMap, Stephenson and Klein datasets, that were present in at least 80% of the LPD restarts. **B)** The differentially expressed genes in the DESNT group compared to the non-DESNT groups, across the MSKCC, CancerMap, Stephenson and Klein datasets, that were present in at least 20%, 40%, 60% and 80% of the LPD restarts.



5.3.5 Pathway Analysis

Dr Bogdan Luca previously analysed the Gene Ontology (GO) [101], Kyoto Encyclopedia of Genes and Genomes (KEGG) [159] and Reactome [102] pathways that were significantly under/over-represented in the set of 45 genes associated with the DESNT signature. The following is a summary of his findings and an independent assessment by Prof. Dylan Edwards.

An independent analysis was performed using all the available pathways annotated in GO, KEGG and Reactome using the *clusterProfiler* R Bioconductor package [160]. The *p*-values were adjusted for multiple comparisons using the FDR method at a 5% level and restricted using a confidence limit of 0.05. In total over 200 GO biological processes, nine KEGG pathways and nine Reactome pathways were identified as being over-represented in the DESNT signature. The top 20 (ordered by significance) GO processes and the nine KEGG and Reactome pathways are presented in Appendix A.3. Dr Luca consistently identified muscle contraction based pathways among those with the greatest significance across all three of the databases analysed .

Prof. Dylan Edwards, from the School of Biological Sciences, UEA, performed an independent assessment of the possible molecular functions involving the set of 45 DESNT genes. The following is a reproduction of his analysis:

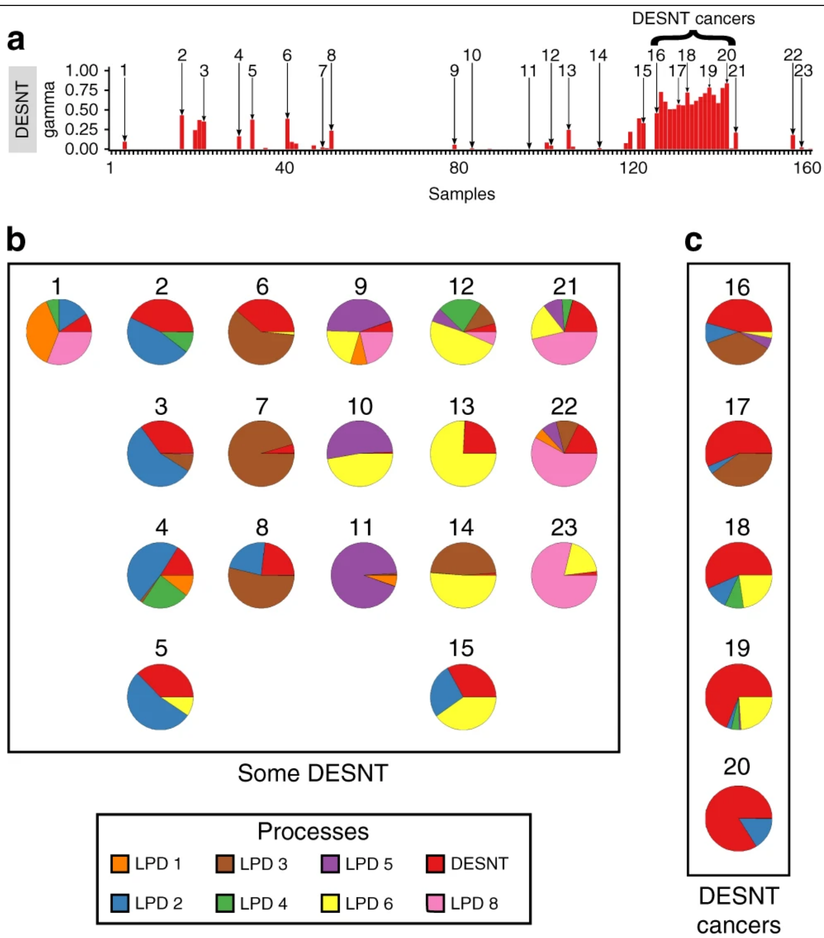
"*Several signature genes encode proteins that are components of the actin cytoskeleton or which regulate its dynamics, including ACTA2, ACTG2, ACTN1, CNN, FLNA, ILK, ITGA5, LMOD1, MYLK, PALLD, VCL, CALD1, CDC42EP3, PDLIM1, SVIL, TNS1, TPM1, TPM2. In particular, actomyosin contractility is highlighted by the presence of myosin light chain kinase (MLCK) and myosin light chain-9 (MYL9) and other molecules such as α-actinin (ACTN1), tensin (TNS1) and calponin (CNN1). Increased malignancy may correlate with increased cell migratory behaviour, which in turn may reflect the deployment of particular types of cell adhesion and cytoskeletal machinery. A high dependency on actomyosin contractility is recognised as a hallmark of amoeboid movement [161], and since this aspect is down-regulated in the poor prognosis signature, it would seem less likely to be the mode of migration employed.*

However, also noteworthy are important focal adhesion components such as integrin α5 (ITGA5), vinculin (VCL) and integrin-linked kinase (ILK), which would be expected to be involved in mesenchymal type migration. It is thus possible that the gene signature favours a collective migration phenotype, typified by maintenance of E-cadherin mediated cell-cell adhesion mechanisms [162]. Also too there are a few transcription factors and an RNA binding protein that will affect translation, thus there could be diverse downstream changes in genetic programmes as a result of the down-regulation of these genes. However, it is hard to predict the consequences here."

5.4 DESNT as a Continuous Variable

In the previous sections of this chapter we have shown that tumours predominately assigned to the DESNT subtype are associated with poorer prognosis. However, considering the full range of DESNT γ values as a continuous variable may identify its use as a predictor of risk in all patients. Based on the current LPD classifications we began to analyse the importance of DESNT γ with BCR failure. To do this we analysed whether BCR failure was related to the proportion of a sample's assignment to the DESNT group. A random selection of samples from the MSKCC dataset are depicted as pie charts in Figure 5.7, demonstrating the varying levels of DESNT γ across all samples.

Fig. 5.7 **A)** Bar chart showing the variable DESNT γ associations from the representative LPD run for the MSKCC dataset. **B)** Pie charts showing how varied the γ associations are for the range of samples highlighted in Figure 5.7-a that are not DESNT dominant. **C)** Pie charts showing how varied the γ associations are for the range of samples highlighted in Figure 5.7-a that are DESNT dominant. Published in Luca et al. (2020) [5]



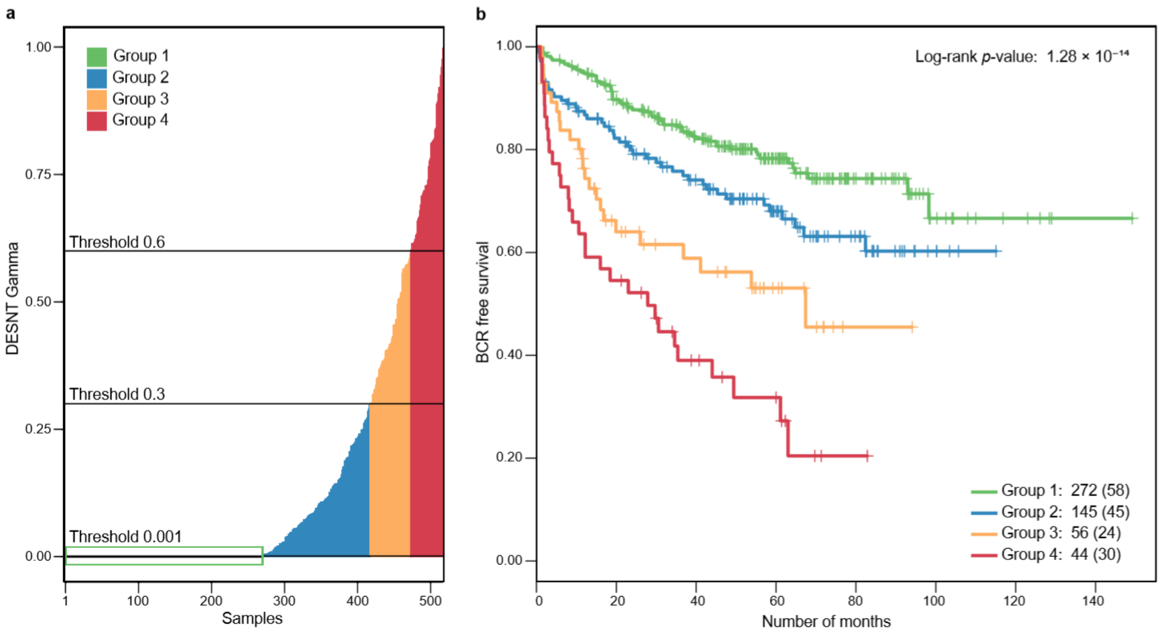
For the purposes of our initial analysis we used all the unique patient samples from the MSKCC, CancerMap, CamCap and Stephenson datasets, removing duplicate patients between datasets randomly. Samples were then split into four groups based on the proportion of their γ assignment to the DESNT group. These proportional groups were:

- **Group 1**: $\gamma < 0.001$
- **Group 2**: $0.001 \leq \gamma < 0.3$
- **Group 3**: $0.3 \leq \gamma < 0.6$
- **Group 4**: $0.6 \leq \gamma$

Kaplan-Meier survival curves were produced for each of the four datasets previously listed (shown in Appendix A.3 A-D). Log-rank p-values were then calculated for each dataset to determine if there was a significant difference between proportional DESNT groups and their associated BCR failure rates. In all of the datasets the p-values were statistically significant (MSKCC 1.74×10^{-3}, CancerMap 8.42×10^{-5}, CamCap 3.16×10^{-5} and Stephenson 1.18×10^{-3}), however some of the groups contained far fewer samples than others.

To ensure the result was robust we combined the four datasets (removing duplicate patients at random) and produced a new Kaplan-Meier plot, as shown in Figure 5.8. We found a strong correlation between an increase in DESNT association and decreased BCR free survival time (p=1.28×10^{-14}; Log-rank test).

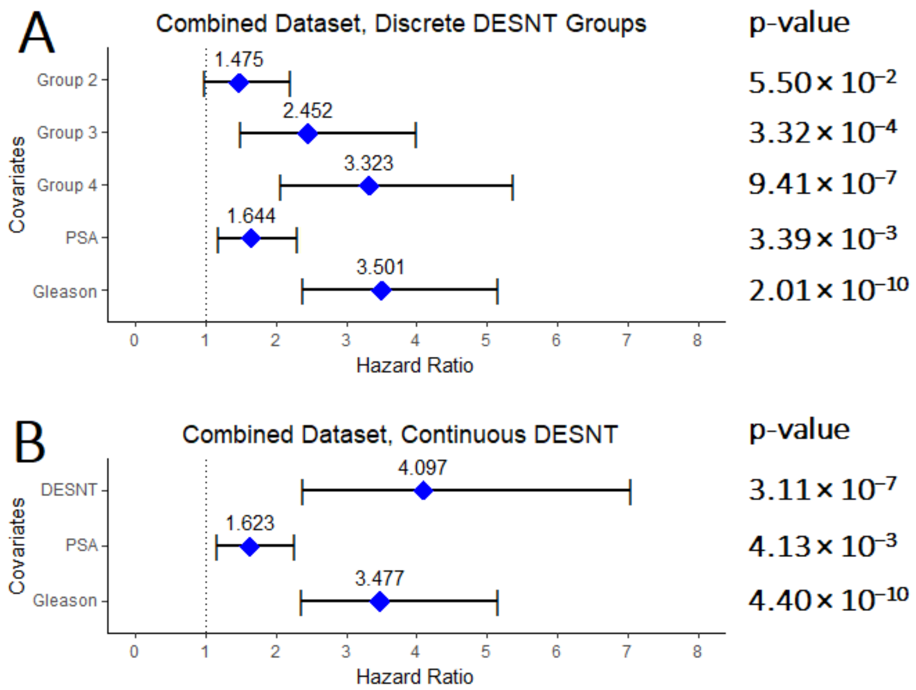
Fig. 5.8 A) An ordered barchart showing the DESNT γ of every sample used in the accompanying Kaplan-Meier survival plot. **B)** A Kaplan-Meier survival plot using all unique samples from the MSKCC, CancerMap, CamCap and Stephenson datasets, split using the four proportional assignment groups. Published in Luca et al. (2020) [5].



A multivariate Cox PH model was produced for the four discretised DESNT groups using the combined dataset, with PSA level and Gleason score as covariates (Figure 5.9 A). The proportional groups were then reassembled to test DESNT γ as a continuous variable.

A second Cox PH model (Figure 5.9 B) was generated from this continuous variable to determine if the proportion of DESNT membership could be used as an independent predictor of BCR failure. The model showed that a high DESNT γ is associated with a higher hazard ratio (HR) than a Gleason score of 8 or higher (DESNT γ HR 4.097 with p-value 3.11×10^{-7} and Gleason HR 3.477 with p-value 4.40×10^{-10}). The log-rank test performed on this second Cox PH model strongly suggests that the proportion of DESNT γ is statistically viable as a predictor of BCR failure, in addition to Gleason score and PSA level.

Fig. 5.9 Cox PH models for the combined prostate cancer dataset, formed from the unique patients in the MSKCC, CancerMap, CamCap and Stephenson datasets, where duplicate patients were removed randomly. The blue markers indicate the hazard ratio for each covariate and the extended bars represent the 95% confidence interval. The log-rank p-value for each covariate is displayed on the right side of the figure. **A)** The covariates were all discretised. The base case for each of the Group variables was $\gamma < 0.001$. Samples were assigned to Group 2, 3 and 4 in the range $0.001 \leq \gamma < 0.3$, $0.3 \leq \gamma < 0.6$ and $0.6 \leq \gamma$ respectively. PSA was split on $(\leq / > 10)$ and Gleason was split on $(\leq / > 7)$. **B)** DESNT represents the continuous range of DESNT γ from 0 - 1. PSA was split on $(\leq / > 10)$ and Gleason was split on $(\leq / > 7)$.



5.5 Biopsy DESNT Analyses

The desired outcome for utilising DESNT in clinical practise is to use DESNT γ to separate the aggressive and non-aggressive tumours prior to treatment. To date all research into DESNT has been performed using samples from patients that have undergone radical prostatectomy. To achieve the desired clinical test, patients must be tested for DESNT status using only their biopsy samples. The result of this test would determine which patients require radical prostatectomy.

Testing new biopsy samples for DESNT status would require rebuilding the LPD model to include the new samples. This would be inefficient from the perspective of both time and computational resource allocation. As a result of rerunning LPD the current model would also undergo small changes and need to be revalidated. To circumvent these issues we employ our novel form of LPD (OAS-LPD), as described in Chapter 3 and in Luca et al. (2020) [5], to classify 20 new biopsy samples for DESNT status. Unfortunately at the time of writing we do not have access to BCR status and metastasis status for these patients and can only use Gleason as a clinical comparison.

5.5.1 Biopsy Samples

FFPE biopsy samples were obtained from 22 unique patients across a range of Gleason scores (Gleason 3+4 to Gleason 5+4). Patient age was distributed between 52 years to 77 years, with a mean age of 70.05 years. Of these samples 20/22 were assessed to have good RNA yields, the remaining 2 samples were removed from the study. Samples were grouped using the new 5 grade group system (Table 5.3). A summary of our assignments can be found in Table 5.4, highlighting a higher proportion of high grade samples.

Gleason Score	New Grade Group
≤ 6	1
3+4	2
4+3	3
8	4
≥ 9	5

Table 5.3 A summary of the new Gleason grade groups. Epstein et al. (2016) [8].

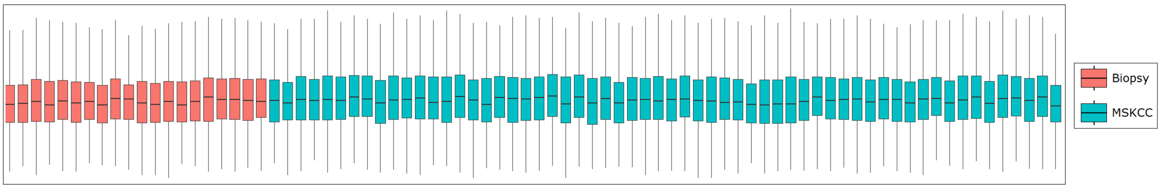
Gleason Grade Group	Number of Biopsy Samples
1	0
2	4
3	2
4	7
5	7

Table 5.4 A summary of the prostate biopsy samples summarised into grade groups.

5.5.2 Applying the DESNT OAS-LPD model

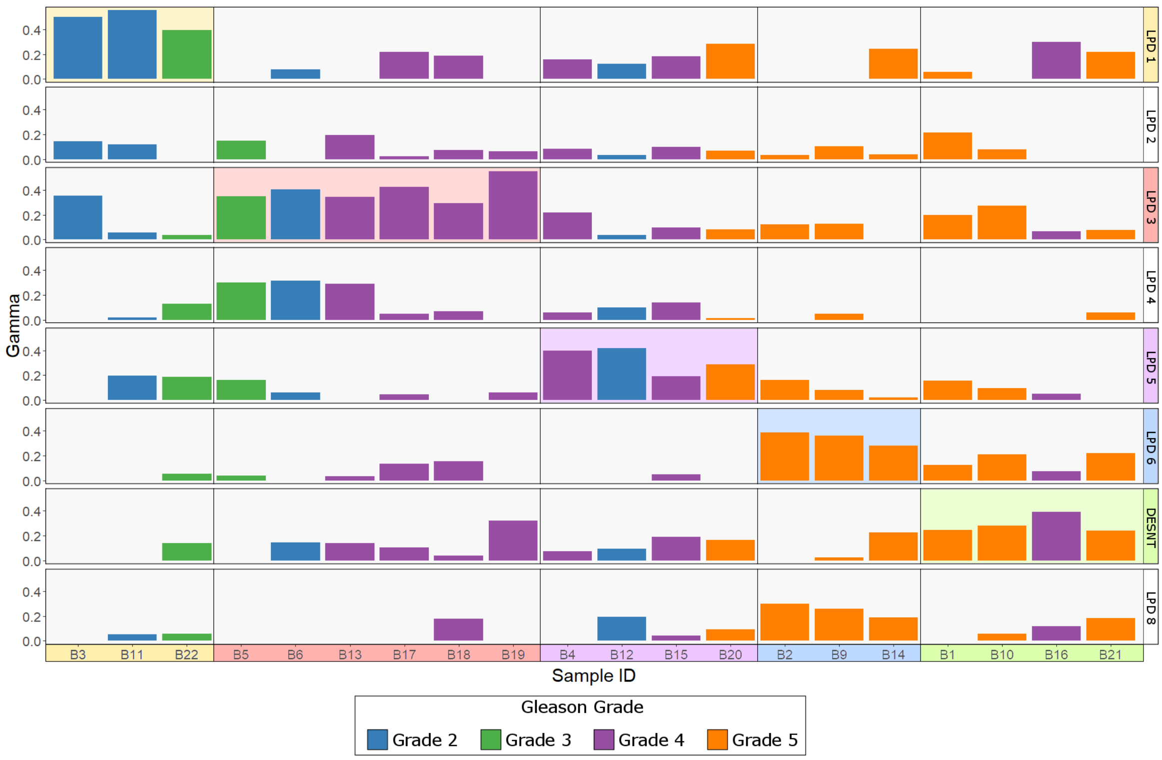
Before classifying the biopsy samples we had to ensure the samples were batch normalised against our original datasets. The process began by grouping the biopsy samples into a new dataset and applying RMA. The resulting expression values were further normalised by application of reference ComBat and reference Quantile normalisation. Within the ComBat normalisation the pre-normalised MSKCC, CamCap, CancerMap, Stephenson, Klein, Erho, Karnes and TCGA datasets were combined into a single reference batch along side the biopsies in a separate new batch. To complete the normalisation we then took the quantile normalised MSKCC data and used it as a reference to quantile normalise the biopsy samples.

Fig. 5.10 Boxplots showing the mean and 95% confidence intervals for the 20 normalised biopsy samples and a random selection of 60 normalised samples from the MSKCC dataset, due to the limited space on the page.



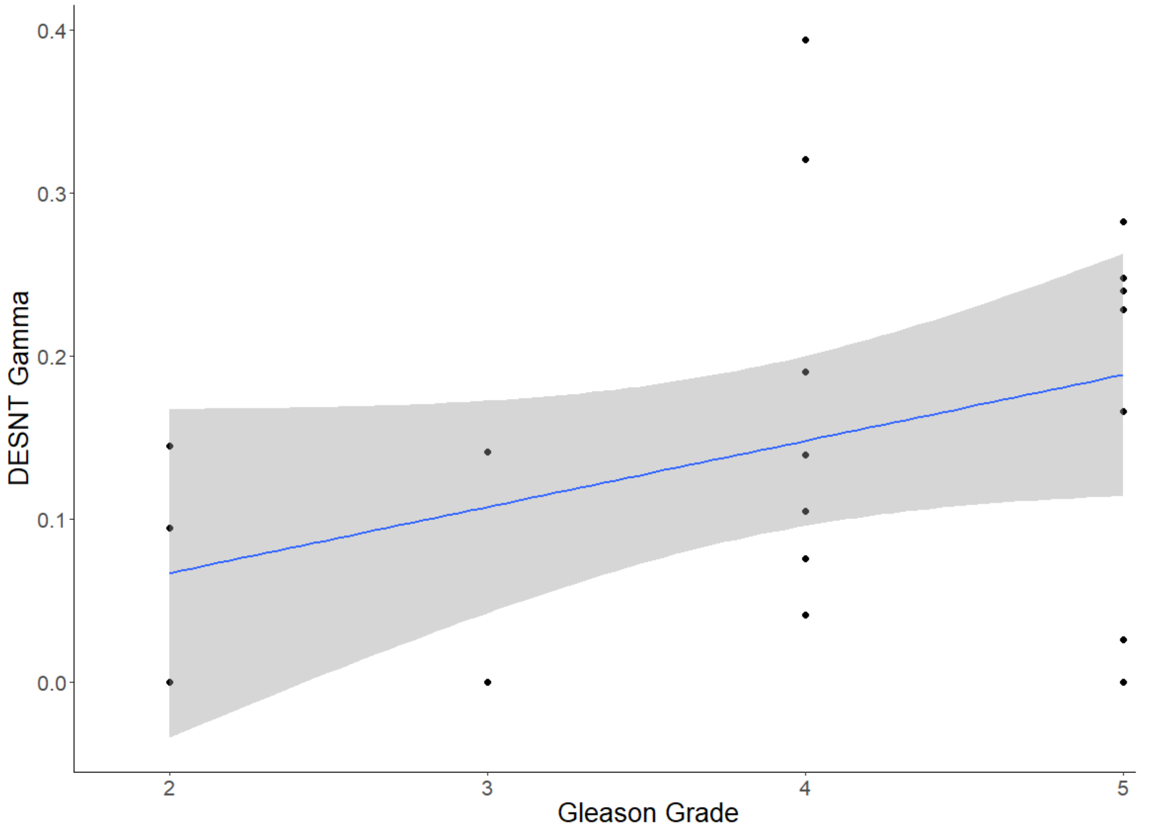
We began to classify the biopsy samples by constructing an OAS-LPD model. To construct the model we derived the original MSKCC representative LPD model's μ_{gk}, σ^2_{gk} and α parameters and set these values to be immutable within an OAS-LPD model. The normalised biopsy samples were then run through OAS-LPD using this model 100 times and the representative of these 100 runs was identified (Figure 5.11).

Fig. 5.11 Bar plots showing the LPD γ values for the association between each biopsy sample and OAS-LPD process. The OAS-LPD process primarily associated with each biopsy sample has also been highlighted, in addition to the Gleason Grade of each given biopsy.



By assigning each sample to LPD process with the greatest γ value we were able to test whether Gleason Grade was dependent on the primary process of each sample. Fisher's exact test (p-value = 0.0256) provided the evidence required to reject the null hypothesis that Gleason Grade was independent of the LPD primary process. This was the expected result based on our previous DESNT work and further supports the idea that biopsies can be used with OAS-LPD to assess patient risk. We also analysed DESNT γ as a continuous variable within the biopsy samples using Pearson's correlation. The result from this test (correlation = 0.395, 95% CI = -0.0573 - 0.713 and p-value = 0.0846) suggests a weak positive correlation between DESNT γ and Gleason Grade within the biopsies (Figure 5.12). However, due to a relatively low number of samples and wide spread of DESNT γ values this correlation is not statistically significant.

Fig. 5.12 Scatter-plot comparing OAS-LPD DESNT γ and Gleason grade for 20 prostate cancer biopsy samples. The blue line denotes the Pearson's correlation and the shaded region the 95% confidence region.



5.6 Discussion

In this chapter we used the 8 process LPD classification by Luca et al. (2017) [17] to further analyse the DESNT subtype using both prostatectomy and biopsy samples. We presented the genes associated with the DESNT subtype and the pathways these genes were involved in. Finally, we presented a novel method (OAS-LPD) for classifying new prostate cancer samples using the existing 8 process LPD classification.

Within the set of genes that comprise the DESNT gene signature are a number of down-regulated genes known to encode proteins that are components of the actin cytoskeleton and facilitate actomyosin contractility. The identification of increased malignancy in DESNT tumours could correlate with an increase to cell migratory behaviours reliant on particular cytoskeletal machinery, however the down-regulation of these genes suggests an alternative migration method is utilised. Within the DESNT signature are a number of genes involved in focal adhesion, *Integrin α5* (ITGA5), *Vinculin* (VCL) and *Integrin-linked Kinase* (ILK). These genes may instead facilitate mesenchymal type migration with E-cadherin mediated cell-cell adhesion mechanisms [162].

In addition to the previously discussed pathways, to which the majority of the DESNT signature genes belong, are a number of genes related to various other transcription factors that can be associated with one or more hallmarks of cancer. One of these genes encodes the *Endothelial PAS Domain Protein* (EPAS1), a transcription factor involved in the induction of oxygen regulated genes implicated in the development of blood vessels [163]. Two other genes of interest are *ETS Proto-Oncogene 2* (ETS2) and *Signal Transducer and Activator of Transcription* (STAT5B), which are partially responsible for regulating apoptosis within cells [163]. These genes may therefore play an important role in the development of the poor prognosis DESNT subtype.

In this chapter we have also demonstrated that the risk of BCR in prostate cancer patients can be determined by analysing the proportion of DESNT γ present in prostatectomy samples.

An increase in DESNT γ is seen to strongly correlate with a decrease in BCR-free survival and is also shown to be an independent predictor of risk, with a greater covariate hazard ratio than current prostate cancer risk measures (PSA and Gleason).

To begin assessing the viability of using LPD to predict risk in a clinical setting we obtained 20 biopsy samples, with good RNA yields. We also modified the LPD algorithm (OAS-LPD) to classify new samples within a pre-existing LPD model. We found that a three stage normalisation process involving RMA, reference ComBat and reference Quantile normalisation was able to adequately normalise new samples' individual gene and sample distributions similar to the levels of the reference dataset(s). This normalisation allowed the new samples to be run through the novel OAS-LPD method to produce a set of Bayesian classifications based on the existing MKSCC LPD model's processes, which were originally published by Luca et al. (2017) [17].

From the OAS-LPD results we were able to establish a correlation between the OAS-LPD processes and Gleason grades in the biopsy samples. This result mirrors the findings found within the prostatectomy samples and warrants the need for further large scale biopsy studies to address the limitations relating to the range of DESNT γ values and restrictive clinical data of this study. Overall we have demonstrated the potential strength of applying LPD to prostate cancer in a clinical setting and believe DESNT γ can be used to improve the targeting of treatment to reduce the over treatment of low risk patients.

Chapter 6

Colorectal Cancer (CRC)

6.1 Summary

In this chapter we discuss key information regarding the colon and colorectal cancer, including risk factors and current disease treatments. This information is vital to understanding the benefits of molecular testing in colorectal cancer, before introducing our own novel molecular classification of colorectal cancer in Chapter 7.

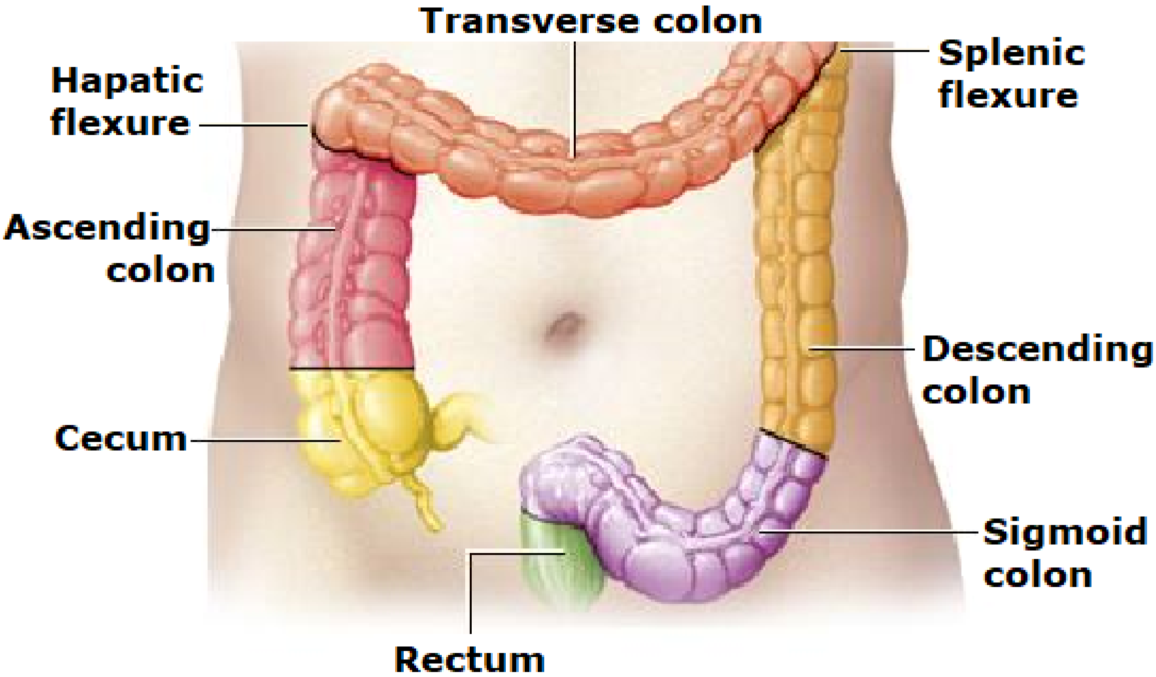
6.2 The Colon

The colon is a long-tube like organ that forms the last part of the gastrointestinal tract. Its purpose is to extract water and electrolytes from solid waste and can be split into two main sections and eight subsections [164] (as shown in Figure 6.1):

- The Proximal colon: Starting at the cecum and ending at the splenic flexure, includes the cecum, ascending colon, hepatic flexure, transverse colon and splenic flexure.

- The Distal colon: Starting from the descending colon and ending at the rectum, includes the descending colon, sigmoid colon and rectum.

Fig. 6.1 A diagram detailing the sections of a colon. Adapted from Mayo Clinic [6].



The cecum is a pouch that connects the small intestines to the large intestines on the right side of the body via the ascending colon. The ascending colon runs up the right side of the body from the cecum to the heptic flexture and is the first section of the colon responsible for extracting water from solid waste. The solid waste is transported up the ascending colon through the process of *peristalsis* (a wave of muscle contraction and relaxation). The transverse colon begins at the hapatic flexure and spans across the abdominal cavity to the splenic flexure. Approximately 41% of CRC cases occur within the proximal colon [165].

The descending colon begins at the splenic flexure and runs down the left side of the body to the sigmoid colon and is responsible for storing faecal matter before it is emptied into the rectum. Below the descending colon is the sigmoid colon, named after its S-shaped structure. The sigmoid colon contains muscular walls that contract to apply pressure on the faecal matter, pushing the compressed stool into the rectum below. Approximately 22% of CRC cases occur within the distal colon down to the rectum with a further 28% of cases occurring within the rectum. The remaining 8% occur in other sites [165].

6.3 Colorectal Cancer

6.3.1 Risk Factors

There are many well established risk factors associated with colorectal cancer. As with many other types of cancers these factors include age, race/ethnicity and family history, however CRC risk has also been associated with specific diets [166] and genetic mutations [167].

Age is a large risk factor in the development of many diseases. In the case of CRC the risk of developing cancer increases significantly after 50 years of age, with the mean age of diagnosis varying globally between 65-75 years [168, 169]. CRC is a disease predominately found in the elderly as demonstrated by the sharply increasing age standardised incidence rates per 100,000 people in the UK between 2015-2017 (1.8, 41.6 and 386.7 cases for age ranges 20-24, 50-54 and 80-84 years respectively) [170].

While older age has been associated with CRC development, it does not explain why incidence rates vary around the world. Race/ethnicity has been shown to be another major risk factor for CRC development, that begins to explain the variable global risk. In the USA the racial group with the highest risk of CRC development are Black people with an age standardised risk of 45.7 per 100,000 people. This figure was 18.7% lower in non-Hispanic White people and 34.4% lower in Asian Americans [171]. In partial contrast, within the UK the racial group with the highest risk of developing CRC was White people, with an average age standardised risk of between 44.1-45.1 per 100,000 people. This was significantly higher than rates found in Black and Asian people, 15.2-22.8 and 25.1-37.7 per 100,000 respectively [170]. It should be noted that the range of standardised rates was attributed to a 17% unknown ethnicity in the population analysed.

Although racial populations show variable relative risk across countries, these changes could be partially explained by the proportion of generations present in each cohort. Studies conducted across the world have concluded that migrates to any given country have a different risk of developing CRC compared to the native population. These differences are reduced over time, with the risk to subsequent generations converging to the average for that country [172, 173].

The differences in generation specific risks are more likely attributable to environmental factors rather than genetic factors due to the relatively short timespan between generations. Of the potential environmental factors, diet stands out as a factor that could have an impact on the health of a patient's colon through direct contact. The main dietary condition widely accepted as causing an increased risk of CRC is the consumption of a western diet (a diet containing high volumes of red and processed meats, high-fat dairy products, refined grains and desserts) [174].

Genetic mutations can also confer a difference in CRC risk, with the main two hereditary conditions being hereditary nonpolyposis colorectal cancer (HNPCC or Lynch syndrome) and familial adenomatous polyposis (FAP). Lynch syndrome is a defined by a mutation in at least one of four mismatch repair genes (*MSH1*, *MSH2*, *MSH6* and *PMS2*), which confers up to an 80% life time risk of developing CRC [175]. FAP is a hereditary condition caused by the mutation of the *APC* gene on chromosome 5q21 [176] and accounts for approximately 1% of all CRC cases [175].

6.3.2 Screening and Early Detection

Population screening has been a controversial topic yielding mixed results for many diseases. However, results from CRC screening have shown improvements to patient survival in many countries [177–179]. The two main screening techniques used to identify new CRC cases in the UK are the faecal occult blood test (FOBT) and faecal immunochemical test (FIT) [180]. These tests measure the amount of blood found in a patient's faeces, with a positive result indicating a significant amount of blood detected. While a positive test can indicate CRC, it can also be caused by ulcers, haemorrhoids, benign polyps, swallowing blood or from inflammatory bowel disease [181]. A positive test therefore requires further tests and examination to confirm the CRC status of a patient.

6.3.3 Diagnosis

Patients presenting with CRC symptoms are initially tested for occult blood in their faeces. NICE recommends that patients with a positive FOB test are then referred for a colonoscopy due to the lack of sensitivity in the initial test [182]. Biopsies may be taken during the colonoscopy for further molecular testing if polyps or other growths are identified. These molecular tests are useful tools to initialise an investigation into hereditary markers (such as those for HNPCC and FAP), or to assess the disease severity using sporadic markers (discussed later in this chapter).

A computed tomography (CT) scan can also be performed in patients with a suspected or confirmed case of CRC to further establish the locations and sizes of their tumours in both the colon and other parts of the body [183, 184]. A virtual colonoscopy using CT scanners can be offered in some cases instead of a physical colonoscopy to reduce the invasiveness of the procedure [183].

6.3.4 Classification criteria

Once a patient has been diagnosed with CRC the severity of the disease must be established in order to guide treatment pathways for the patient. CT scans can advise whether the disease is currently contained locally, or whether it has spread to distant sites around the body. In the latter case analysis of the metastases will drive alternative classification and treatment pathways [185].

Tissue from a colonoscopy biopsy extraction provides a much greater depth of information to explore at both a pathological and molecular level. By studying the biopsy a pathologist is able to determine the key features related to CRC tumours and gain an initial understanding of the disease progression. These pathology results may not be as accurate as those performed later on the surgically removed tumour, as such the biopsy study results are referred to as the *clinical grade/score* and the colectomy study results are referred to as the *pathological grade/score* [186].

6.3.4.1 Tumour Node Metastasis

Tumour Node Metastasis (TNM) classification is the standard pathological system for staging malignant CRC tumours by the American Joint Committee on Cancer (AJCC) and Union for Interventional Cancer Control (UICC). It comprises of three parts that are described by the American Cancer Society [187] and Cancer Research UK (CRUK) [188] as:

- **T:** describes the primary tumour size, whether the tumour has spread into the wall of the colon/rectum and if so how many layers have been invaded.
- **N:** describes the spread into regional lymph nodes.
- **M:** describes the presence or otherwise of distant metastatic spread into distant lymph nodes or other organs.

These three parts can be broken down further to provide detailed descriptions of individual CRC cases:

T - Primary Tumour	
TX -	Primary tumour cannot be assessed.
T0 -	No evidence of primary tumour.
Tis -	Cancer cells only found in the epithelium of the colon.
T1 -	Tumour invading submucosa.
T2 -	Tumour invading the muscularis propria.
T3 -	Tumour penetrating the muscularis propria and the subserosa.

T4 -	Tumour directly invading other organs or structure.
T4a -	Tumour penetrating visceral peritoneum.
T4b -	Tumour directly invading or adhering to other organs or structures.
N - Regional Lymph Nodes	
NX -	Regional lymph nodes cannot be assessed.
N0 -	No regional lymph node metastases.
N1 -	Regional lymph node metastases.
N1a -	Tumour present in 1 regional lymph node.
N1b -	Tumour present in 2 or 3 regional lymph nodes.
N1c -	Tumour present in regional structures that are not lymph nodes.
N2 -	Regional lymph node metastasis in 4 or greater lymph nodes.
N2a -	Tumour present in 4 - 6 regional lymph nodes.
N2b -	Tumour present in 7 or greater regional lymph nodes.
M - Distant Metastasis	
M0 -	No Distant metastases.
M1 -	Distant metastases.
M1a -	Distant metastases in 1 other part of the body.
M1b -	Distant metastases in more than 1 other part of the body.
M1c -	Peritoneal metastases.

Table 6.1 Tumour Node Metastasis (TNM) classification system for CRC.

AJCC TNM staging can be used to stratify patients into similar risk groups based on the progression of the disease. Table 6.2 below outlines the AJCC stages using the detailed TNM information, with higher level stages conveying increased patient mortality risk.

AJCC Stage	TNM Criteria				
0	Tis N0 M0				
I	T1 or T2 N0 M0				
IIA	T3 N0 M0				
IIB	T4a N0 M0				
IIC	T4b N0 M0				
IIIA	T1 or T2 N1 or N1c M0	**OR**	T1 N2a M0		
IIIB	T3 or T4a N1 or N1c M0	**OR**	T2 or T3 N2a M0	**OR**	T1 or T2 N2b M0
IIIC	T4a N2a M0	**OR**	T3 or T4a N2b M0	**OR**	T4b N1 or N2 M0

IVA	Any T Any N M1a
IVB	Any T Any N M1b
IVC	Any T Any N M1c

Table 6.2 AJCC / TNM staging for CRC.

6.3.4.2 Dukes' Staging

While the AJCC TNM staging system is the most common system used to describe CRC progression, some doctors will use a simpler system when discussing results with their patients [189]. One such system is the Dukes' staging system used by some UK doctors. Dukes' staging collapses the TNM stages into the four categories outlined in Table 6.3.

Dukes' Stage	AJCC Stage
A	I
B	IIA, IIB, IIC
C	IIIA, IIIB, IIIC
D	IVA, IVB, IVC

Table 6.3 AJCC stages grouped by Dukes' stage for CRC.

6.3.5 Localised and Regional Disease Treatment

Unless otherwise specified, the following subsections related to disease treatments are based on the NICE guidelines for colorectal cancer treatment [185].

Patients with localised low risk colorecal cancer (T1/T2, N0 and M0) are offered various forms of surgical resection to remove the CRC tissue if the tumour is resectable. This represents the gold standard treatment for localised CRC. NICE do not recommend pre-operative radio-/chemo-therapy for low risk patients, but recognise a small improvement to the relapse free survival of higher risk patients (T1/T2, N1/N2 and M0, or T3/T4, any N and M0) following surgery with pre-operative therapy [190].

6.3.5.1 Surgical Resection

Patients with localised colorectal cancer are offered a selection of three forms of surgical resection. The first two options are called *transanal excision* (TAE) and *endoscopic sub-mucosal dissection* (ESD). Both TAE and ESD are performed using endoscopic surgery to minimise the invasiveness of the procedures and limit the typical hospital stay to just 1-2 days. In each case the aim of surgery is to only remove the cancerous tissue and not the lymph nodes or colon. An additional benefit to TAE and ESD is a reduction to the number of possible complications that could arise from the surgery, however this comes at the risk of requiring further surgery at a later date.

In some cases a decision may be made for a more invasive form of surgical resection called *total mesorectal excision* (TME). The greatest difference in TME surgery, compared to TAE and ESD, is the aim to remove both the cancerous tissue and a portion of the surrounding colon. It is unusual to require further surgery following TME due to the extent of the surgery, however TME does not come without its own risks and negative impacts. Patients opting for TME will typically spend 5 to 7 days in hospital recovering from the surgery and be at risk of more severe complications, such as *anastomtic leaking* (leaking of the bowel into the abdomen). Removing part of the colon may also require the patient to undergo a *colonstomy* to create a *stoma* (redirecting part of the colon to a permanent or temporary bag attached to an opening in the abdomen), resulting in a potentially life long consequence.

6.3.5.2 Radiotherapy and Chemotherapy

Using radiotherapy alone to treat CRC is an uncommon practice. It is typically applied in combination with other treatment options and is most commonly used to treat cancers found in the rectum [191], or in patients with tumours that extend to the lining of the abdomen in combination with chemotherapy before surgery. The intention of this combined therapy is to reduce the size of the tumour and make it easier to remove during surgery. Patients that experience complete clinical and radiological response to neoadjuvant treatment may also be offered the option to defer surgery, but are warned of the additional risk of recurrence.

Chemotherapy is typically used as an adjuvant therapy alongside surgery, since surgery is the gold standard primary treatment in most CRC cases. A report advising NICE [192] considered six independent studies to ascertain that high risk stage II and stage III CRC patients could benefit from adjuvant chemotherapy to reduce overall systemic recurrence. However, the report also warned that the studies themselves were of relatively low quality. Current NICE guidelines recommend only offering adjuvant chemotherapy to patients with stage III colorectal cancers (T1-T4, N1-N2, M0), due to the associated negative side effects of chemotherapy treatment.

Many chemotherapy agents exist for the treatment of CRC including Capecitabine, Oxaliplatin and Fluorouacil. NICE recommends using a combination of Cepecitabine and Oxliplatin (CAPOX) wherever possible due to CAPOX's reduced treatment costs and shorter treatment time compared to other chemotherapy treatments. Where this is not possible because of a patient's histopathology NICE instead recommends the use of Flurouacil combined with Oxaliplatin and Folinic acid (FOLFOX).

6.3.6 Metastatic Disease Treatment

Unless otherwise specified, the following section on metastatic disease treatments is based on the NICE guidelines for colorectal cancer treatment [185].

Metastases are the leading cause of mortality in CRC patients, with a 5-year survival rate of approximately 15% compared to 90% for patients with localised disease [193]. Although overall 5-year survival rates are low, palliative treatment options exist to help alleviate the symptoms of metastatic CRC and in some cases are used to cure specific subtypes. Approximately 20% of patients undergoing systemic therapy with metastatic CRC will experience primary tumour related symptoms such as pain, bleeding and obstruction of the colon. Patients can be offered surgical resection of their primary tumour to prevent these symptoms from manifesting, but at a 5% risk of severe postoperative complications that may delay their systemic therapy.

Patients may be offered additional location specific treatment for their metastases alongside treatment of their primary tumour. The four main metastatic locations with recommended treatments are metastases of the bones, liver, lung and peritoneal. Treatment of bone metastases follows the same treatment pathway as all solid tumours (other than prostate). The treatment involves using bisphosphonate drugs to prevent the loss of bone density. Denosumab is recommended as the primary choice of bisphosphonate even though it has a higher cost, as its overall cost effectiveness is lower than other tested bisphosphonates. This reduced cost is attributed to the statistically significant reduction to first skeletal-related events (HR 0.84, p=7×10^{-4}) [194].

Liver metastases follow a different three layer strategy involving surgery, systemic anticancer therapy (SACT) and selective internal radiation therapy (SIRT). Surgical resection offers the best 5-year survival for CRC liver metastatic patients, with approximately a 30%-50% 5-year survival rate [195]. In addition to resection of both the primary and metastatic liver tumours, NICE recommends the use of SACT to improve overall and disease free survival. Only in special cases when no other treatment options are viable do they advise the use of SIRT, due to inadequate research into the benefits of SIRT [196].

Metastases within the lungs are treated with either metastasectomy (surgical removal), ablation (generating heat through an electrical current), or stereotactic body radiation therapy (SBRT). Due to a lack of high quality evidence NICE do not currently recommend the use of one of these treatments over the others. Unlike other CRC metastases, those found in the peritoneal are primarily advised to undergo SACT, rather than resection. Medical staff are also encouraged to refer patients to specialist centres that can offer potentially curative cytoreductive surgery and hyperthermic intraperitoneal chemotherapy (HIPEC), however NICE acknowledges the mixed results of research in these areas.

6.3.7 Genetic Alterations and Biomarkers

Colorectal cancers can be broadly catergorised into two distinct forms of genetic alteration, those that occur sporadically and those that are hereditary. Sporadic alterations account for approximately 65% of all CRC and can be classified into three distinct mechanisms discussed below [197]. Hereditary alterations can be split into two main conditions, those belonging to *familial adenomatous polyposis* (FAP), including attenuated familial adenomatous polposis (AFAP) and *MYH*-associated polyposis (MAP) and those belonging to hereditary nonpolyposis colorectal cancer (HNPCC).

The first sporadic mechanism is *chromosomal instability* (CIN), which accounts for approximately 80%-85% of sporadic CRCs and can be described as the partial or complete duplication, or deletion, of one or more chromosomes [198–200]. The second mechanism, *microsatellite instability* (MSI), accounts for up to 15% of all sporadic CRCs and can be defined as frequent mutations occurring in microsatellite loci [198, 199]. The third mechanism is known as epigenetic gene silencing, this can also be referred to as *CpG island methylator phenotype* (CIMP) when the mechanism occurs frequently at promoter CpG islands and is present in approximately 30% of CRCs [201]. CIMP status has a large overlap with MSI status and unique combinations of the two mechanisms yield several associations with other clinical variables, as shown in Tables 6.4 & 6.5 [9].

<table>
<tr><th></th><th>CIMP-H</th><th>CIMP-L</th><th>CIMP-0</th></tr>
<tr><td>MSI-H</td><td>Group 1:
10%</td><td colspan="2">Group 2:
5%</td></tr>
<tr><td>MSI-L</td><td rowspan="2">Group 3:
5-10%</td><td>Group 4:
5%</td><td rowspan="2">Group 6:
40%</td></tr>
<tr><td>MSS</td><td>Group 5:
30-35%</td></tr>
</table>

Table 6.4 Table showing the percentages of cases for each group of CIMP and MSI combinations. CIMP-H, CIMP-L and CIMP-0 refer to high, low and very low methylation levels respectively. MSI-H, MSI-L and MSS refer to high, low and no microsatellite instability respectively. Adapted from Ogino 2008 [9].

Group	**Associations**
1	MLH1 and BRAF mutations, CIN negative, proximal colon, elderly females, good prognosis.
2	KRAS mutation, CIN negative, proximal colon, HNPCC.
3	BRAF mutation, CIN negative, right colon, elderly females, poor prognosis.
4	MGMT methylation, KRAS mutation.
5	KRAS mutation, CIN negative, males.
6	Wild-type KRAS and BRAF, CIN positive, distal colon.

Table 6.5 Table highlighting the main CIMP-MSI group associations. Adapted from Ogino 2008 [9].

6.3.7.1 Hereditary CRC

Familial adenomatous polyposis results in the formation of hundreds, or thousands of adenomatous polyps caused by a dominantly inherited mutation of the *APC* gene. Due to this, the risk of developing CRC in patients with untreated FAP is 95% by the age of 50 and almost a 100% risk across their full life time [197].

Cases where the patient presents with only 10-100 adenomatous polyps within the proximal colon may be considered an attenuated form of FAP (AFAP). This condition has a reduced life time risk of 70-80% and can be managed without colectomy in one third of cases. As with FAP, AFAP is caused by dominantly inherited mutations within the *APC* gene that are typically located within the extreme ends of the gene's DNA sequence. *MYH*-associated polyposis is phenotypically similar to AFAP, but is caused by recessive/biallelic mutations within the *MYH* gene. Around 20% of cases involving more than 20 and fewer than 500 colonic adenomatous polyps can be attributed to MAP [202].

Hereditary nonpolyposis colorectal cancer, also known as Lynch syndrome is the most common hereditary form of CRC and accounts for 1-3% of all CRC cases. It is an autosomal dominantly inherited condition that occurs from the mutation of at least one of four mismatch repair (MMR) genes (*MLH1*, *MSH2*, *MSH6* and *PMS2*) [203]. Over 90% of these cases contain mutations in one of the first two genes and 6% of cases are caused by a mutation of the third gene. PMS2 mutations occur rarely within HNPCC, making up only a small percentage of cases, but testing for *PMS2* mutations has also been shown to improve the sensitivity of *MLH1* mutation detection [197, 203]. Due to the loss MMR gene functionality in Lynch syndrome, MSI occurs in over 90% of cases [203].

All patients with confirmed cases of CRC are referred for immunohistochemistry or MSI testing when they are first diagnosed [204]. These tests are performed for two main clinical reasons. First to determine the Lynch syndrome status of the patients and second to determine the MMR and MSI status of the patients. The primarily targets of these tests are the *MLH1* oncogene and *BRAFV600E* mutation [204]. It is important to determine both MMR and MSI status as they each confer a resistance to Flurouacil adjuvant treatment [205–207].

6.3.7.2 Sporadic CRC

Cancers require multiple mutations in order to develop, however the estimated average mutation rate per nucleotide base pair is not sufficient enough to generate all of these mutations [208]. Cells must first develop genomic instability before acquiring the mutations required to progress to carcinoma [209, 210]. The two main pathways to developing sporadic CRC are the adenoma-carcinoma sequence and the serrated pathway [211].

Adenoma-carcinoma describes the progression from normal tissue, to small adenomas, to large adenomas, to eventually forming cancerous tissue and is predominately associated with the development of CIN-positive CRCs [211]. The first mutational step in this sequence involves a mutation within the *APC* gene found on the long arm of chromosome 5. *APC* mutations occur in up to 75% of all sporadic colorectal cancers [198] and result in either a truncated non-functional *APC* protein, or even complete allele loss [209]. Loss of this gene/protein produces an overactivation of the Wnt/β-catenin signalling pathway and causes irregular cell proliferation [211].

Following the loss of *APC*, subsequent mutations in the *KRAS* oncogene encourage adenoma growth. A single nucleotide substitution within the KRAS gene can cause it to bind with Guanoisine-5'-triphosate (GTP), resulting in the propagation of growth factors and an activation cascade within the *MAPK/ERK* signalling pathway [212]. This pathway is normally responsible for regulating signals from cell surface receptors and the nucleus of a cell, such as signals to promote cell division. By destabilising the regulation of these signals the KRAS mutation causes further irregular cell proliferation. The adenomas must then undergo the functional lose of tumour suppressor genes / miss match repair genes such as *Tp53*, *MLH1* and *MGMT* to progress to CRC [7]. While this sequence of events is associated with CIN-positive cancers, it remains unclear whether the mutational sequence encourages CIN development, or whether CIN is a precursor allowing these mutations to occur [211].

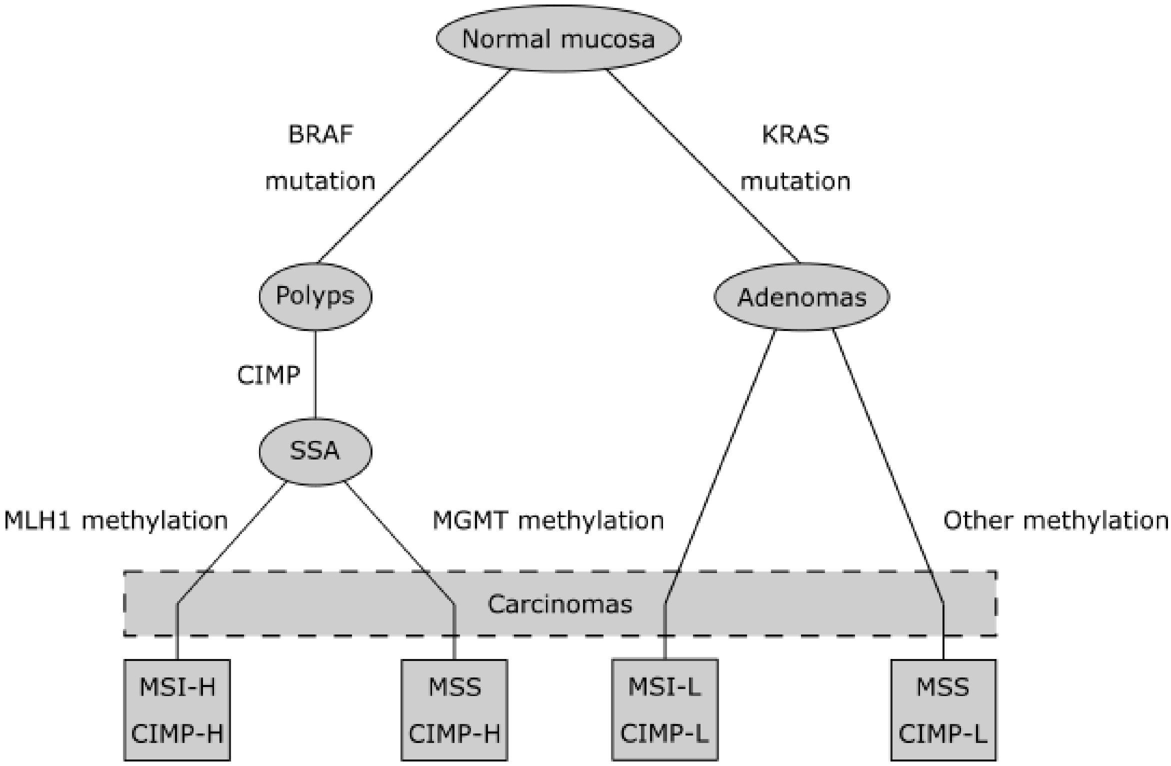
The serrated pathway for CRC development begins with normal tissue growing hyperplastic polyps that can later progress into sessile serrated adenomas, before finally forming cancerous tissue. Hyperplastic polyps are the result of a mutation within the *BRAF* oncogene [213, 211]. In up to 90% of CRC cases involving a BRAF mutation, thymine is substituted with adenine at nucleotide position 1799 [214]. This initial molecular event activates the MAPK pathway in an uncontrolled manor through the binding of BRAF and Adenosine triphosphate. The resulting signal cascade promotes cell proliferation, prevents apoptosis and results in the formation of hyperplastic polyps [211]. The second key molecular event is CIMP, which drives the polyps to serrated adenomas and CRCs [213]. Approximately 75% of sessile serrated adenomas and and 90% of serrated adenocarcinomas present CIMP-positivity [211]. Figure 6.2 provides a visualisation of key molecular events and their resulting carcinomas.

The main sporadic CRC pathways present distinct molecular events and mutations. These unique characteristics allow analysis and classification of patients to better inform their individual clinical management. The KRAS mutations and BRAF mutations found almost exclusively in CIN-positive and CIMP-positive cancers respectively are important markers for predicting patient response to different therapies. Mutation of either the KRAS or BRAF genes confers a resistance to anti-epidermal growth factor receptor monocolonal antibodies (anti-EGFR MoAbs) and patients presenting these mutations must instead undergo alternative treatment [215–217].

6.4 Discussion

In this chapter we have explored the biology of the colon and the risks known to be associated with colorectal cancer. We have reviewed the current diagnosis/classification criteria for

Fig. 6.2 A diagram detailing the sporadic CRC molecular event pathways. Adapted from Szylberg et al. (2015) [7].



colorectal cancer and discussed the key molecular markers guiding patient treatment options. In the next chapter we will produce a novel classification framework for colorectal cancer and demonstrate its usefulness in predicting patient risk alongside current factors.

Chapter 7

Deriving Molecular Subtypes in Colorectal Cancer

7.1 Summary

In the previous chapter we discussed the risk factors currently associated with colorectal cancer (CRC), including the main genetic and epigenetic pathways that commonly lead to the development of CRC. In this chapter we apply the LPD algorithm to over 2,000 colorectal cancer samples to establish a novel classification of the disease. We establish four main subtypes, characterised by unique molecular profiles. We find that each subtype presents an independent clinical outcome and that the poor prognosis subtype (Pericol) could be used to aid clinical decision making by assessing the risk of disease recurrence.

Comparisons with existing publications reveal a large overlap between our Pericol transcriptomic signature and many other published signatures. Two of our other subtypes are also observed to be closely related to two unique groups within the CIMP-MSI model discussed in Chapter 6.3.7. These findings further substantiate the hypothesis that LPD can robustly identify subtypes within heterogeneous diseases and be used to improve the diagnosis of such diseases.

7.2 Materials

7.2.1 Datasets

The work in this chapter was performed using five microarray datasets from the GEO repository: GSE14333 [218], GSE17536 [219], GSE39582 [220], GSE41258 [221] and GSE81653 [222] (Table 7.1). RNA-seq and methylation data from The Cancer Genome Atlas repository (TCGA-COAD [223]) was also used to analyse the novel classifications.

We identified 133 patients in common between the GSE14333 and GSE16536 datasets. To prevent the introduction of bias we removed these duplicate patients from the GSE17536 dataset and combined the remaining samples with GSE14333 to form GSE14333plus. Similarly, 67 samples from the TCGA-COAD dataset were removed due to patient duplication or missing clinical data. Further details regarding these datasets can be found in Table 7.1.

<table>
<tr><th>Dataset</th><th>Samples</th><th>Primary</th><th>Normal</th><th>Tissue Type</th><th>Platform</th></tr>
<tr><td rowspan="2">GSE14333plus [218, 219]</td><td>290</td><td>290</td><td>0</td><td>FF</td><td>Affymetrix HG U133 Plus 2.0</td></tr>
<tr><td>44</td><td>44</td><td>0</td><td>FF</td><td>Affymetrix HG U133 Plus 2.0</td></tr>
<tr><td>GSE39582 [220]</td><td>585</td><td>566</td><td>19</td><td>FF</td><td>Affymetrix HG U133 Plus 2.0</td></tr>
<tr><td>GSE41258 [221]</td><td>240</td><td>186</td><td>54</td><td>FFPE</td><td>Affymetrix HG U133A</td></tr>
<tr><td>GSE81653 [222]</td><td>593</td><td>593</td><td>0</td><td>FFPE</td><td>Affymetrix HG 2.0 ST</td></tr>
<tr><td>TCGA-COAD [223]</td><td>454</td><td>205</td><td>197</td><td>FF/FFPE</td><td>Illumina RNA-Seq</td></tr>
</table>

Table 7.1 Table summarising the unique samples from the datasets used in this chapter.

7.2.2 Clinical Data

All five datasets contain associated data regarding the type of the sample (primary tumour or normal tissue), however all other clinical information was available at varying levels of detail for each dataset. The GSE81653 dataset contained recurrence status without any further follow-up data, or other information regarding known clinical variables. Due to the severely limited clinical data for the GSE81653 dataset we decided to only use it in the production of the LPD models and exclude the samples from any further follow-up analyses.

The GSE14333plus, GSE39582 and GSE41258 datasets all contained relapse-free survival or disease-free survival (DFS) information, allowing us to perform survival based analyses on these datasets using DFS as the endpoint. The TCGA-COAD dataset only contained overall-survival data and could not therefore be included in any survival based analyses using the other datasets. A summary of the other key clinical variables can be found in Table 7.2.

	GSE14333plus	**GSE39582**	**GSE41258**	**GSE81653**	**TCGA-COAD**
Gender					
Male	192	322	108	0	240
Female	142	263	122	0	214
Unknown	0	0	10	593	0
Age					
Median	66	69	65.5	NA	68.81
Mean	65.84	66.95	63.11	NA	67.44
Range	26-92	22-97	19-87	NA	31-90
CRC Location					
Distal	161	351	133	0	171
Proximal	125	232	97	0	263

Table 7.2 continued from previous page

	GSE14333plus	GSE39582	GSE41258	GSE81653	TCGA-COAD
Unknown	48	2	10	593	20
TNM Stage					
I	45	32	35	0	78
II	98	271	58	0	181
III	92	210	60	0	131
IV	99	60	77	0	64
Unknown	0	0	0	593	0
Relapse Events					
Event	55	179	50	234	0
No Event	210	395	180	359	0
Unknown	69	11	10	0	454
MSI					
MSI-H	0	0	41	0	11
MSI-L	0	0	19	0	
MSS	0	0	151	0	81
Unknown	334	585	29	593	362
MMR					
Proficient	0	459	0	0	30
Deficient	0	77	0	0	28
Unknown	334	49	240	593	396
CIMP					
Positive	0	93	0	0	0
Negative	0	420	0	0	0

Table 7.2 continued from previous page

	GSE14333plus	GSE39582	GSE41258	GSE81653	TCGA-COAD
Unknown	334	72	240	593	454
CIN					
Positive	0	369	0	0	0
Negative	0	112	0	0	0
Unknown	334	104	240	593	454
TP53					
Mutant	0	190	119	0	0
Wild-Type	0	161	67	0	0
Unknown	334	234	54	593	454
KRAS					
Mutant	0	217	0	0	22
Wild-Type	0	328	0	0	24
Unknown	334	40	240	593	408
BRAF					
Mutant	0	51	0	0	3
Wild-Type	0	461	0	0	25
Unknown	334	73	240	593	426

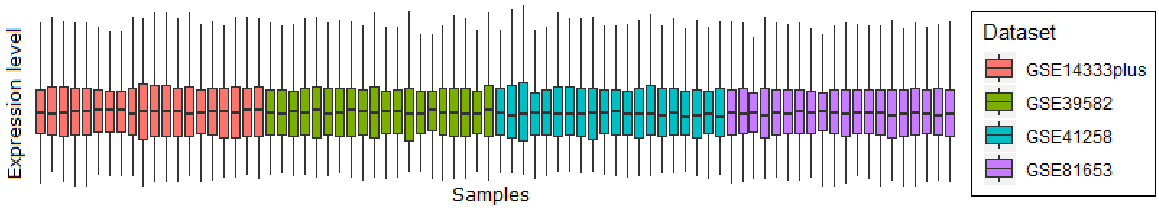
Table 7.2 Table summarising the clinical data associated with the samples used in this chapter.

7.2.3 Dataset Pre-processing

Before beginning the data normalisation we downloaded the raw CEL files for our five microarray datasets, GSE14333, GSE17536, GSE39582, GSE41258 and GSE81653 from the GEO repository. The GSE14333 and GSE17536 datasets were combined as described in Section 7.2.1 to form GSE14333plus. To begin normalising the four microarray datasets we applied the RMA algorithm, described in Section 3.2.2, from the *Oligo* R Bioconductor package [224]. To begin normalising the RNA-seq dataset we applied a variance stabalising and log2 transformation, using the *DESeq2* R Bioconductor package [225].

To further mitigate the dataset specific differences in expression intensities we employed the *ComBat* algorithm, described in Section 3.2.3, from the *sva* R Bioconductor package [226]. We used ComBat on all four of the RMA normalised datasets and the RNA-seq dataset, treating each dataset as an individual batch. Finally we applied Quantile normalisation across all the datasets to ensure a similar distribution of gene expression levels across all samples. A selection of normalised samples can be seen in Figure 7.1.

Fig. 7.1 Boxplots depicting 20 random normalised samples from each of the GSE14333plus, GSE39582, GSE41258 and GSE81653 datasets.



Due to the normal distribution of microarray data and the binomial distribution of RNA-seq data there were concerns as to whether we would be able to use the RNA-seq data with LPD, or be able normalise the two sources of data together. Initially we applied the preceding normalisation steps to the microarray data in isolation. However, after attempting the normalisation with both microarray and RNA-seq data we established this was viable and proceeded with the combined normalisation for the remainder of this thesis. For completeness, the original LPD processes (produced without using RNA-seq data) are provided in Appendix B.1.

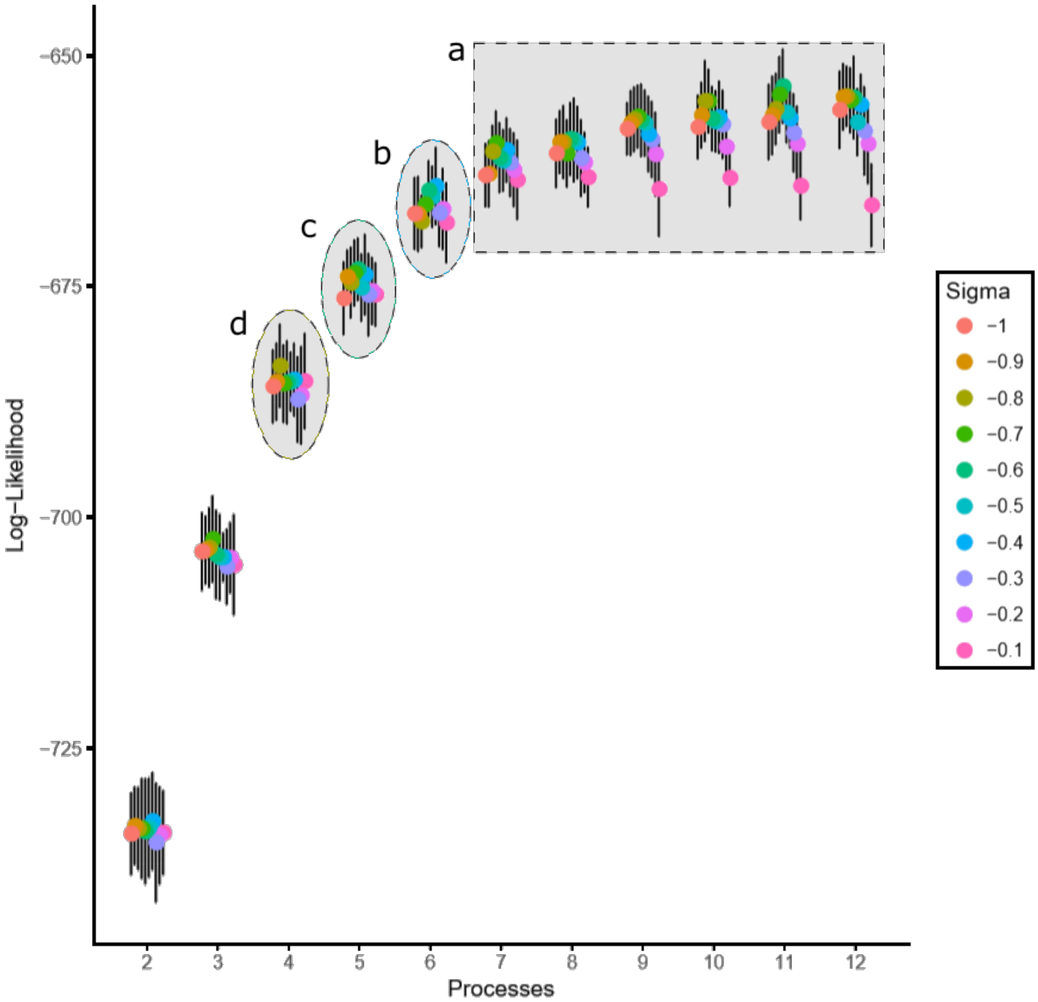
7.3 Creating the LPD Models

In order to classify the CRC samples we first identified the 500 genes with the greatest variance across all our datasets. We calculated the mean expression level of all the probesets that mapped to each of these 500 genes (as discussed in Section 2.5) and used these means as the input for each LPD run. A limit of 500 genes was implemented to counteract the computationally intensive limitation of LPD, which prohibited every gene from being used. The number of input genes was selected based on the previous successful applications of LPD [18, 17, 3] that were able to achieve distinct subtypes using 500 most variable genes.

7.3.1 Choosing LPD Parameters

As discussed in Section 3.3.3 there are two forms of LPD. Out of these two forms the MAP model is more suitable to use as it helps to prevent over-fitting. With this in mind we dedicated additional computing power during the LPD parameter selection phase to allow us to use the MAP model to optimise the parameter selection. We ran LPD 50 times for every combination of σ (between -0.1 to -1.0, increment -0.1) and the number of processes (between 2 to 12, increment 1). The mean log-likelihood was calculated from all the repeats for each parameter combination. We then plotted the log-likelihoods and assessed them. For each number of processes we selected the σ value corresponding to the model with the maximum log-likelihood.

Fig. 7.2 Figure depicting the log-likelihoods of each parameter combination for the GSE14333plus dataset. **a)** The log-likelihood plateau. **b-d)** The three groups of input parameters selected for further analysis.



To avoid over-fitting the data it is important to consider the principle of Occlam's razor, "plurality should not be posited without necessity" [227]. We therefore aimed to select the model with the greatest discriminatory power between individual subtypes, while aiming to minimise the total number of subtypes. To assess each of the models we calculated the Pearson correlations between the subtypes within each of the three models approaching the log-likelihood plateau (Figure 7.2). The number of processes used to generate the model with the lowest mean Pearson correlations between its subtypes was selected as the input for the final model. The input parameters for each datasets' final model are listed in Table 7.3.

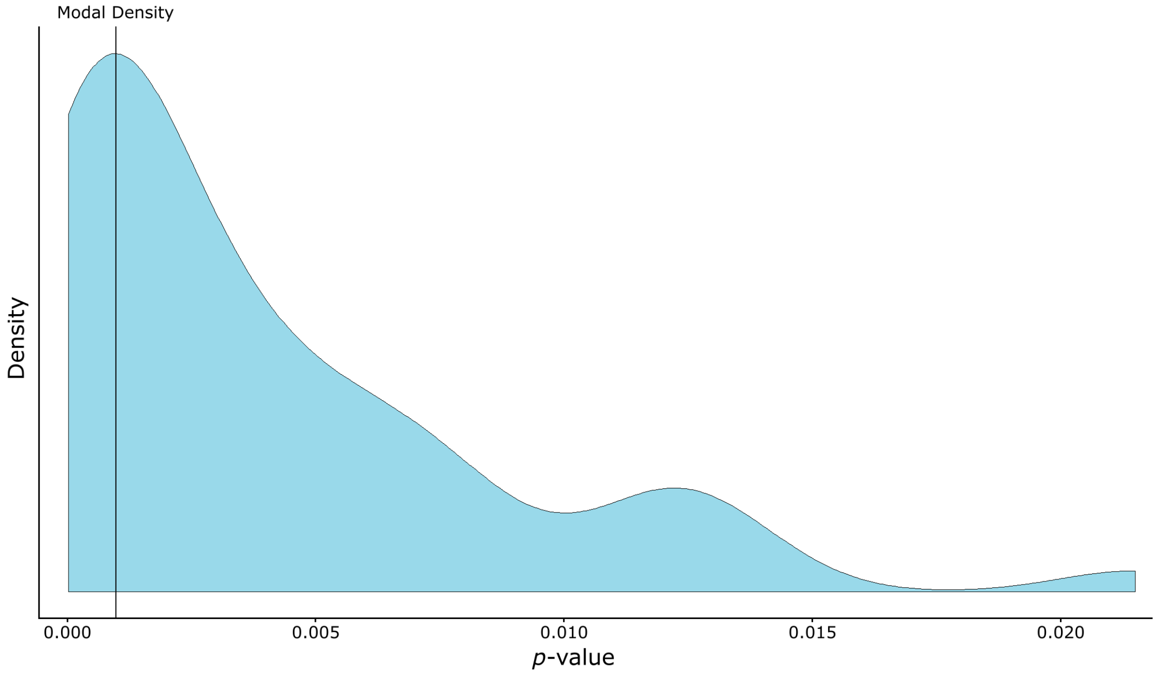
Dataset	σ	**Number of Processes**
GSE14333plus	-0.6	5
GSE39582	-0.5	6
GSE41258	-0.5	5
GSE81653	-0.8	4
TCGA	-0.5	6

Table 7.3 Table summarising the final model parameters for each dataset.

7.3.2 Representative LPD Classification

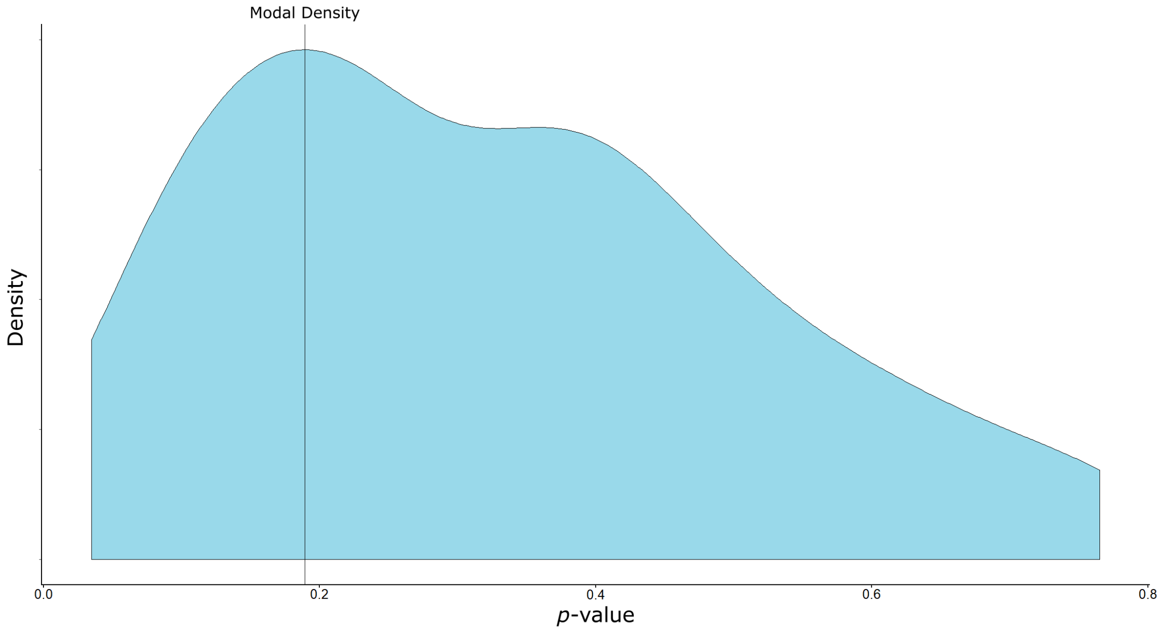
We applied the MAP LPD algorithm (without resampling) a further 100 times per dataset, using the results from Table 7.3, to produce 500 independent classifications of CRC. Each LPD model based on the same dataset exhibited slight variation in sample assignment due to the non-deterministic nature of the LPD algorithm. To account for this variation we selected a representative model of all 100 runs per dataset by performing a log-rank test on all the runs using time to disease relapse as the end point. The LPD model with the log-rank p-value closest to the modal log-rank p-value density was selected as the representative model for each dataset. An example of this for the GSE39582 based models is shown in Figure 7.3 (The remaining density plots for each dataset are shown in Appendix B.2). For the models using the GSE81653 dataset only the disease recurrence status was available, without a measure of time. To determine a representative model for this dataset we substituted the log-rank p-value with the χ^2 p-value, calculated using the contingency table between event status and LPD process.

Fig. 7.3 Figure depicting the identification of a representative LPD run, based on the density of *p*-values from a set of 100 log-rank tests, each performed on an individual LPD model using the GSE39582 dataset. The model with the shortest *p*-value distance to the modal density was selected as the representative LPD run.



A representative run was successfully identified from the set of models based on each independent dataset. However, the LPD runs based on the TCGA-COAD dataset displayed a wide range of log-rank test results (Figure 7.4). While the selected TCGA-COAD LPD run was later identified to strongly correlate with each of the other models' representative runs, we could not confidently state that this run was a true representative for all the TCGA-COAD LPD runs. Due to this, we decided not to use this model in the proceeding analyses.

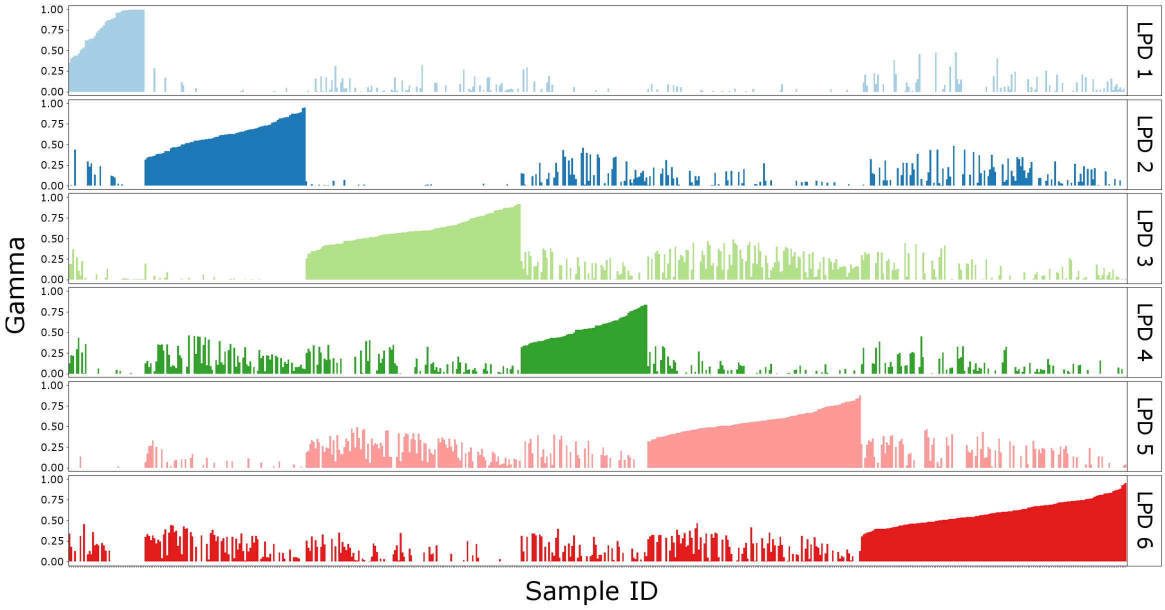
Fig. 7.4 Figure depicting the identification of a representative LPD run, based on the density of p-values from a set of 100 log-rank tests, each performed on an individual LPD model using the TCGA-COAD dataset. The model with the shortest p-value distance to the modal density was selected as the representative LPD run.



7.4 Analysing the LPD Models

Each representative run was made up of unsupervised Bayesian classifications of all samples within each given dataset. As the classifications were Bayesian in nature every sample could belong in part to any number of the derived processes, with each association (γ) to a process comprised of a value between 0 and 1, totalling 1 across all processes (as described in Section 3.3.3). An example of the representative LPD classification results for the GSE39582 dataset are illustrated in Figure 7.5 (all LPD models are shown in Appendix B.3).

Fig. 7.5 Figure depicting the LPD γ values (association between a sample and a process) for each LPD process in the GSE39582 representative run. Samples have been grouped by their process with the greatest γ value for ease of viewing.



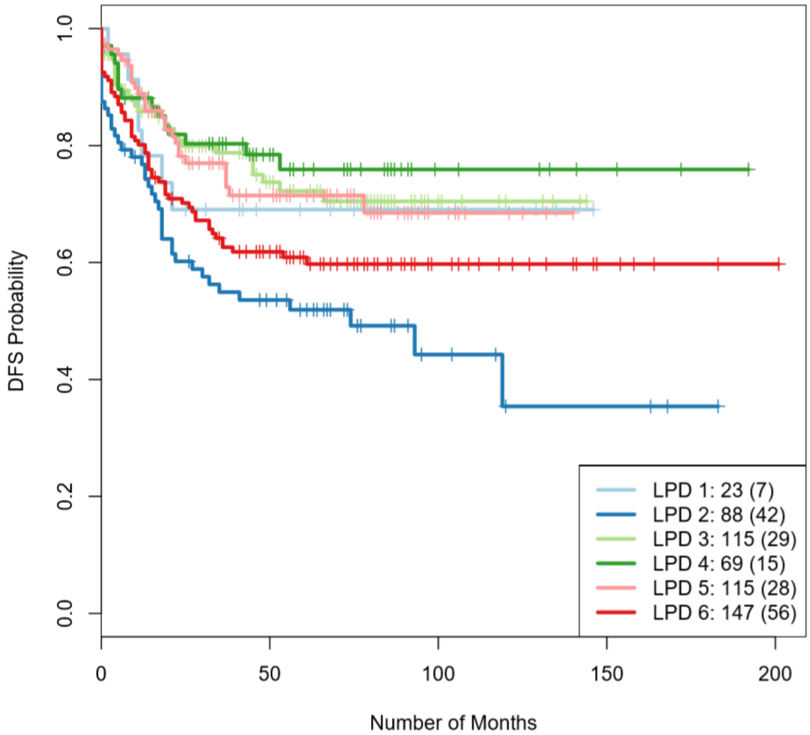
For each sample we examined the set of γ values and consider the process with the greatest γ value the primary process. All 19 normal tissue samples from the GSE39582 dataset were primarily associated with process 1 within the GSE39582 LPD model. A mixture of TNM graded primary tumours were also associated with process 1 within this model. All 54 normal tissue samples from the GSE41258 dataset were primarily associated with processes 2 and 3 in the LPD model based on this dataset, with the majority (76%) of normal samples displaying a greater association to process 3. It is worth highlighting that primary association to each of these two processes was exclusive to normal samples. When GSE41258 was run with fewer processes the two exclusively normal processes remained distinct, suggesting an underlying molecular difference in these normal samples.

7.4.1 Comparing LPD Process Survival

The classifications of CRC samples were derived entirely from the gene expression data without influence from any samples' associated clinical data. We were therefore interested in analysing whether the molecular subtypes exhibited differences in survival. We performed survival analyses using the three datasets where appropriate clinical data was available (GSE14333plus, GSE39582 and GSE41258), assigning samples to their primary process.

We produced Kaplan-Meier survival curves (see Chapter 3.4.1) based on these assignments to determine whether there was a significant difference in survival time between the LPD processes. A representative example can be found for the GSE39582 dataset in Figure B.13. Within this . We found that two of the three models demonstrated a significant difference in survival time between the processes (GSE14333plus log-rank p-value 0.0345, GSE39582 log-rank p-value 9.64×10^{-4}). The samples from the GSE41258 dataset showed markedly better survival times than the other datasets (median DFS = 55 months, compared with 34 and 43 months for GSE14333plus and GSE39582 respectively) and did not display a statistically significant difference in survival between the non-normal processes (log-rank p-value 0.171).

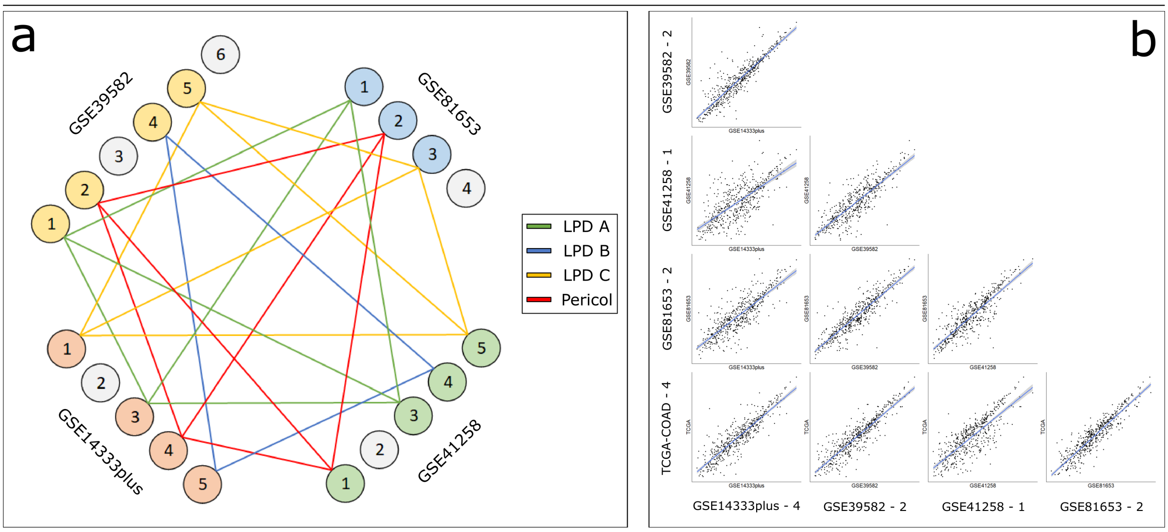
Fig. 7.6 Kaplan-Meier survival curves showing the disease-free survival of the the six LPD processes from the representative run for the GSE39582 dataset. The total number of samples (with DFS information) for each process is shown in the bottom right, with the number of DFS events displayed in brackets.



7.4.2 Identifying Conserved Processes

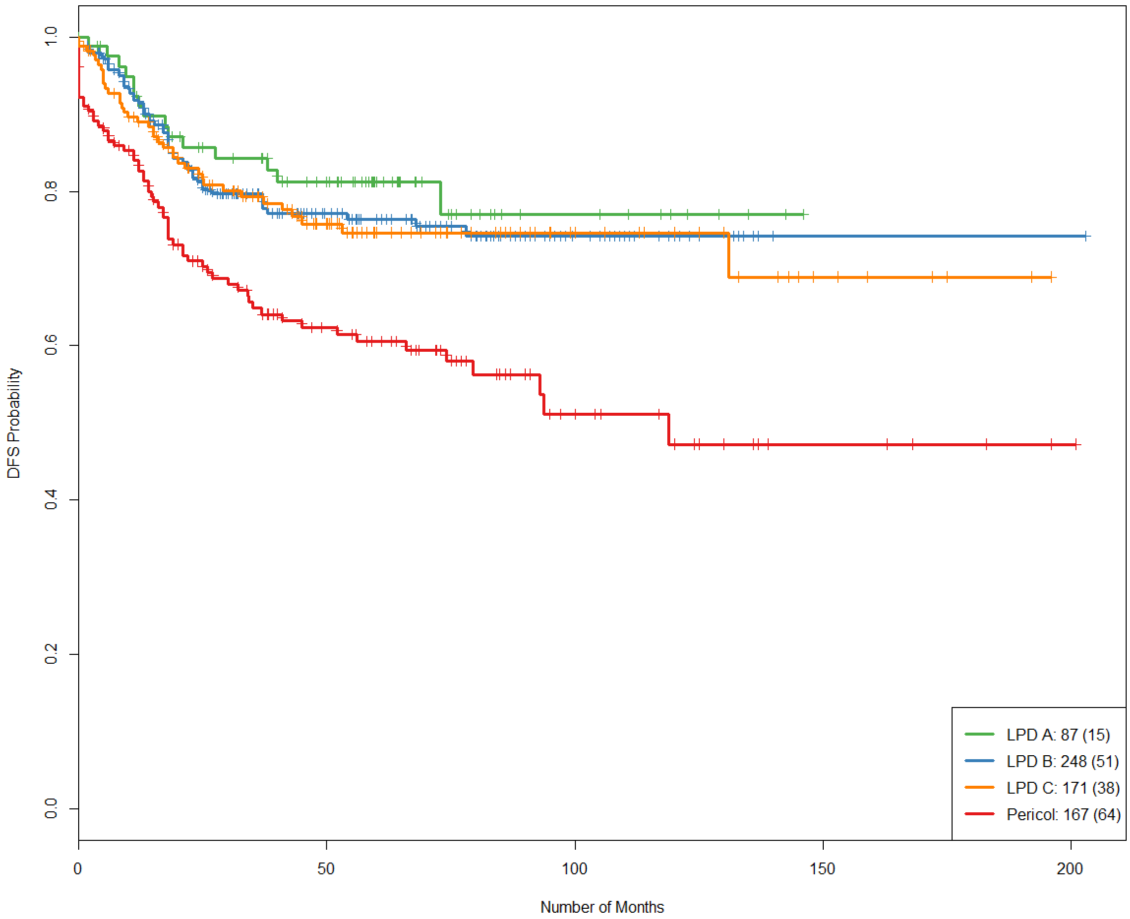
It was important to determine whether any of the processes derived from entirely independent datasets correlated between the representative models. We would expect common molecular processes driving the development of colorectal cancers to be present in all datasets. Identifying common molecular processes would therefore enable us to have greater confidence in the classifications and show that LPD was not just modelling dataset specific artefacts. We calculated the Pearson correlations between the gene expression profiles of all LPD processes from each of the representative models and identified four common subtypes ($r > 0.5$). The non-representative TCGA-COAD based model processes were also compared with each of the representative models to see whether they would have also correlated. They were found to strongly correlate with the four common subtypes, but were not used to derive these common subtypes due to our earlier decision to exclude them from the analysis. Figure 7.7-a depicts the strong positive correlations between each process, highlighting these four common processes. Three of these common processes, henceforth called LPD A, LPD C and Pericol, were present in all four microarray based models as well as the selected TCGA-COAD model. The fourth process, LPD B, was found to correlate significantly between all of the selected models, with the exception of the GSE81753 derived model.

Fig. 7.7 a) Correlation map where each line represents a statistically strong positive correlation between two processes from independent representative models. b) An example of the correlations between all four microarray and TCGA-COAD based models for the Pericol colorectal subtype.



When the datasets were combined on the four common subtypes, they showed a significantly different survival curves (log-rank p-value 8.24×10^{-5}), with Pericol exhibiting significantly worse survival times than the other subtypes (Figure 7.8).

Fig. 7.8 Kaplan-Meier survival plot showing the disease-free survival of the four common processes from the representative LPD runs for the GSE14333plus, GSE39582 and GSE41258 datasets. The total number of samples (with DFS information) for each process is shown in the bottom right, with the number of DFS events displayed in brackets.



7.5 Developing a Consensus OAS-LPD Model

To test samples in a clinical setting would require the use of a fixed/consistent model for all patients. One of the main limitations of using LPD is the need to derive processes from scratch every time new data is added, making it inappropriate in a clinical setting. To overcome this problem we modified the LPD algorithm to create a novel form of LPD called OAS-LPD, as described in Section 3.3.4. This modified version makes the gene expression profiles of a pre-existing LPD model's processes immutable and enables new samples to be assigned to these processes. By removing the need to derive new processes, samples can also be classified in a fraction of the time required to generate the original model. The γ values from an OAS-LPD model are also very stable, reducing the need for multiple reruns.

While the four common processes previously identified showed strong correlations between the four representative models, their expression profiles were not identical. To account for the variation between models we decided to create a consensus model, consisting of four OAS-LPD models, each based on one of the original four representative models. An equally weighted vote/consensus could then be calculated to determine whether all four models would derive the same primary process in each sample.

To begin constructing a consensus model we first extracted the μ_{gk}, σ^2_{gk} and α variables from each representative LPD model. These values were set to be immutable in four new OAS-LPD models in order to conserve the original LPD processes. Each sample from all five datasets was then run through all four OAS-LPD models, resulting in four unique Bayesian classifications per sample.

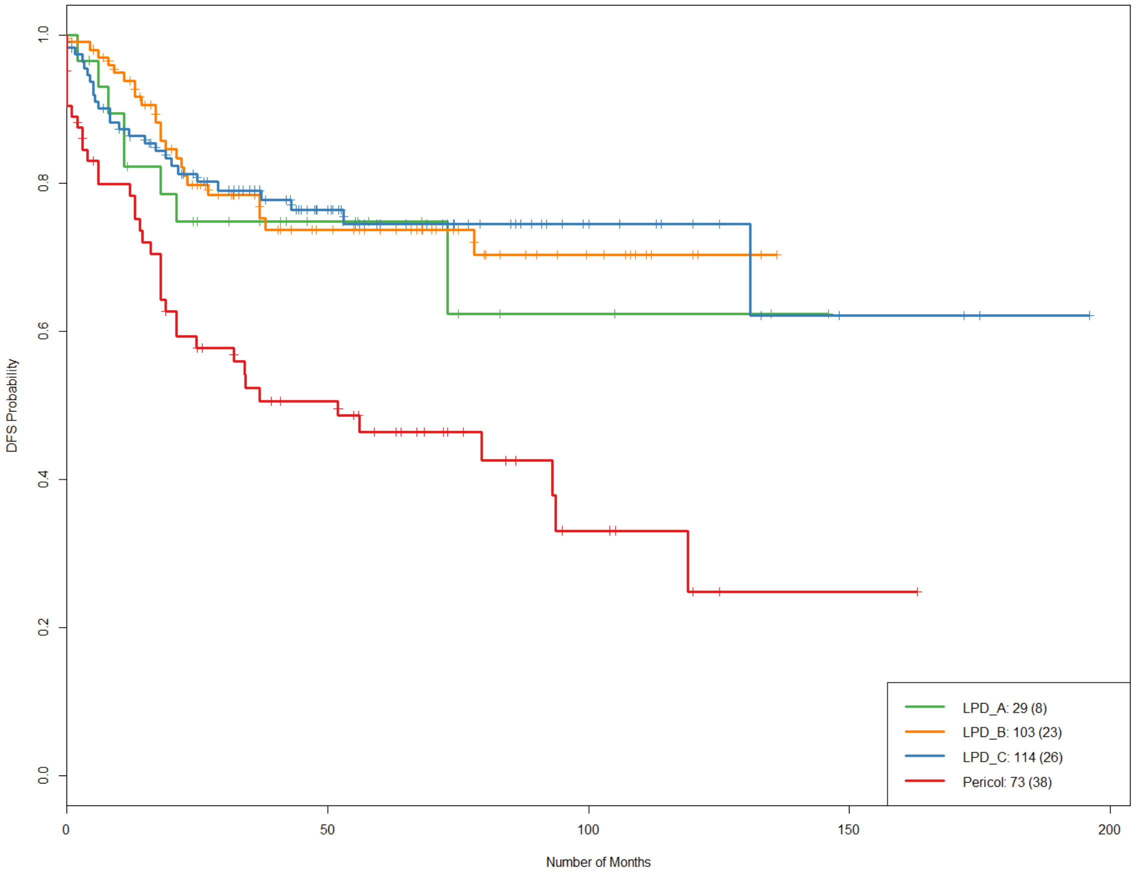
For the purposes of survival analyses we followed the same process as before, assigning each sample to its primary process in each model. The four independent primary process assignments per sample (one for each model) were then assessed to determine whether the four models reached a consensus. If a consensus was reached (all four models agreed on the same subtype) then the sample was deemed to truly belong to the assigned subtype. Table 7.4 summaries the number of samples assigned to each subtype by consensus vote.

	LPD A	**LPD B**	**LPD C**	**Pericol**
GSE14333plus	11	19	37	21
GSE39582	37	63	53	49
GSE41258	39	26	25	9
GSE81653	25	28	30	20
TCGA-COAD	15	31	58	19
Total	127	167	203	118

Table 7.4 A summary of the OAS-LPD consensus assignments.

By generating KM-survival curves from the consensus results it became clear that the four subtypes exhibited significantly different disease-free survival curves to one another (log-rank p-value = 1.43×10^{-5}). The survival curve of the consensus Pericol subtype had also dropped significantly compared to that of the original individual LPD models (Figure 7.9), emphasising the severity of the novel CRC subtype.

Fig. 7.9 Kaplan-Meier survival plot showing the disease-free survival of the four common processes from the consensus OAS-LPD models. The total number of samples (with DFS information) for each process is shown in the bottom right, with the number of DFS events displayed in brackets.



7.6 Analyses of the CRC Subtypes

7.6.1 Novel CRC Subtypes' Clinical Associations

Having assigned the samples from our five datasets to our four CRC subtypes, we began to identify the characteristics of each subtype. We performed a Fisher's exact test on the distribution of each available clinical variable, described in Table 7.2, within our four subtypes (Figure 7.10). Normal tissue samples were exclusively assigned to the LPD A subtype, resulting in an over representation of normal tissue in LPD A (p-value = 2.2×10^{-16}) and the under representation of normal tissue in LPD B, LPD C, and Pericol. Nevertheless, it is important to note that LPD A did not solely consist of normal tissue. LPD A also appeared to have an over representation of TNM stage 1 patients, however this was shown to be narrowly outside the 0.05 confidence cut-off (p-value = 0.0557). The LPD A subtype was not found to have any other significant clinical associations.

A wide range of clinical associations were demonstrated in the LPD B subtype. The subtype showed a strong association with the development of tumours in the distal colon compared to the proximal colon (p-value = 1.28×10^{-6}), but did not show signs of an over representation in any TNM group (p-value = 0.911). When assessing the MSI and MMR status of LPD B the subtype was seen to be microsatellite stable (p-value = 0.0274) and in line with these findings it contained proficient mismatch-repair genes (p-value = 8.08×10^{-5}). CpG islands were not found to be hyper-methylated (p-value = 6.38×10^{-6}), instead LPD B exhibited chromosomal instability (p-value = 1.33×10^{-5}) and an over representation of TP53 mutations (p-value = 1.50×10^{-4}). Finally LPD B was observed to consist of predominately wild-type BRAF tumours (p-value = 1.07×10^{-7}) and did not display any significant difference in mutant and wild-type KRAS (p-value = 0.508).

The LPD C subtype displayed an almost polar opposite set of clinical associations to the LPD B subtype. It contained an over representation of tumours located in the proximal colon

(p-value = 1.63×10^{-10}) and TNM stage 2 tumours, while TNM grades 1 and 4 were found to be under represented (p-value = 1.03×10^{-3}). We identified a strong association between LPD C and deficient MMR genes (p-value = 4.11×10^{-6}) that resulted in high microsatellite instability (p-value = 6.38×10^{-6}). Tumours of this subtype did not exhibit chromosomal instability (p-value = 1.34×10^{-8}) and predominately contained wild-type TP53 (p-value = 2.14×10^{-3}). Hyper-methylation in LPD C CpG islands was significantly overrepresented (p-value = 2.44×10^{-11}) in addition to an over representation of BRAF mutations (p-value = 1.74×10^{-6}). KRAS mutations were not found to be either under or over represented in LPD C (p-value = 0.666).

Pericol tumours were found to mainly be higher TNM stage tumours, with an over representation of stages 3 and 4 (p-value = 0.0200). The mismatch repair genes were found to be proficient (p-value = 2.67×10^{-3}), however MSI status did not exhibit a significant difference in the number of high, low or stable samples (p-value = 0.593). The greatest difference in the Pericol subtype compared to the other three LPD subtypes was the over representation of KRAS mutations (p-value = 0.0486). No significant difference was found in CIMP, CIN, TP53 or BRAF status in the Pericol subtype (p-value = 0.546, 0.322, 0.300 and 0.670 respectively).

Fig. 7.10 CRC subtype clinical associations. The green up arrows highlight the over-representation of a given clinical factor, while the red down arrows highlight the factors as under-represented.

	Tissue	Location	TNM Grade	MSI	MMR	CIMP	CIN	TP53	KRAS	BRAF
Pericol	▲ Tumour ▼ Normal	No current evidence	▲ III & IV ▼ I & II	No current evidence	▲ + ▼ -	No current evidence	No current evidence	No current evidence	▲ Mut ▼ WT	No current evidence
LPD C	▲ Tumour ▼ Normal	▲ Proximal ▼ Distal	▲ II ▼ I & IV	▲ MSI-H ▼ MSS	▲ - ▼ +	▲ + ▼ -	▲ - ▼ +	▲ WT ▼ Mut	No current evidence	▲ Mut ▼ WT
LPD B	▲ Tumour ▼ Normal	▲ Distal ▼ Proximal	No current evidence	▲ MSS ▼ MSI-H	▲ + ▼ -	▲ - ▼ +	▲ + ▼ -	▲ Mut ▼ WT	No current evidence	▲ WT ▼ Mut
LPD A	▲ Normal ▼ Tumour	No current evidence	▲ I ▼ II	No current evidence	No current evidence	No current evidence	No current evidence	No current evidence	No current evidence	No current evidence

7.6.2 Novel CRC Subtypes' Differentially Expressed Genes

In this section we identify sets of genes that were differentially expressed in the samples from each of our four CRC subtypes compared with the samples in each of the other three subtypes. To derive these differentially expressed genes (DEGs) we imposed a stringent set of requirements to help ensure the genes were as robust as possible. The genes had to be differentially expressed in all representative models, they had to be differentially expressed in all 100 LPD repeats and have a false discovery rate below 0.01. The mean expression levels of all the probesets mapping to any given gene were used to derive the DEGs for each subtype. We used all the available probesets within the normalised datasets and did not limit this analysis to only the probesets that mapped to the 500 genes used within the LPD classification.

For the LPD A subtype we identified 139 DEGs within the GSE14333plus model, 5460 DEGS within the GSE39582 model, 5683 DEGs within the GSE41258 model and 2360 DEGs within the GSE81653 model. By calculating the intersection we found a total of 86 genes shared across all the representative models for LPD A (Figure 7.11-a, Table 7.5, Appendix Table B.1). Among the differentially expressed genes identified in LPD A is *TIMP1*. This gene is part of the TIMP gene family, whose normal function involves the degradation of the extracellular matrix, promoting cell proliferation and may also have anti-apoptotic functionality [228]. Unsurprisingly this gene is differentially expressed in many different types of cancer and has been proposed as a non-invasive screening tool in CRC [229]. Another known gene within this set of DEGs is *FCGBP*, which was previously associated with metastatic disease and a decreased overall survival time [230]. This is somewhat surprising given LPD A's relatively good prognosis compared with Pericol.

For the LPD B subtype we identified 3137 DEGs within the GSE14333plus model, 3703 DEGS within the GSE39582 model, 3276 DEGs within the GSE41258 model and 4993 DEGs within the GSE81653 model. By calculating the intersection we found a total of

330 genes shared across all the representative models for LPD B (Figure 7.11-b, Table 7.5, Appendix Table B.2). Within the set of differentially expressed genes is *SPON1*. This gene is mainly expressed in smooth muscle tissue, which surrounds the human colon. The Human Protein Atlas observed a raised expression of this gene in 25% of colorectal cancer patients and found that a high expression conferred better 5-year survival in renal cancer patients (80% compared to 65% in low expression patients) [231]. On the other-hand raised expression of this gene results in a reduction in the 5-year survival of urothelial cancer patients with high *SPON1* expression from 55% to 27% [231].

For the LPD C subtype we identified 787 DEGs within the GSE14333plus model, 2285 DEGS within the GSE39582 model and 114 DEGs within the GSE41258 model. By calculating the intersection we found a total of 26 genes shared across all the representative models for LPD C (Figure 7.11-c, Table 7.5, Appendix Table B.3). The raised expression of the *GBP1* gene in LPD C is another example of LPD identifying subtypes with well documented genes within CRC. The Cancer Genome Atlas Consortium showed that high *GBP1* expression was associated with a reduction in the aggressiveness of CRC tumours [223].

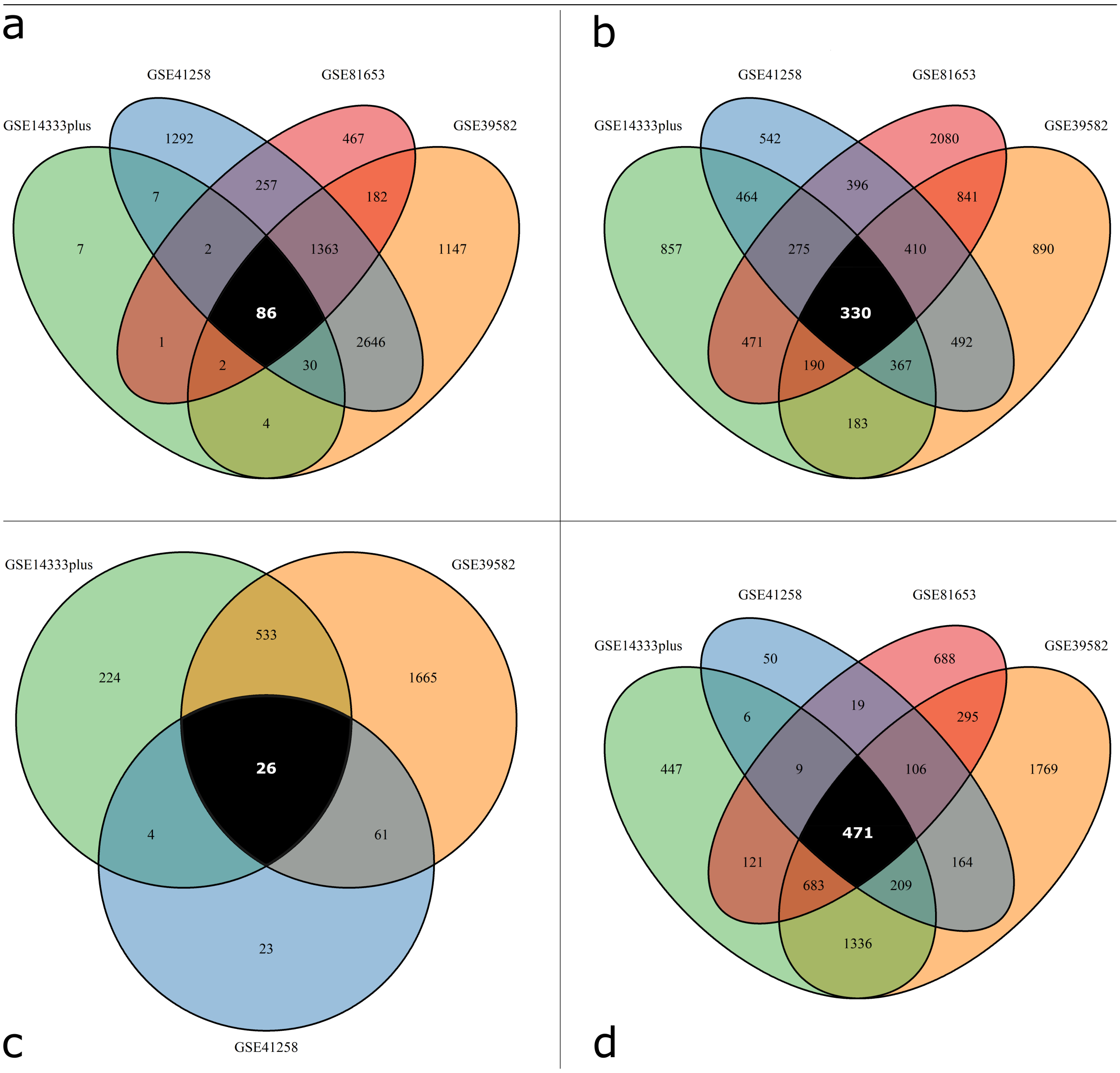
	LPD A	**LPD B**	**LPD C**	**Pericol**
GSE14333plus	139	3137	787	3282
GSE39582	5460	3703	2285	5033
GSE41258	5683	3276	114	1034
GSE81653	2360	4993	NA	2392
Intersection	86	330	26	471

Table 7.5 Table summarising the intersection of differentially expressed genes.

For the Pericol subtype we identified 3282 DEGs within the GSE14333plus model, 5033 DEGS within the GSE39582 model, 1034 DEGs within the GSE41258 model and 2392

DEGs within the GSE81653 model. By calculating the intersection we found a total of 471 genes shared across all the representative models (Figure 7.11-d, Table 7.5, Appendix Table B.4). Alongside LPD A, *TIMP1* was also identified as being differentially expressed in the Pericol subtype. Many of the Pericol DEGs including *TIMP1* were seen to be present in existing colorectal cancer signatures, we discuss this overlap later in this chapter after exploring the enriched pathways these DEGs belong to.

Fig. 7.11 Venn diagrams showing the intersection of differentially expressed genes across each representative model in our four CRC subtypes. a) LPD A. b) LPD B. c) LPD C. d) Pericol



7.6.3 Novel CRC Subtype Pathways

We wanted to identify the biological pathways that were over represented in our sets of differentially expressed genes to help understand the biological mechanisms that were driving the development of our subtypes. To accomplish this we used version 6.8 of the publicly available DAVID Functional Annotation Tool [232, 233]. We used this tool to assess the Gene Ontology (GO) biological processes [234], Kyoto Encyclopedia of Genes and Genomes (KEGG) [159] and Reactome [235] pathways.

Within the LPD A subtype we identified 20 GO biological processes involving the set of differentially expressed genes (Appendix B.5). These processes were primarily involved in the transportation and regulation of salts and other molecules, metabolic processes and organisation of collagen fibrils. We were able to identify six KEGG pathways that were over represented in LPD A (Appendix B.6). These KEGG pathways were also involved in metabolic processes and the reabsorption/reclamation of sodium and biocarbonates. In total five Reactome pathways were observed to be over represented (Appendix B.7), including collagen synthesis and glycosylation.

The LPD B subtype was associated with significantly more processes than LPD A, with 72 GO biological processes observed to be over represented in the set of LPD B DEGs (Appendix B.8). A significant number of these processes were related to angiogenesis and apoptotic regulation. Other processes included extracellular matrix (ECM) organisation, endothelial cell proliferation and organ regeneration. A total of 11 KEGG pathways were identified in LPD B (Appendix B.9). Among these pathways were genes involved in the MAPK signaling pathway and others known to be involved in transcriptional misregulation and proteoglycans found in cancer. Five Reactome pathways were identified to be over represented in LPD B (Appendix B.10). These pathways were primarily associated with cell surface interactions and elastic fibre formation.

No pathways were found to be associated with the genes from LPD C, however Pericol was found to contain the most biological pathways out of all four of our subtypes. A total of 193 GO biological processes were observed within Pericol covering a wide range of functions (Appendix B.11). These pathways can be broadly categorised into processes controlling the organisation, regulation and growth of new and existing cells. Among the top 20 GO biological pathways were genes focusing on angiogenesis, immune responses and ECM organisation. We identified genes involved in 26 KEGG pathways, including ECM receptor interaction, PI3K-Akt signaling and proteoglycans found in cancer (Appendix B.12). Genes involved in small cell lung cancer were also over represented.

A total of 45 Reactome pathways were found to be over represented in Pericol, however many of these were involved in the processes of other diseases (Appendix B.13). Among the top 20 pathways were processes involving ECM proteoglycans, collagen assembly and degregation and the biosynthesis of chondroitin sulfate. We also compared the set of Pericol DEGs with the Online Mendelian Inheritance in Man (OMIM) database of diseases. The presence of differentially expressed *COL6A1*, *COL6A2* and *COL6A3* genes within Pericol suggested a possible association with Bethlem myopathy and Ullrich congenital muscular dystrophy (hereditary conditions involving the progressive weakening of skeletal muscles and connective tissue [236, 237]).

7.6.4 Intersection of Pericol Genes and Published Signatures

The Pericol subtype was identified as offering a poorer prognosis for CRC patients. Focusing on this group and understanding the mechanisms that drive its development therefore have the greatest chance of yielding significant benefits to CRC patients. In this section we compared the 471 differentially expressed genes from the Pericol subtype with 791 unique genes from other published prognostic CRC signatures. We collected published signatures from Oncotype DX [238], Chen et al. (2017) [239], ColoPrint [240], ColDX [241], Gao et al. (2018) [242], Oh et al. (2012) [243], D-Sun et al. (2018) [244], Shu et al. (2018) [245], Chunsheng et al. (2018) [246], D-Sun et al. (2019) [247], Pagnotta et al. (2013) [248] and Pan et al. (2017) [249].

We identified a total of 19 genes shared by any two published signatures (*AKAP12*, *ARHGEF2*, *AXIN2*, *CFTR*, *CYFIP2*, *CYP1B1*, *FAP*, *GPX3*, *KLK10*, *KRT17*, *NHLRC3*, *NT5E*, *POSTN*, *PPARA*, *QPRT*, *REG4*, *SCG2*, *SFRP2* and *SPON1*). Only the *KLK10* gene was shared by three published signatures (Gao et al, D-Sun et al and Shu et al) and has additionally been shown to play a role in the suppression of tumourigenesis in breast and prostate cancers [250].

We then compared each publications' prognostic genes with the DEGs from our Pericol subtype and identified 62 genes in common (Figure 7.12). Four genes among these 62 genes (*AKAP12*, *CYP1B1*, *FAP* and *POSTN*) were seen to be shared between Pericol and at least 2 other publications. The *AKAP12* gene was shared between Pericol, Oncotype DX and Pagnotta et al. (2013) where it was found to be significantly associated with prognosis, however Pagnotta et al. did not include it in their final signature. The *FAP* gene was shared between Pericol, Oncotype DX and Oh et al. (2012). Both the *CYP1B1* and *POSTN* genes were shared between Pericol, Oh et al. (2012) and ColDx.

The total number of genes shared with each publication and the total number of genes from each publication are included with a list of the shared genes below:

- Oncotype DX [15/48]: *AKAP12, AKT3, ANTXR1, BGN, COL1A1, FAP, IGFBP3, IGFBP7, INHBA, PDGFC, SFRP4, SPARC, TGFB3, TIMP2* and *TIMP3*
- ColDx [25/584]: *ATP2C2, BICD1, CLCN2, CTGF, CTSD, CYP1B1, DGAT1, DHRS11, FRAT2, GLUL, GREM1, HIF1A, ICAM1, MACF1, PEA15, PHF21A, PIGR, POSTN, PPFIBP1, RHOQ, ROBO1, SERPINE1, SLA, SLC2A3* and *WSB1*

- Oh et al. (2012) [17/87]: *ADAM12, ANXA1, BCAT1, COL11A1, CXCR4, CYP1B1, FAP, GAS1, LOX, LRRC19, OLR1, POSTN, RGS2, SLC26A3, SPP1, TMEM45A* and *TNFAIP6*

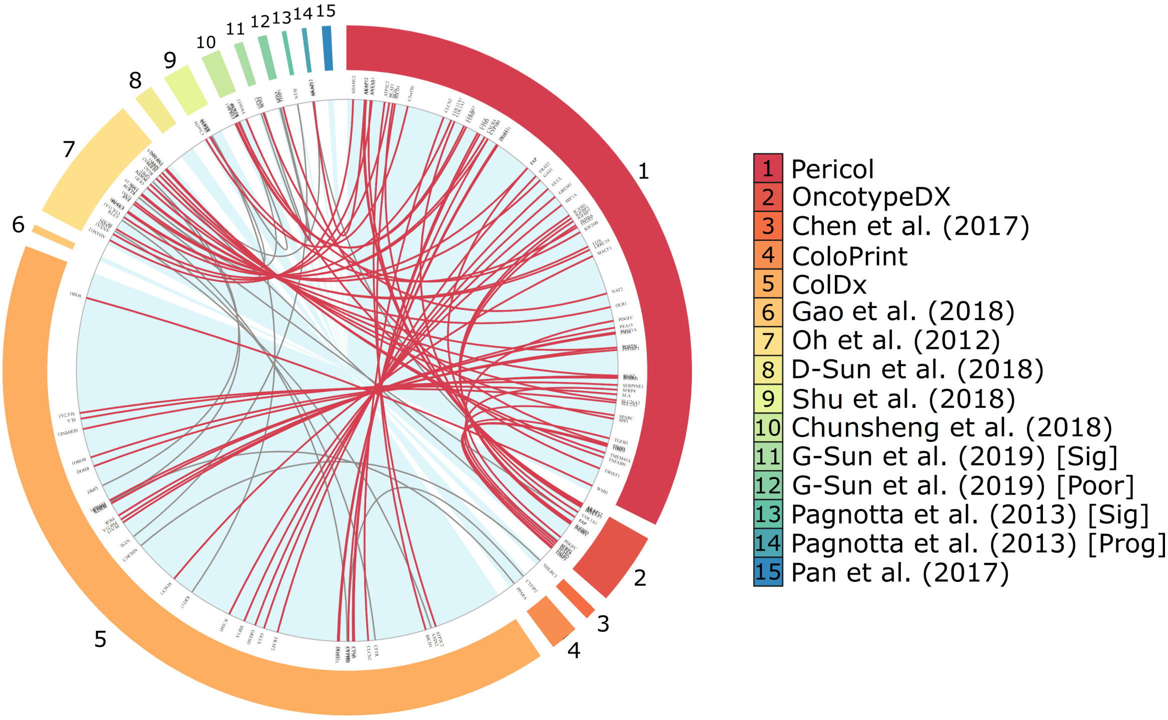
- Shu et al. (2018) [2/17]: *C5orf30* and *ITGA5*

- Chunsheng et al. (2018) [4/12]: *COL8A1, COMP, KIF26B* and *TWIST1*

- G-Sun et al. (2018) [2/8]: *NAT2* and *TIMP1*

- Pagnotta et al. (2013) [1/4]: *AKAP12*

Fig. 7.12 Circos plot highlighting the genes shared between any two signatures. Genes shared with Pericol are highlighted red.



7.6.5 Methylation of Genes in Pericol

In Sections 2.3.3 and 6.3.7 we explained the significant effects that epigenetic alterations can have on the development of cancers. To understand the role of these alterations in the Pericol subtype we performed a differential methylation analysis (using Limma [65] and methylGSA [251]) on the set of 471 genes previously identified to be differentially expressed.

Using the methylation data from TCGA-COAD a total of 1,692 CpG sites corresponding to regions within 380 genes were identified to be differentially methylated in the set of Pericol DEGs. By analysing these sites we identified 98 genes where the majority of CpG sites were hyper-methylated and the genes were under expressed and 87 genes where the majority of CpG sites were hypo-methylated and the genes were over expressed. The differential methylation results for all CpG sites corresponding to these 185 genes have been included in Appendix B.7.

We analysed these sets of hyper/hypo-methylated genes using version 6.8 of the publicly available DAVID Functional Annotation Tool [232, 233]. When assessing the GO biological processes [234], 11 GO terms were identified as being significantly over-represented (Benjamini and Hochberg adjusted p-values ≤ 0.05). The results of this gene set enrichment analysis can be found in Appendix B.8. Among the 87 hypo-methylated genes: 29 were associated with cell adhesion, 20 were associated with extracellular matrix organisation, 13 were associated with angiogenesis and 12 were associated with collagen catabolic processes.

7.7 Pericol, a Continuous Predictor of Recurrence?

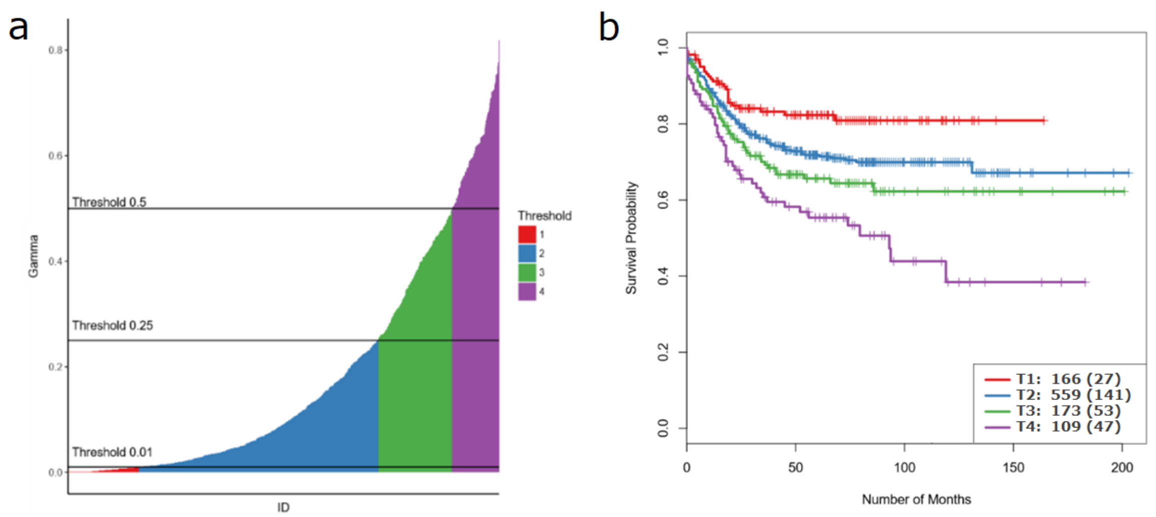
In this chapter we identified four CRC subtypes including Pericol, a subtype with significantly poorer disease-free survival rates than our other three subtypes. Overall only one third of samples could be assigned to a single primary subtype using our consensus model, making it difficult to use in the clinical decision making process for the other two thirds of patients. To overcome this limitation we set out to analyse the Pericol subtype in all samples, treating the Pericol γ value as a continuous variable between 0 and 1. Each sample's Pericol γ value was treated as the mean value across our four OAS LPD models used in our consensus model.

We initially grouped the Pericol γ values into four discrete groups for ease of viewing. These groups were:

- No Pericol ($\gamma < 1\%$).
- Low Pericol ($1\% \leq \gamma < 25\%$)
- Moderate Pericol ($25\% \leq \gamma < 50\%$)
- High Pericol ($\gamma \geq 50\%$)

By calculating and plotting KM-survival curves for these four categories of Pericol γ, using the GSE14333plus, GSE39582 and GSE41258 datasets, we found that Pericol γ showed a strong inverse correlation with disease-free survival time (Log-rank p-value = 5.03×10^{-6}) (Figure 7.13), indicating its potential use as a clinical predictor of risk.

Fig. 7.13 a) Mean Pericol γ taken from all four OAS LPD models for each sample in the GSE14333plus, GSE39582 and GSE41258 datasets, coloured according to the four discrete Pericol γ groups. b) Kaplan-Meier survival curves for the discretised Pericol γ groups.



To further assess the viability of using Pericol γ to predict patient risk we generated a Cox proportional hazard model. This was created using Pericol γ as a continuous variable alongside tumour location, TNM stage and patient age and gender as model covariates. We found that Pericol γ was an independent predictor of disease recurrence with a hazard ratio of 2.71 (p-value = 2.79×10^{-4}, 95% CI = 1.58-4.63).

7.8 Discussion

In this chapter we have presented our work on classifying colorectal cancer. We demonstrated how the LPD algorithm can be applied to a range of colorectal cancer transcriptome datasets to produce novel unsupervised classifications of the disease. By correlating the LPD processes we established four robust categories of colorectal cancer that were consistently identified in multiple independent datasets, each containing fresh frozen and formalin fixed tissue.

Analysis of our novel colorectal cancer classifications showed that each subtype held clinically relevant associations, despite LPD having no access to this information during the classification phase. This became especially interesting when examining each subtype in the wider context of the current colorectal cancer literature. In Chapter 6.3.7 we explained how the majority of sporadic colorectal cancers can be separated into six distinct groups based on their CIMP and MSI status (CMS). When comparing these groups to our subtypes we saw a striking similarity between LPD B and LPD C with CMS group 6 and group 1 respectively.

Within the CMS grouping system CMS group 6 is primarily described as MSI-L/MSS and CIMP-0, while in our own classifications LPD B was also shown to be MSS and CIMP negative. When analysing CMS group 6 in greater detail it is seen to consist of tumours containing both wild-type BRAF and KRAS genes that are located in the distal colon. The same characteristics are observed in LPD B, with the small exception of LPD B exhibiting neither primarily mutant or wild-type KRAS status.

CMS group 1 and LPB C each consist of primarily MSI-H and CIMP positive tumours. These subtypes are most commonly located in the proximal colon of elderly female patients. While these facts hold true in LPB C it must be noted that the over representation of female patients was not statistically significant (p-value = 0.156, Fisher's exact test). However, both of these colorectal cancer subtypes do exhibit a statistically significant over representation of mutated BRAF genes.

LPD's ability to derive these established subtypes in an unsupervised fashion further substantiates the other subtypes within the LPD classification of colorectal cancer. Pericol was the only LPD subtype with a significant association with KRAS mutated tumours. This subtype also exhibited significantly worse disease-free survival times, which may in part be linked to the abundance of KRAS mutations [252]. According to the work of Deng et al. (2015) [252] these patients with increased KRAS mutations could also benefit from adjuvant FOLFOX treatment to a greater degree than the patients with non-Pericol tumours, which do not exhibit an increase in KRAS mutations.

Furthermore, the overall poor prognosis of the Pericol group could be compounded by the effect of KRAS mutations within the tumours located in the distal colon of primary Pericol patients [253]. We observed a 0.721 Cox proportional hazard ratio within the Pericol tumours located in the proximal colon compared to distal based Pericol tumours, however this was not statistically significant (p-value = 0.332, 95% CI = 0.372-1.40).

When analysing the differentially expressed gene signatures of each LPD subtype we found a large intersect between the Pericol subtype and many existing published poor prognosis signatures. Before considering Pericol, we could only establish 19 unique genes that were in common between any two of these publications, whereas Pericol shared 62 unique genes with the same publications.

Most notable within this subset of genes were the *AKAP12*, *CYP1B1*, *FAP* and *POSTN* genes as they formed an intersect between three independent signatures. This was not something that was observable prior to the inclusion of Pericol and further emphasises the overlap with this novel subtype. Another notable intersection is that of Pericol and Oncotype DX, which is currently the only commercial test with level 1 evidence [19] (obtained from at least one properly designed randomized controlled trial). We found 15 genes in common between these signatures, which is even more startling given the unsupervised nature of LPD.

During our own analysis of the Pericol subtype we established its association with significantly poorer patient prognosis. By considering each patient's association (γ) with Pericol we found that Pericol γ can be used as an independent predictor of disease relapse. This novel measurement could be used alongside current standard clinical indicators (TNM stage and tumour location) to provide additional evidence during the clinical decision making progress. By identifying tumours that are more (or less) aggressive we hope to reduce the unnecessary side effects of avoidable treatment in patients with non-aggressive colorectal cancers.

Chapter 8

Conclusions and Future Work

8.1 Summary

In this thesis we have presented new classifications of prostate and colorectal cancers through the application of latent process decomposition. We have identified common subtypes between independent datasets, including aggressive forms of each disease that can be used to describe patient risk of disease recurrence. We now discuss potential avenues of further research that build upon the work presented in this thesis.

8.2 Prostate Cancer - DESNT

8.2.1 Biochemical Risk Assessment

Within the fields of prostate cancer diagnosis and prognosis prediction lies the problem of highly variable patient outcomes. At present the primary screening and diagnosis tool used to identify prostate cancer is the prostate specific antigen (PSA) test. When applied as a screening test it has been shown to reduce the cancer specific mortality by up to 21% [254], but also results in significant overdiagnosis [255]. The significant over diagnosis of indolent prostate cancer is a major issue and leads to significant over-treatment of low risk patients.

In an attempt to reduce the over-treatment of low risk patients we set-out to determine the feasibility of using the DESNT subtype of prostate cancer as a measure of biochemical recurrence (BCR) risk in prostate cancer patients. The DESNT subtype was previously shown to be associated with BCR and we believed that accurately assessing a patient's risk of BCR could be used to better inform the decision making between patients and doctors when determining treatment options. To begin this work we first reproduced the LPD classifications containing DESNT. This was achieved by following the processes described in Luca et al. (2017) [17] to normalise the gene expression microarray datasets and optimise the LPD initialisation parameters.

Upon obtaining the DESNT classification we began to assess the correlation between each samples' association with DESNT. We established a strong inverse correlation between increasing DESNT γ (the measure of confidence between a sample and an LPD process) and patient BCR-free survival. We also demonstrated that DESNT γ is an independent predictor of BCR. These findings could have a direct influence on the clinical discussions between doctors and patients regarding the choice of watchful waiting and radical treatment.

One of the longest running randomised trials to assess the benefits of radical prostatectomy vs watchful waiting was recently concluded by the Swedish Cancer Society [256]. During their 29 year study Bill-Axelson et al. (2018) established a mean gain of 2.9 additional years alive for patients with localised disease that underwent radical prostatectomy, compared to those that remained on watchful waiting over the course of 23 years. They also found that 8.4 patients were required to undergo radical treatment to prevent the death of one patient with localised disease. Overall Bill-Axelson et al. (2018) concluded that patients should be carefully selected for treatment and that low-risk tumours should not undergo radical treatment.

The ability of LPD to predict the risk of BCR through the analysis of DESNT is therefore of vital importance during the decision making process, as BCR precedes metastasis in 24% - 34% of patients [257] and cancer specific death in 53% of patients [258] within 15 years. While BCR is a useful measure of disease prognosis, using the time to metastasis would have been the ideal surrogate measure for cancer survival. Unfortunately this data was not available to us at this time, but should be considered in future studies. Ultimately it will remain the patient's decision whether or not to undergo radical treatment, however we foresee DESNT γ being a useful tool at the time of diagnosis to aid the decision making process alongside existing risk matrices to reduce the overall treatment of low risk patients.

8.2.2 OAS-LPD Classification of Biopsy Samples

To begin transferring our research into a clinical setting we obtained 20 prostate cancer biopsy samples covering a broad range of Gleason grades as part of a pilot study aiming to identify DESNT in prostate cancer biopsies. The main reason for this pilot study was to overcome the limitation of all previous DESNT research where analyses were conducted using prostatectomy samples. A second change to enable the analysis of DESNT in a clinical setting was the modification of the LPD algorithm by Rogers et al. (2005) [3] to allow samples to be classified into LPD processes from a pre-generated LPD model (OAS-LPD, Chapter 3.3.4).

We began to study the 20 biopsy samples by reference normalising them against the datasets used to generate the original DESNT classification by Luca et al. (2017) [17]. An OAS-LPD model was then constructed from the representative model processes based on the MSKCC dataset. The reference normalised biopsies were then run through this OAS-LPD model to produce the Bayesian classifications between biopsies and LPD processes.

We analysed the biopsy classification results to determine whether the DESNT subtype or any other subtypes had been identified in the biopsies. In this admittedly small cohort, we found that the biopsy γ values were comparatively lower than those of the original prostatectomy samples, however the DESNT subtype was identified as the primary subtype within four biopsy samples. All four of these samples were associated with high Gleason grades. This observation corresponds with our previous work on the DESNT subtype that identified an over representation of high Gleason grade cancers in the DESNT subtype.

Unfortunately Gleason grade was the only clinical variable currently available for these samples and this represents a major limitation of the pilot study. In future studies we intend to obtain comprehensive clinical follow-up data for patients, including BCR and metastasis status. The current sample size was another major limiting factor and while a range of Gleason grades were available, low Gleason grade samples were unrepresented in the pilot study. Both of these limitations will need to be addressed in future studies to accurately understand the prevalence and distribution of DESNT tumours within biopsies.

8.3 LPD and Consensus OAS-LPD Classification of Colorectal Cancer

For the heterogeneous disease called colorectal cancer (CRC) we aimed to establish a novel classification using latent process decomposition. To begin this project we gathered multiple gene expression microarray datasets alongside data from The Cancer Genome Atlas (TCGA). These datasets were normalised using RMA, ComBat and Quantile normalisation prior to being used in LPD.

We began the construction of CRC LPD models by refining the optimisation stage of the LPD initialisation parameter selection. We opted to use the MAP version of LPD to optimise both the number of processes and the value of σ simultaneously. This was achieved by repeatedly running every combination of these two variables and identifying the log-likelihood plateau before finally minimising the mean internal model process Pearson correlations. By performing these steps we were able to objectively define the optimal number of processes within each dataset.

Once an LPD model had been generated for each of our four gene expression microarray datasets we calculated the Pearson correlations between every pair-wise combination of processes. By analysing the Pearson correlations between each processes' expression profile it became clear that there were four common processes, with three processes strongly correlated between all four models and a fourth process strongly correlated between three models.

While identification of these common processes is a promising sign that LPD was not modelling dataset specific noise, one result that we cannot fully explain is the variable number of processes identified in each dataset, which ranged between four and six processes. The GSE14333plus dataset contained four common processes and a fifth dataset specific process. This fifth GSE14333plus process was weakly or moderately correlated with the poor prognosis process called Pericol in each of the other dataset models, suggesting an unknown underlying difference that separated these samples from Pericol.

The fifth GSE14333plus process was also moderately correlated with a dataset specific process within the GSE39582 based LPD model, however the GSE39582 specific process did not demonstrate a similarity to any other process. Similarly the dataset specific process within the GSE81653 did not correlate with any process from any model. These unique processes cannot be explained by dataset size or by microarray platform, but we speculate that in some cases these processes could be formed of samples that are in the early stages of multiple other processes.

Among the processes that correlated between datasets, namely LPD A, LPD B, LPD C and Pericol, we observed significantly poorer survival in patients primarily assigned to the Pericol subtype. Pericol γ was identified as an independent predictor of disease recurrence when combined with existing risk factors (TNM stage, tumour location, patient age and gender). This poor prognosis subtype was also the only subtype to be associated with an over representation of KRAS mutations, a mutation previously reported to reduce the survival of CRC patients [252]. However, due to these KRAS mutations patients with Pericol tumours may benefit from adjuvant FOLFOX treatment to a greater extent than those with wild-type KRAS [252]. The predictive power of LPD and Pericol could therefore be a useful tool to aid the clinical decision making process.

The differentially expressed genes present in the Pericol subtype were observed to overlap with a wide number of published signatures. While these signatures shared a total of 19 unique genes with each other, the intersect with Pericol's differentially expressed genes (DEGs) was more than three times greater (62 unique DEGs). This large overlap with published signatures combined with consistently identifying Pericol across multiple independent datasets demonstrates LPD's ability to identify robust subtypes. These findings encourage the development and use of related techniques to classify other heterogeneous diseases where current techniques have struggled to produce consistent results.

8.4 Consensus Molecular Subtypes

There have been several attempts at the unsupervised classification of colorectal cancer in recent years [259, 260, 220, 261–264, 223]. However, a recent study by Guinney et al. (2015) [20], using an aggregated dataset of 4,151 normalised samples, examined six of the previous models [259, 260, 220, 262–264] and established these models were dissimilar. One reason for the varying results could be attributed to the use of hierarchical clustering in six of the eight studies, which ignores the underlying heterogeneity.

Guinney et al. (2015) then attempted to produce a robust model using the original independent models as a starting point. Each of these models consisted of three to six subtypes, creating a collection of 27 unique subtype labels. These labels were treated as part of a consensus using a Markov cluster algorithm (MCL), applied on the network of associated Jaccard distances. By optimising the inflation factor within the MCL, Guinney et al. (2015) were able to establish four robust consensus subtypes.

Observing the identification of four robust subtypes in such a large study raises an interesting question regarding how similar our own four subtype consensus OAS-LPD model could be to the MCL consensus model. Among the 18 datasets used by Guinney et al. (2015) five were from proprietary sources and four of the remaining 13 public datasets were used in this thesis. A potential future study to access the similarity between these two independent models and techniques could provide further evidence that these models offer an accurate description of colorectal cancer. Observing such a result would also demonstrate the robustness of LPD in the classification of heterogeneous diseases. Undertaking this project would represent a large international collaboration involving many collaborators, with the potential to establish the most robust classification of CRC to date.

8.5 Development of Clinical Tests

The identification of DESNT and Pericol's prognostic power encourages the development of clinical tests. These tests could be used to identify the high risk patients that are in the early stages of each disease and help to reduce the over treatment of low risk patients currently self-electing for treatment.

Preventing irreversible surgery (prostatectomy or colectomy) in low risk patients requires the development of a clinical test using biopsy samples, or a less invasive material source such as a blood or urine samples. In the context of DESNT and Pericol, a less invasive test using these subtypes is entirely theoretical and would require extensive further study to develop. However, a study by Connell et al. in 2020 [265] combined urine-derived cell-free messenger RNA (cf-RNA) and urine cell DNA methylation data to produce a risk score capable of predicting whether a TRUS biopsy would contain Gleason score $\geq 3+4$ prostate tumours. While Connell et al. (2020) were able to identify prostate cancer positive patients, reducing the need for biopsies in up to 75% of patients, their test was unable to distinguish between the severity of the disease within each patient. The development of a non-invasive test capable of distinguishing between the severity of tumours would mark a major breakthrough in prostate and colorectal cancer research.

Further Prostate cancer studies are also required to validate the identification of DESNT tumours using biopsy samples. In the case of colorectal cancer we have already demonstrated Pericol γ's ability to independently predict disease recurrence using biopsy samples. Due to this success we will focus our discussion on the development of a biopsy based test for colorectal cancer.

One of the largest points of contention when developing a test is the choice of technology. In our exploratory work with colorectal cancer we employed gene microarrays and exon microarrays, as microarray studies are abundant and fulfil LPD's assumption that the gene expression follows a normal/log-normal distribution. We also employed RNA-seq and showed the expression profiles of the LPD processes still correlated with those generated from the microarrays, however there was a greater level of variability between RNA-seq LPD runs. A third potential technique to process the biopsy samples is through quantitative reverse transcription polymerase chain reaction (RT-qPCR), which has been shown to produce results comparable to microarrays [266, 267].

To translate this work into a clinical test we must first consider a number of factors, namely the cost, the expected turnaround time and the number of genes to analyse. These decisions are important to determine the appropriate technology to measure each sample's gene expression levels. These decisions are also connected, with the cost of each available technology varying in relation to the number of genes.

Among the clinical tests currently available for colorectal cancer is the Onocotype DX Colon Recurrence Score Test. This test assesses the gene expression levels of 12 genes to establish the likelihood of recurrence within three years of surgery [19]. RT-PCR is a fast and cost effective method for small sets of genes, making it an appropriate choice for the Oncotype DX test [268, 269].

In contrast to the small number of genes used by Oncotype DX, our LPD classifications were derived from 500 unique genes obtained from microarrays. It would be necessary to measure the expression levels of all 500 of these genes to use our CRC consensus OAS-LPD model. The relatively large number of genes is likely to rule out the use of RT-PCR, instead custom arrays or RNA-seq could be more appropriate options to use alongside OAS-LPD. Within a hospital setting the application of whole transcriptome based tests may require additional infrastructure to facilitate their use. An alternative solution would be to use 3rd party laboratories that have been accredited by the relevant governing bodies, such as UKAS within the UK [270]. While the current DESNT and Pericol results are promising, only large scale clinical trials will offer a definitive assessment of the predictive power and potential benefits of using OAS-LPD with these subtypes in a clinical setting.

8.6 Improved Versions of LPD

In this thesis we used a version of LPD developed by Rogers et al. [3], to reproduce the DESNT subtype of prostate cancer and to produce a novel classification of colorectal cancer. One of the reasons for using this particular version of LPD was that our research lab had previously used it to produce the DESNT classification [17] and for their classification of breast cancer data [18]. LPD was also selected

Several other versions of LPD have also been developed. One of these proposed models, created by Ying et al. [271], uses an improved framework for parameter estimation. This new model uses the marginalised variational Bayes (MVB) framework instead of the standard variational Bayes (VB) method used by Rogers et al. [3], which has been shown to produce mathematically better solutions [271].

The LPD model has been further improved by Masada et al. [272], using a new parameter estimation framework, known as MVB+. This version of LPD allows for the model hyper-parameters (such as σ) to be re-estimated during model training. This removes the need to perform additional optimisation steps when choosing the LPD parameters (described in Section 5.3.1), as sigma is now automatically optimised.

In addition to producing models that better fit the data, the authors of these new versions of LPD claim that they are also capable of generating models in significantly less time. In our analyses, using the version of LPD by Rogers et al. (2005) [3], we required approximately 24 hours to fit an LPD model on 320 samples using 500 genes. In comparison to this Masada et al. (2009) [272] required only 174 minutes (on average) to fit a model on 286 samples using 17,816 genes. This impressive improvement to performance is further emphasised by their use of far more genes, providing a greater level of detail by removing the need to reduce the set of input genes.

A future investigation of interest would be to use these new versions of LPD with our existing prostate and colorectal cancer datasets. The models generated by these techniques could then be compared to our current prostate and colorectal cancer LPD models. Empirical analysis of these new models may further support the mathematical improvement claims made by Ying et al. (2007) [271] and Masada et al. (2009) [272]. If the mathematical claims held true we could expect to see improvements to the models, which may result in greater confidence between samples and LPD processes and fewer samples changing their primary subtype between LPD runs.

Employing the new algorithms' greatly reduced computational times would allow us to generate LPD models for many more datasets in a far more manageable time-frame. By comparing these models to our existing classifications we could further demonstrate the widespread nature of the LPD processes, with few processes that are dataset specific artefacts.

8.7 Conclusion

In this thesis we applied LPD to two heterogeneous diseases known as prostate and colorectal cancers. We identified the potential use of the DESNT subtype's γ value as an indicator of patient biochemical relapse risk in prostate cancer [5]. If this test could be validated in a clinical setting our findings could significantly reduce the over treatment of low risk patients and help to identify high risk patients that were previously viewed as low risk.

We have also produced a new classification framework for colorectal cancer using a consensus OAS-LPD approach. Among our four CRC subtypes we observed a significantly poorer prognosis in patients that displayed a predominantly Pericol expression signature. Pericol was shown to be an independent predictor of disease recurrence alongside existing risk measures and could be used to further inform the choice of treatments through its association with KRAS mutations. We also established sets of known clinical covariates within two of our other CRC subtypes, providing further evidence that LPD is able to extract the heterogeneous structure of diseases in an unsupervised manor.

These findings emphasise the importance of using more advanced techniques, such as LPD, when analysing heterogeneous diseases. By using LPD instead of other algorithms, such as Naive-Bayes, dendrograms or Gaussian mixture models, we can reduce the risk of overfitting, unambiguously determine the number of processes and remove the need to preselect genes prior to feature extraction [3].

While beyond the scope of this thesis, studying the molecular changes that drive the development of each cancer subtype identified in this thesis may reveal new therapeutic targets. Such discoveries could in turn enable new radical treatments to be developed, or lead to the personalisation of treatment pathways. These outcomes would represent a significant step forward in the classification and treatment of heterogeneous diseases.

References

[1] S Caul and J Broggio. Cancer registration statistics, England: 2017. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/cancerregistrationstatisticsengland/2017, 2019. Accessed: June 2019.

[2] Susan. C Lester. *Manual of Surgical Pathology*. Saunders, 3rd edition, 2010.

[3] S. Rogers, M. Girolami, C. Campbell, and R. Breitling. The latent process decomposition of cDNA microarray data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2(2):143–156, 2005.

[4] AcademLib. Prostate Cancer. http://academlib.com/5561/health/prostate_cancer, 2014. Accessed: May 2017.

[5] Bogdan-Alexandru Luca, Vincent Moulton, Christopher Ellis, Dylan R Edwards, Colin Campbell, Rosalin A Cooper, Jeremy Clark, Daniel S Brewer, and Colin S Cooper. A novel stratification framework for predicting outcome in patients with prostate cancer. *British Journal of Cancer*, 2020. https://doi.org/10.1038/s41416-020-0799-5.

[6] Liza Torborg. Mayo CLinic Q and A: Preventing colon cancer with screening, early detection. https://newsnetwork.mayoclinic.org/discussion/mayo-clinic-q-and-a-preventing-colon-cancer-with-screening-early-detection/, 2019. Accessed: March 2020.

[7] Łukasz Szylberg, Marlena Janiczek, Aneta Popiel, and Andrzej Marszałek. Serrated Polyps and Their Alternative Pathway to the Colorectal Cancer: A Systematic Review. *Gastroenterology Research and Practice*, 2015:573814, 2015.

[8] J. I Epstein, M. J Zelefsky, D. D Sjoberg, J. B Nelson, L Egevad, Magi-Galluzzi C, Vickers A. J, A. V Parwani, V. E Reuter, S.W Fine, J. A Eastham, P Wiklund, M Han, C. A Reddy, J. P Ciezki, T Nyberg, and Klein E.A. A Contemporary Prostate Cancer Grading System: A Validated Alternative to the Gleason Score. *European Urology*, 69(3):428–435, 2016.

[9] Shuji Ogino and Ajay Goel. Molecular Classification and Correlates in Colorectal Cancer. *Journal of Molecular Diagnostics*, 10(1):13–27, 2008.

[10] National Cancer Institute. cancer. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/cancer, 2020. Accessed: July 2020.

[11] Global Cancer Observatory. Cancer fact sheets: All cancers. http://gco.iarc.fr/today/data/factsheets/cancers/39-All-cancers-fact-sheet.pdf, 2019. Accessed: June 2019.

[12] P Krzyszczyk, A Acevedo, EJ Davidoff, LM Timmins, I Marrero-Berrios, M Patel, C White, C Lowe, JJ Sherba, C Hartmanshenn, KM O'Neill, ML Balter, ZR Fritz, IP Androulakis, RS Schloss, and ML Yarmush. The growing role of precision and personalized medicine for cancer treatment. *Technology (Singapore World Science)*, 6(3):79–100, 2018.

[13] Amy V Kapp, Stefnie S Jeffrey, Anita Langerød, Anne-Lise Børresen-Dale, Wonshik Han, Dong-Young Noh, Ida R K Bukholm, et al. Discovery and validation of breast cancer subtypes. *BMC Genomics*, 7(1):231, 2006.

[14] F Bertucci, P Finetti, and D Birnbaum. Basal Breast Cancer: A Complex and Deadly Molecular Subtype. *Current Molecular Medicine*, 12(1):96–110, 2012.

[15] William D Foulkes, Ingunn M Stefansson, Pierre O Chappuis, Louis R Bégin, John R Goffin, Nora Wong, Michel Trudel, and Lars A Akslen. Germline BRCA1 mutations and a basal epithelial phenotype in breast cancer. *Journal of the National Cancer Institute*, 95(19):1482–1485, 2003.

[16] Prasanna Alluri and Lisa Newman. Basal-like and Triple Negative Breast Cancers: Searching For Positives Among Many Negatives. *Surgical Oncology Clinics of North America*, 23(3):567–577, 2015.

[17] Bogdan-Alexandru Luca, Daniel S Brewer, Dylan R Edwards, Sandra Edwards, Hayley C Whitaker, Sue Merson, Nening Dennis, Rosalin A Cooper, Steven Hazell, Anne Y Warren, Rosalind Eeles, Andy G Lynch, Helen Ross-Adams, Alastair D Lamb, David E Neal, Krishna Sethia, Robert D Mills, Richard Y. Ball, Helen Curley, Jeremy Clark, Vincent Moulton, and Colin S. Cooper. DESNT: A Poor Prognosis Category of Human Prostate Cancer. *European Urology Focus*, 2017.

[18] Luke Carrivick, Simon Rogers, Jeremy Clark, Colin Campbell, Mark Girolami, and Colin Cooper. Identification of prognostic signatures in breast cancer microarray data using Bayesian techniques. *Journal of the Royal Society: Interface*, 3(8):367–381, 2006.

[19] OncotypeIQ. Oncotype DX Colon Recurrence Score. https://www.oncotypeiq.com/en-US/colon-cancer/healthcare-professionals/oncotype-dx-colon-recurrence-score/about-the-test, 2020. Accessed: May 2020.

[20] Justin Guinney, Rodrigo Dienstmann, Xin Wang, Aurélien De Reyniès, Andreas Schlicker, Charlotte Soneson, Laetitia Marisa, et al. The consensus molecular subtypes of colorectal cancer. *Nature Medicine*, pages 1350–1356, 2015.

[21] Harvey Lodish, Arnold Berk, Lawrence Zipursky, Paul Matsudaira, David Baltimore, and James Darnell. *Molecular Cell Biology*. W. H. Freeman, 4 edition, 1999.

[22] Tobias Maier, Marc Güell, and Luis Serrano. Correlation of mrna and protein in complex biological samples. *FEBS Letters*, 583(24):3966–3973, 2009.

[23] Lucy W Barrett, Sue Fletcher, and Steve D Wilton. Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cellular and Molecular Life Sciences*, 69(21):3613–3634, 2012.

[24] Douglas Hanahan and Robert A Weinberg. The Hallmarks of Cancer. *Cell*, 100(1):57–70, 2000.

[25] T Dunn. Oxygen and cancer. *North Carolina medical journal*, 58(2):140–143, 1997.

[26] Michael Papetii and Ira M Herman. Mechanisms of normal and tumour-derived angiogenesis. *The American Journal of Physiology: Cell Physiology*, 282:947–970, 2002.

[27] Douglas Hanahan and Robert A Weinberg. Hallmarks of Cancer: The Next Generation. *Cell*, 144(5):646–674, 2011.

[28] HarperCollins Publishers. Definition of gloycolysis. https://www.collinsdictionary.com/dictionary/english/glycolysis, 2020. Accessed: June 2020.

[29] Bahar Yetkin-Arik, Ilse M. C. Vogels, Patrycja Nowak-Sliwinska, Andrea Weiss, Riekelt H. Houtkooper, Cornelis J. F. Van Noorden, Ingeborg Klaassen, and Reinier O. Schlingemann. The role of glycolysis and mitochondrial respiration in the formation and functioning of endothelial tip cells during angiogenesis. *Scientific Reports*, 9(1):12608, 2019.

[30] Kelly M Kennedy and Mark W Dewhirst. Tumor Metabolism of Lactate: The Influence and Therapeutic Potential for MCT and CD147 Regulation. *Future Oncology*, 6(1):127–148, 2010.

[31] Matthew G Vander Heiden, Lewis C Cantley, and Craig B Thompson. Understanding the Warburg Effect: The Metabolic Requirements of Cell Proliferation. *Science*, 324(5930):1029–1033, 2009.

[32] Caroline Jochems and Jeffrey Schlom. Tumor-infiltrating immune cells and prognosis: the potential link between conventional cancer therapy and immunity. *Experimental biology and medicine (Maywood)*, 236(5):567–579, 2011.

[33] F Pagés, J Galon, M C Nieu-Nosjean, E Tartour, C Sautés-Fridman, and W H Fridman. Immune Infiltration in Human Tumors: A Prognostic Factor That Should Not Be Ignored. *Oncogene*, 29(8):1093–1020, 2009.

[34] Claire M Vajdic and Marina T van Leeuwen. Cancer Incidence and Risk Factors After Solid Organ Transplantation. *International Journal of Cancer*, 125(8):1747–1754, 2009.

[35] Magali Olivier, Monica Hollstein, and Pierre Hainaut. TP53 Mutations in Human Cancers: Origins, Consequences, and Clinical Use. *Cold Spring Harbor Perspectives in Biology*, 2(1):a001008, 2010.

[36] R Fodde. The APC gene in colorectal cancer. *European Journal of Cancer*, 38(7):867–871, 2002.

[37] M Oh, N Alkhushaym, S Fallatah, A Althagafi, R Alsowaida, J Jeter, J. R Martain, H. M Babiker, A McBride, and I Abraham. The association of BRCA1 and BRCA2 mutations with prostate cancer risk, frequency, and mortality: A meta-analysis. *The Prostate*, 79(8):880–895, 2019.

[38] Rohini Roy, Jarin Chun, and N. Simon Powell. BRCA1 and BRCA2: different roles in a common pathway of genome protection. *Nature Reviews Cancer*, 12(1):68–78, 2012.

[39] Stefan Fröhling and Hartmut Döhner. Chromosomal Abnormalities in Cancer. *New England Journal of Medicine*, 359(7):722–734, 2008.

[40] Sylvan C. Baca, Davide Prandi, Michael S. Lawrence, Juan Miguel Mosquera, Alessandro Romanel, Yotam Drier, Kyung Park, Naoki Kitabayashi, Theresa Y. MacDonald, Mahmoud Ghandi, Eliezer Van Allen, Gregory V. Kryukov, Andrea Sboner, Jean-Philippe Theurillat, T. David Soong, Elizabeth Nickerson, Daniel Auclair, Ashutosh Tewari, Himisha Beltran, Robert C. Onofrio, Gunther Boysen, Candace Guiducci, Christopher E. Barbieri, Kristian Cibulskis, Andrey Sivachenko, Scott L. Carter, Gordon Saksena, Douglas Voet, Alex H Ramos, Wendy Winckler, Michelle Cipicchio, Kristin Ardlie, Philip W. Kantoff, Michael F. Berger, Stacey B. Gabriel, Todd R. Golub, Matthew Meyerson, Eric S. Lander, Olivier Elemento, Gad Getz, Francesca Demichelis, Mark A. Rubin, , and Levi A. Garraway. Punctuated Evolution of Prostate Cancer Genomes. *Cell*, 153(3):666–677, 2013.

[41] MF Berger, MS Lawrence, F Demichelis, Y Drier, K Cibulskis, AY Sivachenko, A Sboner, R Esgueva, D Pflueger, C Sougnez, R Onofrio, SL Carter, K Park, L Habegger, L Ambrogio, T Fennell, M Parkin, G Saksena, D Voet, AH Ramos, TJ Pugh, J Wilkinson, S Fisher, W Winckler, S Mahan, K Ardlie, J Baldwin, JW Simons, N Kitabayashi, TY MacDonald, PW Kantoff, L Chin, SB Gabriel, MB Gerstein, TR Golub, M Meyerson, A Tewari, ES Lander, G Getz, MA Rubin, and LA Garraway. The genomic complexity of primary human prostate cancer. *Nature*, 470(7333):214–220, 2011.

[42] Mireille Crampe, Karl Haslam, Emma Groarke, Eileen Kelleher, Derville O'Shea, Eibhlin Conneally, and Stephen E. Langabeer. Chronic Myeloid Leukemia with an e6a2 BCR-ABL1 Fusion Transcript: Cooperating Mutations at Blast Crisis and Molecular Monitoring. *Case Reports in Hematology*, 2017:1–5, 2017. Art ID. 9071702.

[43] Maria-Magdalena Georgescu. PTEN Tumor Suppressor Network in PI3K-Akt Pathway Control. *Genes and Cancer*, 1(12):1170–1177, 2010.

[44] Michele Milella, Italia Falcone, Fabiana Conciatori, Ursula Cesta Incani, Anais Del Curatolo, Nicola Inzerilli, Carmen M. A. Nuzzo, Vanja Vaccaro, Sabrina Vari, Francesco Cognetti, and Ludovica Ciuffreda. PTEN: Multiple Functions in Human Malignant Tumors. *Frontiers in Oncology*, 5:24, 2015.

[45] Matthew Aguirre, Manuel A Rivas, and James Priest. Phenome-wide Burden of Copy-Number Variation in the UK Biobank. *The American Journal of Human Genetics*, 105(2):373–383, 2019.

[46] Suzanne Clancy. Copy Number Variation. *Nature Education*, 1(1):95, 2008.

[47] Linfan Zhang, Nikta Feizi, Chen Chi, and Pingzhao Hu. Association Analysis of Somatic Copy Number Alteration Burden With Breast Cancer Survival. *Frontiers in Genetics*, 9:421, 2018.

[48] Theresa Phillips. The role of methylation in gene expression. *Nature Education*, 1(1):116, 2008.

[49] M. Gardiner-Garden and M. Frommer. CpG Islands in Vertebrate Genomes. *Journal of Molecular Biology*, 196(2):261–282, 1987.

[50] Jörg Tost. DNA Methylation: An Introduction to the Biology and the Disease-Associated Changes of a Promising Biomarker. *Molecular Biotechnology*, 44(1):71–81, 2010.

[51] Niaz Mahmood and Shafaat A Rabbani. DNA Methylation Readers and Cancer: Mechanistic and Therapeutic Applications. *Frontiers in Oncology*, 9:489, 2019.

[52] Roberta Rosa, Francesca Monteleone, Nicola Zambrano, and Roberto Bianco. In Vitro and In Vivo Models for Analysis of Resistance to Anticancer Molecular Therapies. *Current Medical Chemistry*, 21(14):1595–1606, 2014.

[53] ThermoFisher. RNAlater-ICE Frozen Tissue Transition Solution. https://www.thermofisher.com/uk/en/home/references/ambion-tech-support/rna-buffers-chemicals/tech-notes/thaw-frozen-tissues–without-damaging-rna.html, 2020. Accessed: February 2020.

[54] Florenza Lüder Ripoli, Annika Mohr, Susanne Conradine Hammer, Saskia Willenbrock, Marion Hewicker-Trautwein, Silvia Hennecke, Hugo Murua Escobar, , and Ingo Nolte. A Comparison of Fresh Frozen vs. Formalin-Fixed, Paraffin-Embedded Specimens of Canine Mammary Tumors via Branched-DNA Assay. *International Journal of Molecular Science*, 17(5), 2016. Art ID. 724.

[55] Pan Zhang, Brian D. Lehmann, Yu Shyr, and Yan Guo. The Utilization of Formalin Fixed-Paraffin-Embedded Specimens in High Throughput Genomic Studies. *International Journal of Genomics*, 2017:1–9, 2017. Art ID. 1926304.

[56] EF Gaffney, PH Riegman, WE Grizzle, and PH Watson. Factors that drive the increasing use of FFPE tissue in basic and translational cancer research. *Biotechnic & Histochemistry*, 93(5):373–386, 2018.

[57] Sigma-Aldrich. Primary Cell Culture Basics. https://www.sigmaaldrich.com/technical-documents/articles/biology/primary-cell-culture.html, 2020. Accessed: February 2020.

[58] Jennifer L. Wilding and Walter F. Bodmer. Cancer Cell Lines for Drug Discovery and Development. *American Association for Cancer Research: Cancer Research*, 74(9):2377–2384, 2014.

[59] Han-Ming Liu, Dan Yang, Zhao-Fa Liu, Sheng-Zhou Hu, Shen-Hai Yan, and Xian-Wen He. Density distribution of gene expression profiles and evaluation of using maximal information coefficient to identify differentially expressed genes. *PLOS ONE*, 14:1–28, 2019.

[60] M Grunstein and D S Hogness. Colony hybridization: a method for the isolation of cloned DNAs that contain a specific gene. *Proceedings of the National Academy of Sciences of the United States of America*, 72(10):3961–3965, 1975.

[61] Mark Schena, Dari Shalon, Ronald W Davis, and Patrick O Brown. Quantitive Monitorng of Gene Expression Patterns with a Complementary DNA Microarry. *Science*, 270(5235):467–470, 1995.

[62] Roger Bumgarner. Overview of dna microarrays: Types, applications, and their future. *Current Protocols in Molecular Biology*, 101(1):22.1.1–22.1.11, 2013.

[63] National Human Genome Research Institute. DNA Microarray Technology. https://www.genome.gov/10000533/dna-microarray-technology/, 2015. Accessed: May 2017.

[64] Jahangheer S. Shaik and Mohammed Yeasin. A unified framework for finding differentially expressed genes from microarray experiments. *BMC Bioinformatics*, 8(1):347, 2007.

[65] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015.

[66] Victor Trevino, Francesco Falciani, and Hugo A Barrera-Salda na. DNA Microarrays: a Powerful Genomic Tool for Biomedical and Clinical Research. *Molecular Medicine*, 13(9-10):527–541, 2007.

[67] ThermoFisher. GeneChip™ Human Exon 1.0 ST Array. https://www.thermofisher.com/order/catalog/product/900651, 2017. Accessed: May 2017.

[68] Ye Deng, Zhili He, Joy D Van Nostrand, and Jizhong Zhou. Design and analysis of mismatch probes for long oligonucleotide microarrays. *BioMed Central Genomics*, 9(491), 2008.

[69] Stephen R Piccolo, Ying Sun, Joshua D Campbell, Marc E Lenburg, Andrea H Bild, and W Evan Johnson. A single-sample microarray normalization method to facilitate personalized-medicine workflows. *Genomics*, 100(6):337–344, 2012.

[70] Christian Müller, Arne Schillert, Caroline Röthemeier, David-Alexandre Trégouët, Carole Proust, Harald Binder, Norbert Pfeiffer, Manfred Beutel, Karl J. Lackner, Renate B. Schnabel, Laurence Tiret, Philipp S. Wild, Stefan Blankenberg, Tanja Zeller, and Andreas Ziegler. Removing Batch Effects from Longitudinal Gene Expression - Quantile Normalization Plus ComBat as Best Approach for Microarray Transcriptome Data. *PLOS ONE*, 11(6):1–23, 2016.

[71] Shoba Ranganathan, Kenta Nakai, Christian Schönbach, and Michael Gribskov. *Encyclopedia of Bioinformatics and Computational Biology*, pages 467–468. Elsevier, 2018.

[72] B. M Bolstrad, R. A Irizarry, M Åstrand, and T. P Speed. Analysis of Data From Vival DNA Microchips. *Journal of the American Statistical Association*, 19(2):185–193, 2003.

[73] Amaratunga Dhammika and Cabrera Javier. Analysis of Data From Vival DNA Microchips. *Journal of the American Statistical Association*, 96(456):1161–1170, 2001.

[74] R. A Irizarry, B Hobbs, F Collin, Y. D Beazer-Barclay, K. J Antonellis, U Scherf, and T. P Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 64(2):249–264, 2003.

[75] Cheng Li and Wing Hung Wong. Model-basedanalysisofoligonucleotidearrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences*, 98(1):31–36, 2001.

[76] John W Tukey. *Exploratory data analysis*. Addison-Wesley Series in Behavioral Science. Addison-Wesley, Reading, MA, 1977.

[77] W. Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[78] Emmanuel Gbenga Dada, Joseph Stephen Bassi, Haruna Chiroma, Shafi'i Muhammad Abdulhamid, Adebayo Olusola Adetunmbi, and Opeyemi Emmanuel Ajibuwa. Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon*, 5(6):e01802, 2019.

[79] J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare. Credit card fraud detection using machine learning techniques: A comparative analysis. In *2017 International Conference on Computing Networking and Informatics (ICCNI)*, pages 1–9, 2017.

[80] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, and Dimitrios I. Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17, 2015.

[81] Mohamed Alloghani, Dhiya Al-Jumeily, Jamila Mustafina, Abir Hussain, and Ahmed Aljaaf. *A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science*, pages 3–21. Springer International Publishing, 2020.

[82] Olcay Akman, Timothy Comar, Daniel Hrozencik, and Josselyn Gonzales. Chapter 11 - data clustering and self-organizing maps in biology. In *Algebraic and Combinatorial Computational Biology*, MSE/Mathematics in Science and Engineering, pages 351–374. Academic Press, 2019.

[83] E. Forgy. Cluster analysis of multivariate data: efficiency vs. interpretability of classification. *Biometrics*, 21:768–780, 1965.

[84] S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[85] J. B. MacQueen. Some Methods for classification and Analysis of Multivariate Observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[86] D. Raja Kishor and N. B. Venkateswarlu. Hybridization of Expectation-Maximization and K-Means Algorithms for Better Clustering Performance. *Cybernetics and Information Technologies*, 16(2):16–34, 2016.

[87] Chuong B Do and Serafim Batzoglou. What is the expectation maximization algorithm? *Nature Biotechnology*, 26:897–899, 2008.

[88] Pasi Fränti and Sami Sieranoja. How much can k-means be improved by using better initialization and repeats? *Pattern Recognition*, 93:95–112, 2019.

[89] Liu Lin, Tang Lin, Dong Wen, Yao Shaowen, and Zhou Wei. An overview of topic modeling and its current applications in bioinformatics. *Springerplus*, 5(1):1608, 2016.

[90] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[91] Thomas Hofmann. Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, 42:177–196, 2001.

[92] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

[93] E Mazaris and A Tsiotras. Molecular Pathways in Prostate Cancer. *Nephro-Urology Monthly*, 5(3):792–800, 2013.

[94] Colin S Cooper, Rosalind Eeles, David C Wedge, Peter Van Loo, Gunes Gundem, Ludmil B Alexandrov, Barbara Kremeyer, et al. Analysis of the Genetic Phylogeny of Multifocal Prostate Cancer Identifies Multiple Independent Clonal Expansions in Neoplastic and Morphologically Normal Prostate Tissue. *Nature Genetics*, 47(4):367–372, 2015.

[95] K.G.M. Moons, A. Rogier T. Donders, E.W. Steyerberg, and F.E. Harrell. Penalized maximum likelihood estimation to directly adjust diagnostic and prognostic prediction models for overoptimism: a clinical example. *Journal of Clinical Epidemiology*, 57(12):1262 – 1270, 2004.

[96] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37:183–233, 1993.

[97] David G Kleinbaum and Mitchel Klein. *Survival Analysis: A Self-Learning Text*. Springer-Verlag New York, 2005.

[98] D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[99] T M Therneau and P M Grambsch. *The Cox Model*, pages 39–77. Springer, 2000.

[100] National Human Genome Research Institute. Biological Pathways Fact Sheet. https://www.genome.gov/about-genomics/fact-sheets/Biological-Pathways-Fact-Sheet, 2015. Accessed: March 2020.

[101] The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, 47(D1):D330–D338, 2018.

[102] Bijay Jassal, Lisa Matthews, Guilherme Viteri, Chuqiao Gong, Pascual Lorente, Antonio Fabregat, Konstantinos Sidiropoulos, Justin Cook, Marc Gillespie, Robin Haw, Fred Loney, Bruce May, Marija Milacic, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Solomon Shorser, Thawfeek Varusai, Joel Weiser, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The reactome pathway knowledgebase. *Nucleic Acids Research*, 48(D1):D498–D503, 2019.

[103] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353–D361, 2016.

[104] Isabelle Rivals, Léon Personnaz, Lieng Taing, and Marie-Claude Potier. Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, 23(4):401–407, 2006.

[105] William K Oh, Mark Hurwitz, Anthony V D'Amico, Jerome P Richie, and Philip W Kantoff. *Cancer Medicine*. BC Decker, 6 edition, 2003.

[106] Joaquin J Garcia, Hikmat A Al-Ahmadie, Anuradha Gopalan, Satish K Tickoo, Peter T Scardino, Victor E Reuter, and Samson W Fine. Do Prostatic Transition Zone Tumours Have A Distinct Morphology? *The American Journal of Surgical Pathology*, 32(11):1709–1714, 2008.

[107] Muhammad Naeem Bashir. Epidemiology of Prostate Cancer. *The Asian Pacific Journal of Cancer Prevention*, 16(13):5137–5141, 2015.

[108] J Cuzick, MA Thorat, G Andriole, OW Brawley, PH Brown, Z Culig, RA Eeles, LG Ford, FC Hamdy, L Holmberg, D Ilic, TJ Key, C La Vecchia, H Lilja, M Marberger, FL Meyskens, LM Minasian, C Parker, HL Parnes, S Perner, H Rittenhouse, J Schalken, HP Schmid, BJ Schmitz-Dräger, FH Schröder, A Stenzl, B Tombal, TJ Wilt, and A Wolk. Prevention and early detection of prostate cancer. *The Lancet Oncology*, 15(11):484–492, 2014.

[109] Michael F Leitzmann, , and Sabine Rohrmann. Risk factors for the onset of prostatic cancer: age, location, and behavioral correlates. *Journal of Clinical Epi*, 4(1):1–11, 2012.

[110] Henrik Gröberg. Prostate cancer epidemiology. *The Lancet*, 361(9360):859–864, 2003.

[111] Alexandre R. Zlotta, Shin Egawa, Dmitry Pushkar, Alexander Govorov, Takahiro Kimura, Masahito Kido, Hiroyuki Takahashi, Cynthia Kuk, Marta Kovylina, Najla Aldaoud, Neil Fleshner, Antonio Finelli, Laurence Klotz, Jenna Sykes, Gina Lockwood, and Theodorus H. van der Kwast. Prevalence of Prostate Cancer on Autopsy: Cross-Sectional Study on Unscreened Caucasian and Asian Men. *Journal of the Nation Cancer Institute*, 104(14):1050–1058, 2013.

[112] Yoav Ben-Shlomo, Simon Evans, Fowzia Ibrahim, Biral Patel, Ken Anson, Frank Chinegwundoh, Cathy Corbishley, Danny Dorling, Bethan Thomas, David Gillatt, Roger Kirby, Gordon Muir, Vinod Nargund, Rick Popert, Chris Metcalfe, and Raj Persad. The Risk of Prostate Cancer amongst Black Men in the United Kingdom: The PROCESS Cohort Study. *European Urology*, 153(1):99–105, 2008.

[113] Aurora Perez-Cornago, Timothy J Key, Naomi E Allen, Georgina K Fensom, Kathryn E Bradbury, Richard M Martin, and Ruth C Travis. Prospective investigation of risk factors for prostate cancer in the UK Biobank cohort study. *British Journal of Cancer*, 117(10):1562–1571, 2017.

[114] Ola Bratt, Linda Drevin, Olof Akre, Hans Garmo, and Pär Stattin. Family History and Probability of Prostate Cancer, Differentiated by Risk Category: A Nationwide Population-Based Study. *JNCI: Journal of the National Cancer Institute*, 108(10), 2016. djw110.

[115] James Ted McDonald, Michael Farnworth, and Zikuan Liu. Cancer and the healthy immigrant effect: a statistical analysis of cancer diagnosis using a linked Census-cancer registry administrative database. *BMC Public Health*, 17(1):a296, 2017.

[116] Jennifer H. Cohen, Alan R. Kristal, and Janet L. Stanford. Fruit and Vegetable Intakes and Prostate Cancer Risk. *JNCI: Journal of the National Cancer Institute*, 92(1):61–68, 2000.

[117] W H Lee, R A Morton, J I Epstein, J D Brooks, P A Campbell, G S Bova, W S Hsieh, W B Isaacs, and W G Nelson. Cytidine methylation of regulatory sequences near the π-class glutathione S-transferase gene accompanies human prostatic carcinogenesis. *Proceedings of the National Academy of Sciences*, 91(24):11733–11737, 1994.

[118] Christopher A. Moskaluk, Paul H. Duray, Kenneth H. Cowan, Marston Linehan, and Maria J. Merino. Immunohistochemical expression of π-class glutathione s-transferase is down-regulated in adenocarcinoma of the prostate. *Cancer*, 79(8):1595–1599, 1997.

[119] A Bagheri, SM Nachvak, M Rezaei, M Moravridzade, M Moradi, and M Nelson. Dietary patterns and risk of prostate cancer: a factor analysis study in a sample of Iranian men. *Health Promotion Perspectives*, 8(2):133–138, 2018.

[120] DC Muller, G Severi, L Baglietto, K Krishnan, DR English, JL Hopper, and GG Giles. Dietary patterns and prostate cancer risk. *Cancer Epidemiology, biomarkers and prevention*, 18(11):3126–3129, 2009.

[121] American Cancer Society. What Tests Can Detect Prostate Cancer Early? https://www.cancer.org/cancer/prostate-cancer/early-detection/tests.html, 2016. Accessed: May 2017.

[122] Amitava Dasgupta and Amer Wahed. *Clinical Chemistry, Immunology and Laboratory Quality Control*. Elsevier Inc, 2014.

[123] Hans Lilja, David Ulmert, and Andrew J Vickers. Prostate-specific antigen and prostate cancer: prediction, detection and monitoring. *Nature Reviews Cancer*, 8(4):268–278, 2008.

[124] Ruth Etzioni, David F. Penson, Julie M. Legler, Dante di Tommaso, Rob Boer, Peter H. Gann, and Eric J. Feuer. Overdiagnosis due to prostate-specific antigen screening: Lessons from u.s. prostate cancer incidence trends. *JNCI: Journal of the National Cancer Institute*, 94(13):981, 2002.

[125] Kathleen Y Wolin, Jason Luly, Siobhan Sutcliffe, Gerald L Andriole, and Adam S Kibel. Risk of Urinary Incontinence Following Prostatectomy: The Role of Physical Activity and Obesity. *Journal of Urology*, 183(2):629–633, 2010.

[126] Andrew R McCullough. Sexual Dysfunction after Radical Prostatectomy. *Reviews in Urology*, 7(Suppl 2):S3–S10, 2005.

[127] William T. Lowrance, Elena B. Elkin, Lindsay M. Jacks, MS, David S. Yee, Thomas L. Jang, Vincent P. Laudone, and Bertrand D. Guillonneau and. Comparative Effectiveness of Surgical Treatments for Prostate Cancer: A Population-Based Analysis of Postoperative Outcomes. *Journal of Urology*, 183(4):1366–1372, 2010.

[128] Debra E. Irwina, Ian Milsom, Steinar Hunskaar, Kate Reilly, Zoe Kopp, Sender Herschorn, Karin Coyne, Con Kelleher, Christian Hampel, Walter Artibani, and Paul Abrams. Population-Based Survey of Urinary Incontinence, Overactive Bladder, and Other Lower Urinary Tract Symptoms in Five Countries: Results of the EPIC Study. *European Urology*, 50(6):1306–1315, 2006.

[129] Ian M. Thompson, Donna K. Pauler, Phyllis J. Goodman, Catherine M. Tangen, M. Scott Lucia, Howard L. Parnes, Lori M. Minasian, Leslie G. Ford, Scott M. Lippman, E. David Crawford, John J. Crowley, and Charles A. Jr. Coltman. Prevalence of prostate cancer among men with a prostate-specific antigen level $\leq$ 4.0 ng per milliliter. *New England Journal of Medicine*, 350(22):2239–2246, 2004.

[130] Axel Heidenreich, Patrick J Bastian, Joaquim Bellmunt, Michel Bolla, Steven Joniau, Theodor van der Kwast, Malcolm Mason, et al. Eau guidelines on prostate cancer. part 1: Screening, diagnosis, and local treatment with curative intent—update 2013. *European Urology*, 65(1):124–137, 2014.

[131] National Institute for Health and Care Excellence. Prostate cancer: diagnosis and management. https://www.nice.org.uk/guidance/cg175/resources/%20prostate-cancer-diagnosis-and-management-35109753913285, 2014. Accessed: May 2017.

[132] A V Taira, G S Merrick ad R W Galbreath, H Andreini, W Taubenslag, R Curtis, W M Butler, et al. Performance of transperineal template-guided mapping biopsy in detecting prostate cancer in the initial and repeat biopsy setting. *Prostate Cancer and Prostatic Diseases*, 13(1):71–77, 2010.

[133] Hashim U Ahmed, Ahmed El-Shater Bosaily, Louise C Brown, Rhian Gabe, Richard Kaplan, Mahesh K Parmar, Yolanda Collaco-Moraes, et al. Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. *The Lancet*, 389(10071):815–822, 2017.

[134] Donald F Gleason. Histologic grading of prostate cancer: A perspective. *Human Pathology*, 23(3):273–279, 1992.

[135] Jennifer R Stark, Sven Perner, Meir J Stampfer, Jennifer A Sinnott, Stephen Finn, Anna S Eisenstein, Jing Ma, et al. Gleason Score and Lethal Prostate Cancer: Does 3 + 4 = 4 + 3? *Journal of Clinical Oncology*, 27(21):3459–3464, 2009.

[136] James Brierley, M. K. Gospodarowicz, and Ch Wittekind. *TNM classification of malignant tumours*. John Wiley & Sons, Inc., 2017.

[137] Geore Rodrigues, Padraig Warde, Tom Pickles ad Juanita Crook, Michael Brundage, Luis Souhami, and Himu Lukka. Pre-treatment risk stratification of prostate cancer patients: A critical review. *Canadian Urological Associaton Journal*, 6(2):121–127, 2012.

[138] Laurence Klotz, Liying Zhang, Adam Lam, Robert Nam, Alexandre Mamedov, and Andrew Loblaw. Clinical results of long-term follow-up of a large, active surveillance cohort with localized prostate cancer. *Journal of Clinical Oncology*, 28(1):126–131, 2010.

[139] National Institute for Health and Care Excellence. Low dose rate brachytherapy for localised prostate cancer. https://www.nice.org.uk/guidance/ipg132, 2005. Accessed: February 2021.

[140] Juanita Mary Crook, Alfonso Gomez-Iturriaga, Kris Wallace, Clement Ma, Sharon Fung, Shabbir Alibhai, Michael Jewett, and Neil Fleshner. Comparison of health-related quality of life 5 years after spirit: Surgical prostatectomy versus interstitial radiation intervention trial. *Journal of Clinical Oncology*, 29(4):362–368, 2011.

[141] Dirk Böhmer, Manfred Wirth, Kurt Miller, Volker Budach, Axel Heidenreich, and Thomas Wiegel. Radiotherapy and Hormone Treatment in Prostate Cancer. *Deutsches Ärzteblatt International*, 113(14):235–241, 2016.

[142] Fernando J Bianco, Peter T Scardino, and James A Eastham. Radical prostatectomy: long-term cancer control and recovery of sexual and urinary function ("trifecta"). *Urology*, 66(5 Suppl):83—94, November 2005.

[143] Channing J Paller and Emmanuel S Antonarakis. Management of Biochemically Recurrent Prostate Cancer After Local Therapy: Evolving Standards of Care and New Directions. *Clinical Advances in Hematology and Oncology*, 11(1):14–23, 2013.

[144] Culley Carson III and Roger Rittmaster. The role of dihydroteserone in benign prostatic hyperplasia. *Urology*, 61(4, Supplement 1):2 – 7, 2003.

[145] Johan F. Langenhuijsen, Emile N. van Lin, Aswin L. Hoffmann, Ilse Spitters-Post, J. Alfred Witjes, Johannes H. Kaanders, and Peter F. Mulders. Neoadjuvant androgen deprivation for prostate volume reduction: The optimal duration in prostate cancer radiotherapy. *Urologic Oncology: Seminars and Original Investigations*, 29(1):52 – 57, 2011. Plagiarism.

[146] Leoonard G Gomella. Hormone therapy in the management of prostate cancer: evidence-based approaches. *Therapeitic Advances in Urology*, 2(4):171–181, 2010.

[147] Gary H Lyman. *Oxford American Handbook of Oncology*. Oxford University Press, 2 edition, 2015.

[148] Michel Bolla, Dionisio Gonzalez, Padraig Warde, Jean Bernard Dubois, René-Olivier Mirimanoff, Guy Storme, Jacques Bernier, et al. Improved survival in patients with locally advanced prostate cancer treated with radiotherapy and goserelin. *New England Journal of Medicine*, 337(5):295–300, 1997.

[149] Michel Bolla, Laurence Collettea, Léo Blank, Padraig Warde, Jean Bernard Dubois, René-Olivier Mirimanoff, Guy Storme, et al. Long-term results with immediate androgen suppression and external irradiation in patients with locally advanced prostate cancer (an EORTC study): a phase III randomised trial. *The Lancet*, 360(9327):103–108, 2002.

[150] Norra MacReady. Quality of Life Declines With Androgen Deprivation Therapy. http://www.medscape.com/viewarticle/819916, 2014. Accessed: May 2017.

[151] Fred Saad and Sebastien J Hotte. Guidelines for the management of castrate-resistant prostate cancer. *Canadian Urological Association Journal*, 4(6):380–384, 2010.

[152] Sok Kuan Wong, Nur-Vaizura Mohamad, Tijjani Rabiu Giaze, Kok-Yong Chin, Norazlina Mohamed, and Soelaiman Ima-Nirwana. Prostate Cancer and Bone Metastases: The Underlying Mechanisms. *International Journal of Molecular Sciences*, 20(10):e2587, 2019.

[153] Christopher Logothetis, Michael J. Morris, Robert Den, and Robert E. Coleman. Current perspectives on bone metastases in castrate-resistant prostate cancer. *Cancer Metastasis Reviews*, 37(1):189–196, 2018.

[154] F Saad, J Carles, S Gillessen, A Heidenreich, D Heinrich, J Gratt, J Lèvy, K Miller, S Nilsson, O Petrenciuc, M Tucci, M Wirth, J Federhofer, and JM O'Sullivan. Radium-223 and concomitant therapies in patients with metastatic castration-resistant prostate cancer: an international, early access, open-label, single-arm phase 3b trial. *The Lancet Oncology*, 17(19):1306–1316, 2016.

[155] Barry S Taylor, Nikolaus Schultz, Haley Hieronymus, Anuradha Gopalan, Yonghong Xaiao, Brett S Carver, Vivek K Arora, et al. Integrative genomic profiling of human prostate cancer. *Cancer Cell*, 18(1):11–22, 2010.

[156] H Ross-Adams, AD Lamb, MJ Dunning, S Halim, J Lindberg, CM Massie, LA Egevad, et al. Integration of copy number and transcriptomics provides risk stratification in prostate cancer: A discovery and validation cohort study. *EBioMedicine*, 2(9):1133–1144, 2015.

[157] Andrew J Stephenson, Alex Smith, Michael W Kattan, Jaya Satagopan, Victor E Reuter, Peter T Scardino, and William L Gerald. Integration of Gene Expression Profiling and Clinical Variables to Predict Prostate Carcinoma Recurrence after Radical Prostatectomy. *Cancer*, 104(2):290–298, 2005.

[158] Eric A Klein, Kasra Yousefi, Zaid Haddad, Voleak Choeurng, Christine Buerki, Andrew J Stephenson, Jianbo Li, et al. A Genomic Classifier Improves Prediction of Metastatic Disease Within 5 Years After Surgery in Node-negative High-risk Prostate Cancer Patients Managed by Radical Prostatectomy Without Adjuvant Therapy. *European Urology*, 67(4):778–786, 2015.

[159] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[160] Guangchuang Yu, Li-Gen Wang, and Giovanni Dall'Olio. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology*, 16(5):284–287, 2014.

[161] Victoria Sanz-Moreno, Gilles Gadea, Jessica Ahn, Hugh Paterson, Pierfrancesco Marra, Sophie Pinner, Erik Sahai, and Christopher J Marshall. Rac activation and inactivation control plasticity of tumor cell movement. *Cell*, 135(3):510–523, 2008.

[162] Peter Friedl, Joseph Locker, Erik Sahai, and Jeffrey E Segall. Classifying collective cancer cell invasion. *Nature Cell Biology*, 14(8):777–783, 2012.

[163] Gil Stelzer, Naomi Rosen, Inbar Plaschkes, Shahar Zimmerman, Michal Twik, Simon Fishilevich, Tsippi Iny Stein, Ron Nudel, Iris Lieder, Yaron Mazor, Sergey Kaplan, Dvir Dahary, David Warshawsky, Yaron Guan-Golan, Asher Kohn, Noa Rappaport, Marilyn Safran, and Doron Lancet. BRAF gene: From human cancers to developmental syndromes. *Saudi Journal of Biological Sciences*, 54(1):1.30.1–1.30.33, 2016.

[164] Ruth Hull. *ANATOMY & PHYSIOLOGY for therapists and healthcare professionals*. The Write Idea Ltd, 2009.

[165] John. M Carethers. Risk factors for colon location of cancer. *Translational Gastroenterology and Hepatology*, 3(76), 2018.

[166] Kannan Thanikachalam and Gazala Khan. Colorectal Cancer and Nutrition. *Nutrients*, 11(1):164, 2019.

[167] H. J Schmoll, E Van Cutsem, A Stein, V Valentini, B Glimelius, K Haustermans, B Nordlinger, C. J van de Velde, J Balmana, J Regula, I. D Nagtegaal, R. G Beets-Tan, D Arnold, F Ciardiello, P Hoff, D Kerr, C. H Köhne, R Labianca, T Price, W Scheithauer, A Sobrero, J Tabernero, D Aderka, S Barroso, G Bodoky, J. Y Douillard, H El Ghazaly, J Gallardo, A Garin, R Glynne-Jones, K Jordan, A Meshcheryakov, D Papamichail, P Pfeiffer, I Souglakos, S Turhal, and A Cervantes. ESMO Consensus Guidelines for management of patients with colon and rectal cancer. a personalized approach to clinical decision making. *Annals of Oncology*, 23(10):2479–2516, 2012.

[168] ASCO. Colorectal Cancer: Risk Factors and Prevention. https://www.cancer.net/cancer-types/colorectal-cancer/risk-factors-and-prevention, 2019. Accessed: March 2020.

[169] Akiko Tamakoshi, Koshi Nakamura, Shigekazu Ukawa, Emiko Okada, Makoto Hirata, Akiko Nagai, Koichi Matsuda, Yoichiro Kamatani, Kaori Muto, Yutaka Kiyohara, Zentaro Yamagata, Toshiharu Ninomiya, Michiaki Kubo, Yusuke Nakamura, and BioBank Japan Cooperative Hospital Group. Characteristics and prognosis of Japanese colorectal cancer patients: The BioBank Japan Project. *Journal of Epidemiology*, 27(3):S36–S42, 2017.

[170] Cancer Research UK. Bowel cancer incidence statistics. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer/incidence, 2019. Accessed: March 2020.

[171] Rebecca L. Siegel, Kimberly D. Miller, Ann Goding Sauer, Stacey A. Fedewa, Lynn F. Butterly, Joseph C. Anderson, Andrea Cercek, Robert A. Smith, and Ahmedin Jemal. Colorectal cancer statistics, 2020. *CA: A Cancer Journal for Clinicians*, 70(2), 2020. EARLY PRINT, INSERT UPDATE HERE FOR PAGES - volume expected to be published in April or May.

[172] Jo Cavallo. Study Suggests Risk of Cancer Death Increases With Each Generation of Latinos Born in the United States. https://ascopost.com/News/59438, 2018. Accessed: March 2020.

[173] Junus M. van der Wal, Adee Bodewes, Charles Agyemang, and Anton Kunst. A population-based retrospective study comparing cancer mortality between Moluccan migrants and the general Dutch population: equal risk 65 years after immigration? *BMJ Open*, 9(8):e029288, 2019.

[174] Alan Moss and Kumanan Nalankilli. The Association Between Diet and Colorectal Cancer Risk: Moving Beyond Generalizations. *Gastroenterology*, 152(8):1821–1823, 2017.

[175] Irfan M Hisamuddin and Vincent W Yang. Genetics of Colorectal Cancer. *Medscape General Medicine*, 6(3):13, 2004.

[176] P Galiatsatos and W. D Foulkes. Familial adenomatous polyposis. *The American journal of Gastroenterology*, 101(2):385–398, 2006.

[177] Hermann Brenner, Lina Jansen, Alexis Ulrich, Jenny Chang-Claude, and Michael Hoffmeister. Survival of patients with symptom- and screening-detected colorectal cancer. *Oncotarget*, 7(28):44695–44704, 2016.

[178] Sara Koo, Laura Jane Neilson, Christian Von Wagner, and Colin John Rees. The NHS Bowel Cancer Screening Program: current perspectives on strategies for improvement. *Risk Management and Healthcare Policy*, 10:177–187, 2017.

[179] Ann G Zauber. The Impact of Screening on Colorectal Cancer Mortality and Incidence – Has It Really Made a Difference? *Digestive Diseases and Science*, 60(3):681–691, 2015.

[180] Public Health England. Bowel cancer screening: programme overview. https://www.gov.uk/guidance/bowel-cancer-screening-programme-overview, 2019. Accessed: March 2020.

[181] The Association for Clinical Biochemistry & Laboratory Medicine. Faecal Occult Blood Test and Faecal Immunochemical Test. https://labtestsonline.org.uk/tests/faecal-occult-blood-test-and-faecal-immunochemical-test, 2019. Accessed: March 2020.

[182] Marie Westwood and Shona Lang and Isaac Corro Ramos and Maiwenn Al. Diagnostics Assessment Report commissioned by the NIHR HTA Programme on behalf of the National Institute for Health and Care Excellence – Protocol. https://www.nice.org.uk/guidance/dg30/documents/final-protocol, 2016. Accessed: March 2020.

[183] National Institute for Health and Care Excellence. Computed tomographic colonography (virtual colonoscopy) Interventional procedures guidance [IPG129]. https://www.nice.org.uk/guidance/IPG129, 2005. Accessed: March 2020.

[184] The American Society of Clinical Oncology. Colorectal Cancer: Diagnosis. https://www.cancer.net/cancer-types/colorectal-cancer/diagnosis, 2019. Accessed: March 2020.

[185] National Institute for Health and Care Excellence. Colorectal cancer NICE guideline [NG151]. https://www.nice.org.uk/guidance/ng151, 2020. Accessed: March 2020.

[186] The American Society of Clinical Oncology. Stages of Cancer. https://www.cancer.net/navigating-cancer-care/diagnosing-cancer/stages-cancer, 2018. Accessed: March 2020.

[187] The American Cancer Society. Colorectal Cancer Stages. https://www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/staged.html, 2018. Accessed: March 2020.

[188] Cancer Research UK. TNM Staging. https://www.cancerresearchuk.org/about-cancer/bowel-cancer/stages-types-and-grades/TNM-staging, 2019. Accessed: March 2020.

[189] Cancer Research UK. Dukesśtaging system. https://www.cancerresearchuk.org/about-cancer/bowel-cancer/stages-types-and-grades/dukes-staging, 2018. Accessed: March 2020.

[190] The National Guideline Alliance. Colorectal Cancer (update) [C2] Preoperative radiotherapy and chemoradiotherapy for rectal cancer. https://www.nice.org.uk/guidance/ng151/evidence/c2-preoperative-radiotherapy-and-chemoradiotherapy-for-rectal-cancer-pdf-7029391217, 2020. Accessed: March 2020.

[191] The American Cancer Society. Radiation Therapy for Colorectal Cancer. https://www.cancer.org/cancer/colon-rectal-cancer/treating/radiation-therapy.html, 2018. Accessed: March 2020.

[192] National Institute for Health and Care Excellence. Colorectal cancer: the diagnosis and management of colorectal cancer. https://www.nice.org.uk/guidance/cg131/documents/colorectal-cancer-full-guideline2, 2011. Accessed: March 2020.

[193] The American Society of Clinical Oncology. Colorectal Cancer: Statistics. https://www.cancer.net/cancer-types/colorectal-cancer/statistics, 2020. Accessed: April 2020.

[194] National Institute for Health and Care Excellence. Denosumab for the prevention of skeletal-related events in adults with bone metastases from solid tumours. https://www.nice.org.uk/guidance/ta265/chapter/4-Evidence-and-interpretation, 2012. Accessed: April 2020.

[195] Li-Jun Wang, Hong-Wei Wang, Ke-Min Jin, Juan Li, and Bao-Cai Xing. Comparison of sequential, delayed and simultaneous resection strategies for synchronous colorectal liver metastases. *BMC Surgery*, 20:16, 2020.

[196] National Institute for Health and Care Excellence. Selective internal radiation therapy for unresectable colorectal metastases in the liver. https://www.nice.org.uk/guidance/IPG672/chapter/1-Recommendations, 2020. Accessed: April 2020.

[197] Randell Bert. Inheritance of colorectal cancer. *Drug Discovery Today: Disease Mechanisms*, 4(4):293–300, 2007.

[198] T Philip Chung and James W Fleshman. The Genetics of Sporadic Colon Cancer. *Seminars in Colon and Rectal Surgery*, 15(3):128–135, 2004.

[199] William M Grady and Colin C Pritchard. Molecular Alterations and Biomarkers in Colorectal Cancer. *Toxicologic Pathology*, 42(1):124–139, 2013.

[200] Andy Hin-Fung Tsang, Ka-Ho Cheng, Apple Siu-Ping Wong, Simon Siu-Man Ng, Brigette Buig-Yue Ma, Charles Ming-Lok Chan, Nancy Bo-Yin Tsui, Lawrence Wing-Chi Chan, Benjamin Yat-Ming Yung, and Sze-Chuen Cesar Wong. Current and future molecular diagnostics in colorectal cancer and colorectal adenoma. *World Journal of Gastroenterology*, 20(14):3847–3857, 2014.

[201] Wade S. Samowitz, Hans Albertsen, Jennifer Herrick, Theodore R. Levin, Carol Sweeney, Maureen A. Murtaugh, Roger K. Wolff, and Martha L. Slattery. Evaluation of a Large, Population-Based Sample Supports a CpG Island Methylator Phenotype in Colon Cancer. *Gasteroenterology*, 129(3):837–845, 2005.

[202] Liang Wang, Linnea M. Baudhuin, Lisa A. Boardman, Kelle J. Steenblock, Gloria M. Petersen, Kevin C. Halling, Amy J. French, Ruth A. Johnson, Lawrence J. Burgart, Kari Rabe, Noralane M. Lindor, and Stephen N. Thibodeau. MYH mutations in patients with attenuated and classic polyposis and with young-onset colorectal cancer without polyps. *Gasteroenterology*, 127(1):9–16, 2004.

[203] H F A Vasen, G Möslein, A Alonso, I Bernstein, L Bertario, I Blanco, J Burn, G Capella, C Engel, I Frayling, W Friedl, F J Hes, S Hodgson, J-P Mecklin, P Møller, F Nagengast, Y Parc, L Renkonen-Sinisalo, J R Sampson, A Stormorken, and J Wijnen. Guidelines for the clinical management of Lynch syndrome (hereditary non-polyposis cancer). *Journal of Medical Genetics*, 44(6):353–362, 2007.

[204] National Institute for Health and Care Excellence. Molecular testing strategies for Lynch syndrome in people with colorectal cancer. https://www.nice.org.uk/guidance/dg27/chapter/1-Recommendations, 2017. Accessed: April 2020.

[205] Angelika Copija, Dariusz Waniczek, Andrzej Witkoś, Katarzyna Walkiewicz, and Ewa Nowakowska-Zajdel. Clinical Significance and Prognostic Relevance of Microsatellite Instability in Sporadic Colorectal Cancer Patients. *International Journal of Molecular Sciences*, 18(1):107, 2017.

[206] Won-Seok Jo and John M. Carethers. Chemotherapeutic implications in microsatellite unstable colorectal cancer. *Cancer Biomarkers*, 2(1):51–60, 2006.

[207] Hisato Kawakami, Aziz Zaanan, and Frank A. Sinicrope. MSI testing and its role in the management of colorectal cancer. *Currative Treatment Options in Oncology*, 16(7):30, 2015.

[208] Maria S Pino and Daniel C Chung. The Chromosomal Instability Pathway In Colon Cancer. *Gastroenterology*, 138(6):2059–2072, 2010.

[209] Tannaz Armaghany, Jon D. Wilson, Quyen Chu, and Glenn Mills. Genetic Alterations in Colorectal Cancer. *Gastrointestinal Cancer Research*, 5(1):19–27, 2012.

[210] I Munteanu and B Mastalier. Genetics of colorectal cancer. *Journal of Medince and Life*, 7(4):507–511, 2014.

[211] NaNa Keum and Edward Giovannucci. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. *Nature Reviews Gastroenterology & Hepatology*, 16:713–732, 2019.

[212] Kevin M Haigis. KRAS Alleles: The Devil Is In The Detail. *Trends in Cancer*, 3(10):686–697, 2017.

[213] Fatima Domenica Elisa De Palma, Valeria D'Argenio, Jonathan Pol, Guido Kroemer, Maria Chiara Maiuri, and Francesco Salvatore. The Molecular Hallmarks of the Serrated Pathway in Colorectal Cancer. *Cancers*, 11(7):1017, 2019.

[214] Muhammad Ramzan Manwar Hussain, Mukhtiar Baig, Hussein Sheik Ali Mohamoud, Zaheer Ulhaq, Daniel C Hoessli, Ghaidaa Siraj Khogeer, Ranem Radwan Al-Sayed, and Jumana Yousuf Al-Aamaa. BRAF gene: From human cancers to developmental syndromes. *Saudi Journal of Biological Sciences*, 22(4):359–373, 2015.

[215] M. M Bertagnolli, M Redston, C. C Compton, D Niedzwiecki, R. J Mayer, R. M Goldberg, T. A Colacchio, L. B Saltz, and R. S Warren. Microsatellite instability and loss of heterozygosity at chromosomal location 18q: prospective evaluation of biomarkers for stages II and III colon cancer. *Journal of Clinical Oncology*, 29(23):3153–3162, 2011.

[216] I. J Dahabreh, T Terasawa, P. J Castaldi, and T. A Trikalinos. Microsatellite instability and loss of heterozygosity at chromosomal location 18q: prospective evaluation of biomarkers for stages II and III colon cancer. *Annals of Internal Medicine*, 154(1):37–49, 2011.

[217] T. K Guren, M Thomsen, E. H Kure, H Sorbye, B Glimelius, P Pfeiffer, P Österlund, F Sigurdsson, I. M. B Lothe, A. M Dalsgaard, E Skovlund, T Christoffersen, and K. M Tveit. Microsatellite instability and loss of heterozygosity at chromosomal location 18q: prospective evaluation of biomarkers for stages II and III colon cancer. *British Journal of Cancer*, 116(10):1271–1278, 2017.

[218] RN1 Jorissen, P Gibbs, M Christie, S Prakash, L Lipton, J Desai, D Kerr, LA Aaltonen, D Arango, M Kruhøffer, TF Orntoft, CL Andersen, M Gruidl, VP Kamath, S Eschrich, TJ Yeatman, and OM. Sieber. Metastasis-Associated Gene Expression Changes Predict Poor Outcomes in Patients with Dukes Stage B and C Colorectal Cancer. *Clinical cancer research*, 15(24):7642–7651, 2009.

[219] JJ Smith, NG Deane, F Wu, NB Merchant, B Zhang, A Jiang, P Lu, JC Johnson, C Schmidt, CE Bailey, S Eschrich, C Kis, S Levy, MK Washington, MJ Heslin, RJ Coffey, TJ Yeatman, Y Shyr, and RD Beauchamp. Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. *Gastroenterology*, 138(3):958–968, 2010.

[220] L Marisa, A de Reyniés, A Duval, J Selves, MP Gaub, L Vescovo, MC Etienne-Grimaldi, R Schiappa, D Guenot, M Ayadi, S Kirzin, M Chazal, JF Fléjou, D Benchimol, A Berger, A Lagarde, E Pencreach, F Piard, D Elias, Y Parc, S Olschwang, G Milano, P Laurent-Puig, and V. Boige. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS Medicine*, 10(5):e1001453, 2013.

[221] ML Martin, Z Zeng, M Adileh, A Jacobo, C Li, E Vakiani, G Hua, L Zhang, A Haimovitz-Friedman, Z Fuks, R Kolesnick, and PB. Paty. Logarithmic expansion of $LGR5^+$ cells in human colorectal cancer. *Cellular Signalling*, 42:97–105, 2018.

[222] HH Lin, NC Wei, TY Chou, CC Lin, YT Lan, SC Chang, HS Wang, SH Yang, WS Chen, TC Lin, JK Lin, and JK1. Jiang. Building personalized treatment plans for early-stage colorectal cancer patients. *Oncotarget*, 8(8):13895–13817, 2017.

[223] The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487:330–337, 2012.

[224] Benilton S Carvalho and Rafael A Irizarry. A framework for oligonucleotide microarray preprocessing. *Bioinformatics*, 26(19):2363–2367, 2010.

[225] Michael Love, Constantin Ahlmann-Eltze, Kwame Forbes, Simon Anders, and Wolfgang Huber. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15:550, 2014.

[226] Jeffrey T. Leek, W. Evan Johnson, Hilary S. Parker, Elana J. Fertig, Andrew E. Jaffe, John D. Storey, Yuqing Zhang, and Leonardo Collado Torres. *sva: Surrogate Variable Analysis*, 2019. R package version 3.30.1.

[227] Brian Duignan. Occam's razor. https://www.britannica.com/topic/Occams-razor, 2018. Accessed: June 2020.

[228] RefSeq. TIMP1 TIMP metallopeptidase inhibitor 1. https://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=DetailsSearch&Term=7076, 2020. Accessed: July 2020.

[229] Chunyan Meng, Xiaowei Yin, Jingting Liu, Kaifeng Tang, Hongchao Tang, and Jianhua Liao. TIMP-1 is a novel serum biomarker for the diagnosis of colorectal cancer: A meta-analysis. *PLoS One*, 13(11):e0207039, 2018.

[230] Qi Chong Qi, Hong Liang, Cheng Zhijian, and Yin Qingzhang. Identification of metastasis-associated genes in colorectal cancer using metaDE and survival analysis. *Oncology Letters*, 11(1):568–574, 2015.

[231] Mathias Uhlén, Erik Björling, Charlotta Agaton, Cristina Al-Khalili Szigyarto, Bahram Amini, Elisabet Andersen, Ann-Catrin Andersson, Pia Angelidou, Anna Asplund, Caroline Asplund, Lisa Berglund, Kristina Bergström, Harry Brumer, Dijana Cerjan, Marica Ekström, Adila Elobeid, Cecilia Eriksson, Linn Fagerberg, Ronny Falk, Jenny Fall, Mattias Forsberg, Marcus Gry Björklund, Kristoffer Gumbel, Asif Halimi, Inga Hallin, Carl Hamsten, Marianne Hansson, My Hedhammar, Görel Hercules, Caroline Kampf, Karin Larsson, Mats Lindskog, Wald Lodewyckx, Jan Lund, Joakim Lundeberg, Kristina Magnusson, Erik Malm, Peter Nilsson, Jenny Ödling, Per Oksvold, Ingmarie Olsson, Emma Öster, Jenny Ottosson, Linda Paavilainen, Anja Persson, Rebecca Rimini, Johan Rockberg, Marcus Runeson, Åsa Sivertsson, Anna Sköllermo, Johanna Steen, Maria Stenvall, Fredrik Sterky, Sara Strömberg, Mårten Sundberg, Hanna Tegel, Samuel Tourle, Eva Wahlund, Annelie Waldén, Jinghong Wan, Henrik Wernérus, Joakim Westberg, Kenneth Wester, Ulla Wrethagen, Lan Lan Xu, Sophia Hober, and Fredrik Pontén. A human protein atlas for normal and cancer tissues based on antibody proteomics. *Molecular & Cellular Proteomics*, 4(12):1920–1932, 2005.

[232] Huang DW, Sherman BT, and Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13, 2009.

[233] Huang DW, Sherman BT, and Lempicki RA. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature Protocols*, 4(1):44–57, 2009.

[234] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genetics*, 25(1):25–29, 2000.

[235] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G.R. Gopinath, G.R. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, 33(Database issue):D428–D432, 2005.

[236] Anne Katrin Lampe, Kevin M Flanigan, Katharine Mary Bushby, and Debbie Hicks. Collagen type vi-related disorders. In Margaret P Adam, Holly H Ardinger, Roberta A Pagon, Stephanie E Wallace, Lora JH Bean, Karen Stephens, and Anne Amemiya, editors, *GeneREVIEWS*. University of Washington Press, Seattle, 2012. Available from: https://www.ncbi.nlm.nih.gov/books/NBK1503/.

[237] NIH U.S. National Library of Medicine. Collagen VI-related myopathy. https://ghr.nlm.nih.gov/condition/collagen-vi-related-myopathy, 2020. Accessed: May 2020.

[238] Michael J. O'Connell, Ian Lavery, Greg Yothers, Soonmyung Paik, Kim M. Clark-Langone, Margarita Lopatin, Drew Watson, Frederick L. Baehner, Steven Shak, Joffre Baker, J. Wayne Cowens, and Norman Wolmark. Relationship between tumor gene expression and recurrence in four independent studies of patients with stage ii/iii colon cancer treated with surgery alone or surgery plus adjuvant fluorouracil plus leucovorin. *Journal of Clinical Oncology*, 28(25):3937–3944, 2010.

[239] Huarong Chen, Xiaoqiang Sun, Weiting Ge, Yun Qian, Rui Bai, and Shu Zheng. A seven-gene signature predicts overall survival of patients with colorectal cancer. *Oncotarget*, 8(56):95054–95065, 2017.

[240] R Salazar, P Roepman, G Capella, V Moreno, I Simon, C Dreezen, A Lopez-Doriga, C Santos, C Marijnen, J Westerga, S Bruin, D Kerr, P Kuppen, C van de Velde, H Morreau, L Van, Velthuysen, AM Glas, LJ Van't, Veer, and R Tollenaar. Gene expression signature to improve prognosis prediction of stage ii and iii colorectal cancer. *Journal of Clinical Oncology*, 29(1):17–24, 2011.

[241] Richard D. Kennedy, Max Bylesjo, Peter Kerr, Timothy Davison, Julie M. Black, Elaine W. Kay, Robert J. Holt, Vitali Proutski, Miika Ahdesmaki, Vadim Farztdinov, Nicolas Goffard, Peter Hey, Fionnuala McDyer, Karl Mulligan, Julie Mussen, Eamonn O'Brien, Gavin Oliver, Steven M. Walker, Jude M. Mulligan, Claire Wilson, Andreas Winter, Diarmuid O'Donoghue, Hugh Mulcahy, Jacintha O'Sullivan, Kieran Sheahan, John Hyland, Rajiv Dhir, Oliver F. Bathe, Ola Winqvist, Upender Manne, Chandrakumar Shanmugam, Sridhar Ramaswamy, Eduardo J. Leon, William I. Smith, Ultan McDermott, Richard H. Wilson, Daniel Longley, John Marshall, Robert Cummins, Daniel J. Sargent, Patrick G. Johnston, and D. Paul Harkin. Development and independent validation of a prognostic assay for stage ii colon cancer using formalin-fixed paraffin-embedded tissue. *Journal of Clinical Oncology*, 29(35):4620–4626, 2011.

[242] Ping Gao, Miao Heb, Chunling Zhang, and Changhui Geng. Integrated analysis of gene expression signatures associated with colon cancer from three datasets. *Gene*, 654:95–102, 2018.

[243] Sang Cheul Oh, Yun-Yong Park, Eun Sung Park, Jae Yun Lim, Soo Mi Kim, Sang-Bae Kim, Jongseung Kim, Sang Cheol Kim, In-Sun Chu, J Joshua Smith, R Daniel Beauchamp, Timothy J Yeatman, Scott Kopetz, and Ju-Seog Lee. Prognostic gene expression signature associated with two molecularly distinct subtypes of colorectal cancer. *Gut*, 61(9):1291–1298, 2012.

[244] Dalong Sun, Jing Chen, Longzi Liu, Guangxi Zhao, Pingping Dong, Bingrui Wu, Jun Wang, and Ling Dong. Establishment of a 12-gene expression signature to predict colon cancer prognosis. *PeerJ*, 6:e4942, 2018.

[245] Peng Shu, Jianping Wu, Yao Tong, Chunxia Xu, and Xingguo Zhang. Gene pair based prognostic signature for colorectal colon cancer. *Medicine*, 97(42):e12788, 2018.

[246] Li Chunsheng, Shen Zhen, Zhou Yangyang, and Wei Yu. Independent prognostic genes and mechanism investigation for colon cancer. *Biological Research*, 51(1):10, 2018.

[247] Guangwei Sun, Yalun Li, Yangjie Peng, Dapeng Lu, Fuqiang Zhang, Xueyang Cui, Qingyue Zhang, and Zhuang Li. Identification of a five-gene signature with prognostic value in colorectal cancer. *Journal of Cellular Physiology*, 234(4):3829–3836, 2019.

[248] Stefano Maria Pagnotta, Carmelo Laudanna, Massimo Pancione, Lina Sabatino, Carolina Votino, Andrea Remo, Luigi Cerulo, Pietro Zoppoli, Erminia Manfrin, Vittorio Colantuoni, and Michele Ceccarelli. Ensemble of gene signatures identifies novel biomarkers in colorectal cancer activated through pparγ and tnfα signaling. *PLOS ONE*, 8(8):e72638, 2013.

[249] Yida Pan, Hongyang Zhang, Mingming Zhang, Jie Zhu, Jianghong Yu, Bangting Wang, Jigang Qiu, and Jun Zhang. A five-gene based risk score with high prognostic value in colorectal cancer. *Oncology Letters*, 14(6):6724–6734, 2017.

[250] RefSeq. KLK10 kallikrein related peptidase 10. https://www.ncbi.nlm.nih.gov/gene/5655, 2020. Accessed: July 2020.

[251] Xu Ren and Pei Fen Kuan. methyGSA: a bioconductor package and shiny app for dna methylation data length bias adjustment in gene set testing. *Bioinformatics*, 35(11):e3917, 2019.

[252] Yanhong Deng, Li Wang, Shuyun Tan, George P. Kim, Ruoxu Dou, Dianke Chen, Yue Cai, Xinhui Fu, Lei Wang, Jun Zhu, and Jianping Wang. Kras as a predictor of poor prognosis and benefit from postoperative folfox chemotherapy in patients with stage ii and iii colorectal cancer. *Molecular Oncology*, 9(7):1341–1347, 2015.

[253] H. Blons, J.F. Emile, K. Le Malicot, C. Julié, A. Zaanan, J. Tabernero, E. Mini, G. Folprecht, J.L. Van Laethem, J. Thaler, J. Bridgewater, L. Nørgård-Petersen, E. Van Cutsem, C. Lepage, M.A. Zawadi, R. Salazar, P. Laurent-Puig, and J. Taieb. Prognostic value of kras mutations in stage iii colon cancer: post hoc analysis of the petacc8 phase iii trial dataset. *Annals of Oncology*, 25(12):2378–2385, 2014.

[254] Fritz H Schröder, Jonas Hugosson, Monique J Roobol, Teuvo L J Tammela, Marco Zappa, Vera Nelen, Maciej Kwiatkowski, Marcos Lujan, Liisa Määttänen, Hans Lilja, Louis J Denis, Franz Recker, Alvaro Paez, Chris H Bangma, Sigrid Carlsson, Donella Puliti, Arnauld Villers, Xavier Rebillard, Matti Hakama, Ulf-Hakan Stenman, Paula Kujala, Kimmo Taari, Gunnar Aus, Andreas Huber, Theo H van der Kwast, Ron H N van Schaik, Harry J de Koning, Sue M Moss, and Anssi Auvinen. Screening and prostate cancer mortality: results of the european randomised study of screening for prostate cancer (erspc) at 13 years of follow-up. *The Lancet*, 384(9959):2027–2035, 2014.

[255] Ruth Etzioni, Roman Gukati, Leslie Mallinger, and Jeanne Mandelblatt. Influence of study features and methods on overdiagnosis estimates in breast and prostate cancer screening. *Annals of Internal Medicine*, 158(11):831–838, 2013.

[256] Anna Bill-Axelson, Lars Holmberg, Hans Garmo, Kimmo Taari, Christer Busch, Stig Nordling, Michael Häggman, Swen-Olof Andersson, Ove Andrén, Gunnar Steineck, Hans-Olov Adami, and Jan-Erik Johansson. Radical prostatectomy or watchful waiting in prostate cancer — 29-year follow-up. *New England Journal of Medicine*, 379(24):2319–2329, 2018.

[257] Xiaozeng Lin, Anil Kapoor, Yan Gu, Mathilda Jing Chow, Hui Xu, Pierre Major, and Damu Tang. Assessment of biochemical recurrence of prostate cancer (review). *International Journal of Oncology*, 55(6):1194–1212, 2019.

[258] Stephen J. Freedland, Elizabeth B. Humphreys, Leslie A. Mangold, Mario Eisenberger, and Alan W. Partin. Time to prostate specific antigen recurrence after radical prostatectomy and risk of prostate cancer specific mortality. *The Journal of Urology*, 176(4):1404–1408, 2006.

[259] Eva Budinska, Vlad Popovici, Sabine Tejpar, Giovanni D'Ario, Nicolas Lapique, Katarzyna Otylia Sikora, Antonio Fabio Di Narzo, et al. Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. *The Journal Of Pathology*, 231(1):63–76, 2013.

[260] E Melo de Sousa, X Wang, M Jansen, E Fessler, A trinh, L P de Rooij, J H de Jong, et al. Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nature Medicine*, 19(5):614–618, 2013.

[261] Beatriz Perez Villamil, Alejandro Romera Lopez, Susana Hernandez Prieto, Guillermo Lopez Campos, Antonio Calles, Jose Antonio Lopez Asenjo, Julian Sanz Ortega, Cristina Fernandez Perez, Javier Sastre, Rosario Alfonso, Trinidad Caldes, Fernando Martin Sanchez, and Eduardo Diaz Rubio. Colon cancer molecular subtypes identified by expression profiling and associatedto stroma, mucinous type and different clinical behavior. *BMC Cancer*, 12(1):260, 2012.

[262] Paul Roepman, Andreas Schlicker, Josep Tabernero, Ian Majewski, Sun Tian, Victor Moreno, Mireille H Snel, et al. Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition. *International Journal Of Cancer*, 134(3):552–556, 2014.

[263] Anguraj Sadanandam, Costas A Lyssiotis, Krisztian Homicsko, Eric A Collisson, William J Gibb, Stephan Wullschleger, and Liliane C Gonzalez Ostos. A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nature Medicine*, 19(5):619–625, 2013.

[264] Andreas Schlicker, Garry Beran, Christine M. Chresta, Gael McWalter, Alison Pritchard, Susie Weston, Sarah Runswick, Sara Davenport, Kerry Heathcote, Denis Alferez Castro, George Orphanides, Tim French, and Lodewyk FA Wessels. Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines. *BMC Medical Genomics*, 5(1):66, 2012.

[265] Shea P Connell, Eve O'Reilly, Alexandra Tuzova, Martyn Webb, Rachel Hurst, Robert Mills, Fang Zhao, Bharati Bapat, Colin S Cooper, Antoinette S Perry, Jeremy Clark, and Daniel S Brewer. Development of a multivariable risk model integrating urinary

cell DNA methylation and cell-free RNA data for the detection of significant prostate cancer. *The Prostate*, 80(7):547–558, 2020.

[266] K. Allanach, M. Mengel, G. Einecke, B. Sis, L. G. Hidalgo, T Mueller, and P. F. Halloran. Comparing microarray versus rt-pcr assessment of renal allograft biopsies: Similar performance despite different dynamic ranges. *American Journal of Transplantation*, 8(5):1006–1015, 2008.

[267] Wiguins Etienne, Martha H. Meyer, Johnny Peppers, and Ralph A. Meyer. Comparison of mrna gene expression by rt-pcr and dna microarray. *BioTechniques*, 36(4):618–626, 2004.

[268] Katharina König, Martin Peifer, Jana Fassunke, Michaela A. Ihle, Helen Künstlinger, Carina Heydt, Katrin Stamm, Frank Ueckeroth Claudia Vollbrecht, Marc Bos, Masyar Gardizi, Matthias Scheffler, Lucia Nogova, Frauke Leenders, Kerstin Albus, Lydia Meder, Kerstin Becker, Alexandra Florin, Ursula Rommerscheidt-Fuss, Janine Altmüller, Michael Kloth, Peter Nürnberg, Thomas Henkel, Sven-Ernö Bikár, Martin L. Sos, William J. Geese, Lewis Strauss, Yon-Dschun Ko, Ulrich Gerigk, Margarete Odenthal, Thomas Zander, Jürgen Wolf, Sabine Merkelbach-Bruse, Reinhard Buettner, and Lukas C. Heukamp. Implementation of amplicon parallel sequencing leads to improvement of diagnosis and therapy of lung cancer patients. *Journal of Thoracic Oncology*, 10(7):1049–1057, 2015.

[269] Gowri Raman, Byron Wallace, Kamal Patel, Joseph Lau, and Thomas A. Trikalinos. *Update on Horizon Scans of Genetic Tests Currently Available for Clinical Use in Cancers*, chapter Appendix A, pages One–page summaries of the genetic tests for cancers. The name of the publisher, 2011.

[270] UKAS. Molecular testing strategies for Lynch syndrome in people with colorectal cancer. https://www.ukas.com/services/accreditation-services/medical-laboratory-accreditation-iso-15189/, 2020. Accessed: June 2020.

[271] Yiming Ying, Peng Li, and Colin Campbell. A marginalized variational bayesian approach to the analysis of array data. In *Proceedings of the Machine Learning in Systems Biology 2007 workshop*, 2008.

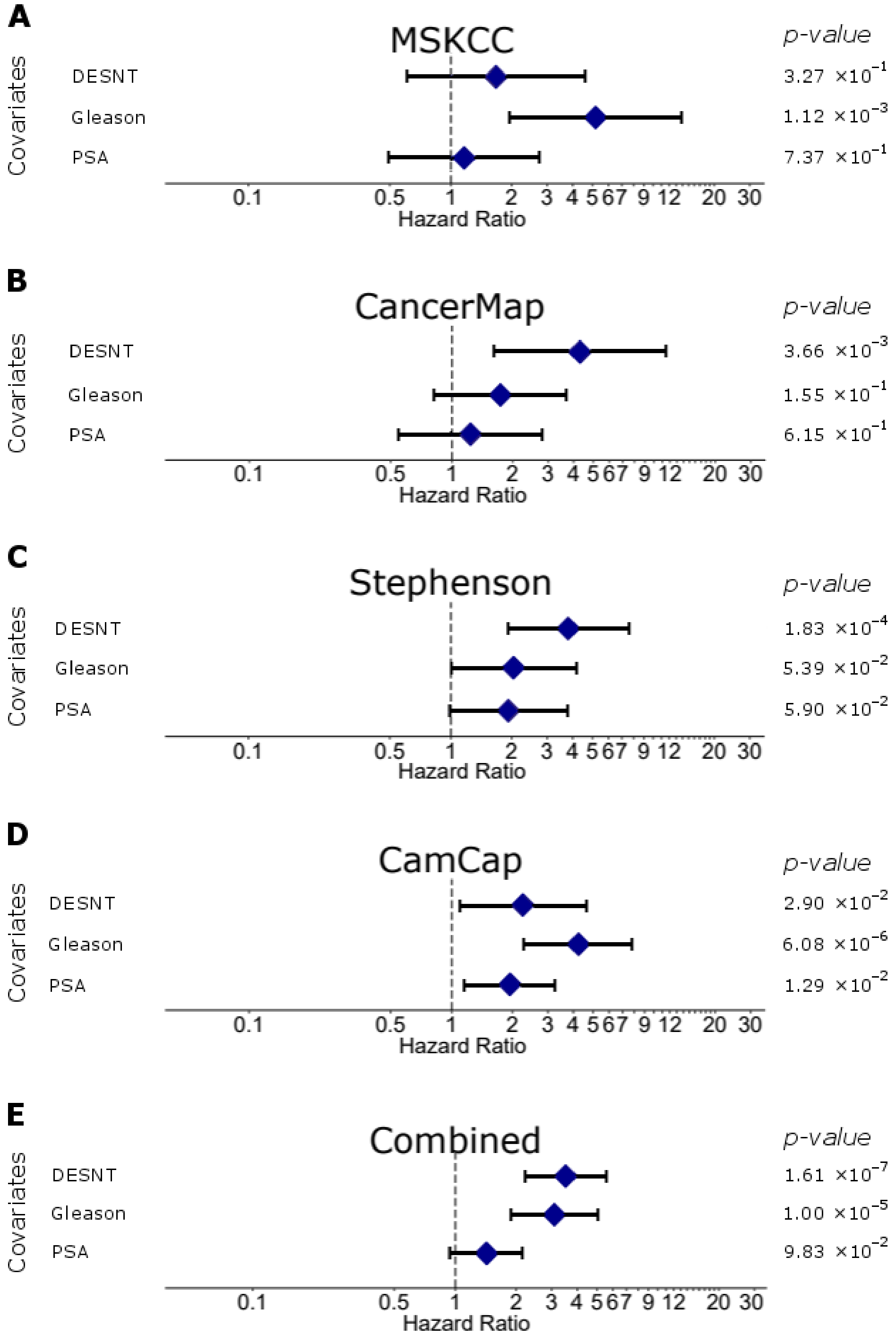
[272] Tomonari Masada, Tsuyoshi Hamada, Yuichiro Shibata, and Kiyoshi Oguri. Bayesian multi-topic microarray analysis with hyperparameter reestimation. In *Proceedings of the 5th International Conference on Advanced Data Mining and Applications*, ADMA '09, pages 253–264, Berlin, Heidelberg, 2009. Springer-Verlag.

Appendix A

Appendix A

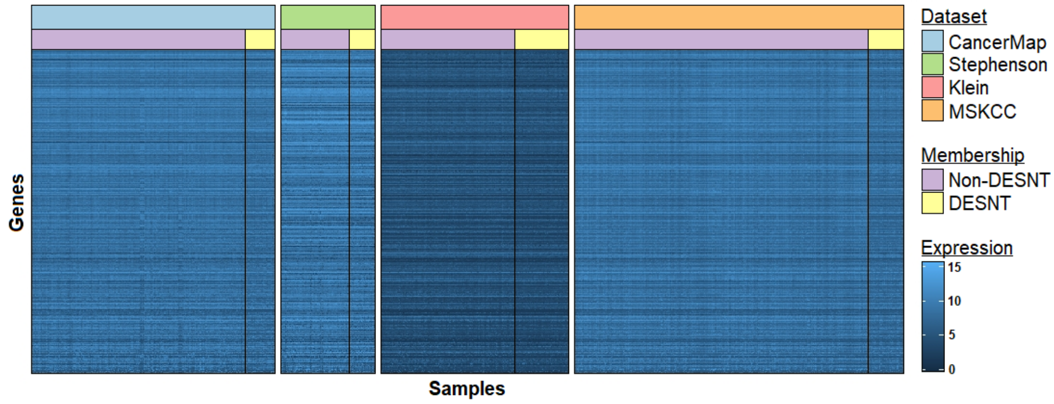
A.1 Discrete Multivariate Cox PH Models

Fig. A.1 Results from the multivariate Cox PH models, using the MSKCC (A), CancerMap (B), CamCap (C), Stephenson (D) datasets and a combination of the previous four datasets (E). The blue markers denote the hazard ratio for each covariate and the extended bars denote the 95% confidence interval. The log-rank p-value for each covariates' hazard ratio is listed on the right side of the figure. PSA level was split on $\leq / > 10$, Gleason score was split on $\leq / > 7$ and DESNT was split on non-DESNT/DESNT membership.



A.2 Gene Expression Levels

Fig. A.2 A heatmap depicting the gene expression levels of the 500 genes, used in the LPD classification process, for the CancerMap, Stephenson, Klein and MSKCC datasets.



A.3 DESNT Over-Represented Pathways

Table A.1 Top 20 GO pathways over-represented in the DESNT signature.

Pathway ID	Description	GeneRatio	*p*-value	*p*-adjusted
GO:0009611	response to wounding	19/44	3.2×10^{-13}	6.33×10^{-10}
GO:0003012	muscle system process	13/44	9.3×10^{-13}	9.2×10^{-10}
GO:0006936	muscle contraction	12/44	2.2×10^{-12}	1.45×10^{-9}
GO:0042060	wound healing	15/44	4.16×10^{-11}	2.06×10^{-8}
GO:0030029	actin filament-based process	13/44	5.31×10^{-10}	1.92×10^{-7}
GO:0009653	anatomical structure morphogenesis	24/44	5.81×10^{-10}	1.92×10^{-7}
GO:0048856	anatomical structure development	31/44	2.04×10^{-9}	5.43×10^{-7}
GO:0034329	cell junction assembly	9/44	2.33×10^{-9}	5.43×10^{-7}
GO:0030036	actin cytoskeleton organization	12/44	2.47×10^{-9}	5.43×10^{-7}

GO:0044707	single-multicellular organism process	35/44	3.36×10^{-9}	6.53×10^{-7}
GO:0007596	blood coagulation	12/44	3.88×10^{-9}	6.53×10^{-7}
GO:0007599	hemostasis	12/44	4.29×10^{-9}	6.53×10^{-7}
GO:0050817	coagulation	12/44	4.29×10^{-9}	6.53×10^{-7}
GO:0050878	regulation of body fluid levels	13/44	5.24×10^{-9}	7.41×10^{-7}
GO:0034330	cell junction organization	9/44	6.24×10^{-9}	8.24×10^{-7}
GO:0032501	multicellular organismal process	35/44	1.01×10^{-8}	1.23×10^{-6}
GO:0048468	cell development	20/44	1.05×10^{-8}	1.23×10^{-6}
GO:0032989	cellular component morphogenesis	17/44	1.87×10^{-8}	2.06×10^{-6}
GO:0003008	system process	19/44	2.29×10^{-8}	2.39×10^{-6}
GO:0031589	cell-substrate adhesion	9/44	2.65×10^{-8}	2.63×10^{-6}

Table A.2 KEGG pathways over-represented in the DESNT signature.

Pathway ID	**Description**	**GeneRatio**	***p*-value**	***p*-adjusted**
hsa04270	Vascular smooth muscle contraction	6/26	3.86×10^{-6}	1.99×10^{-4}
hsa04510	Focal adhesion	7/26	6.13×10^{-6}	1.99×10^{-4}
hsa04520	Adherens junction		$4/261.43 \times 10^{-4}$	3.11×10^{-3}
hsa04670	Leukocyte transendothelial migration	4/26	8.56×10^{-4}	1.28×10^{-2}
hsa04810	Regulation of actin cytoskeleton	5/26	9.85×10^{-4}	1.28×10^{-2}

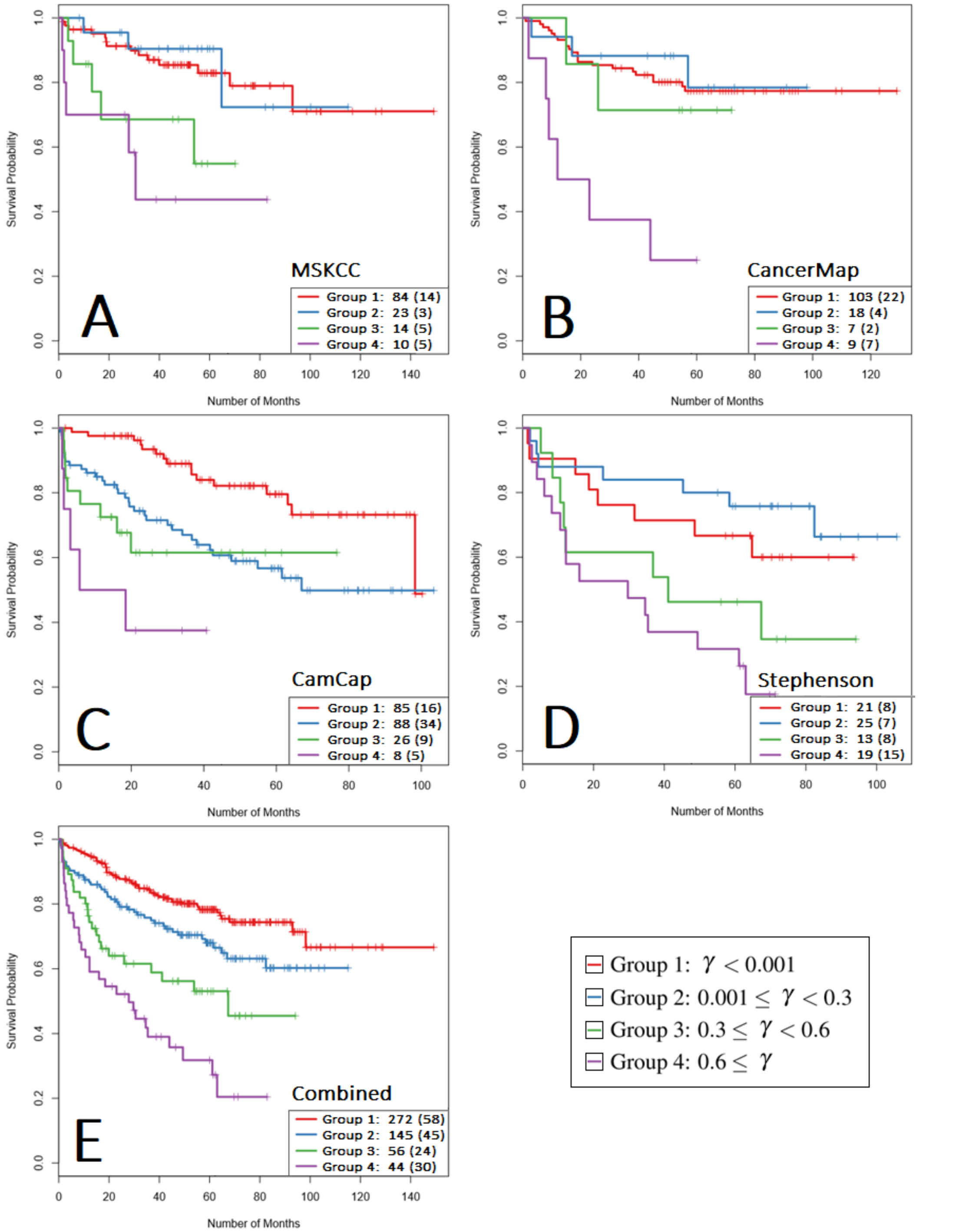
hsa05100	Bacterial invasion of epithelial cells	3/26	2.86×10^{-3}	2.78×10^{-2}
hsa04022	cGMP-PKG signaling pathway	4/26	3.08×10^{-3}	2.78×10^{-2}
hsa05410	Hypertrophic cardiomyopathy (HCM)	3/26	3.42×10^{-3}	2.78×10^{-2}
hsa05414	Dilated cardiomyopathy	3/26	4.3×10^{-3}	3.1×10^{-2}

Table A.3 Reactome pathways over-represented in the DESNT signature.

Pathway ID	**Description**	**GeneRatio**	***p*-value**	***p*-adjusted**
445355	Smooth Muscle Contraction	10/28	4.67×10^{-18}	4.2×10^{-16}
397014	Muscle contraction	10/28	3.24×10^{-14}	1.46×10^{-12}
446353	Cell-extracellular matrix interactions	4/28	8.8×10^{-7}	2.64×10^{-5}
5627123	RHO GTPases activate PAKs	3/28	1.07×10^{-4}	2.41×10^{-3}
446728	Cell junction organization	4/28	2.6×10^{-4}	4.46×10^{-3}
114608	Platelet degranulation	4/28	3.16×10^{-4}	4.46×10^{-3}
109582	Hemostasis	8/28	3.47×10^{-4}	4.46×10^{-3}
76005	Response to elevated platelet cytosolic Ca2+	4/28	3.98×10^{-4}	4.48×10^{-3}
1500931	Cell-Cell communication	4/28	1.63×10^{-3}	1.63×10^{-2}

A.4 Discretised Proportional DESNT Assignment Kaplan-Meier Curves

Fig. A.3 Kaplan-Meier survival curves comparing the discretised DESNT γ groups for the MSKCC (A), CancerMap (B), CamCap (C), Stephenson (D) and merged dataset (E), using BCR failure as the event. The number of cancer samples in each group is indicated at the bottom right corner, alongside the number of BCR failures in parentheses.



Appendix B

Appendix B

B.1 LPD Models Normalised Without TCGA Samples

Fig. B.1 Figure depicting the LPD γ values (association between a sample and a process) for each LPD process in the GSE14333plus representative run when normalising the data without TCGA samples. Samples have been coloured by their Dukes' stage assignment.

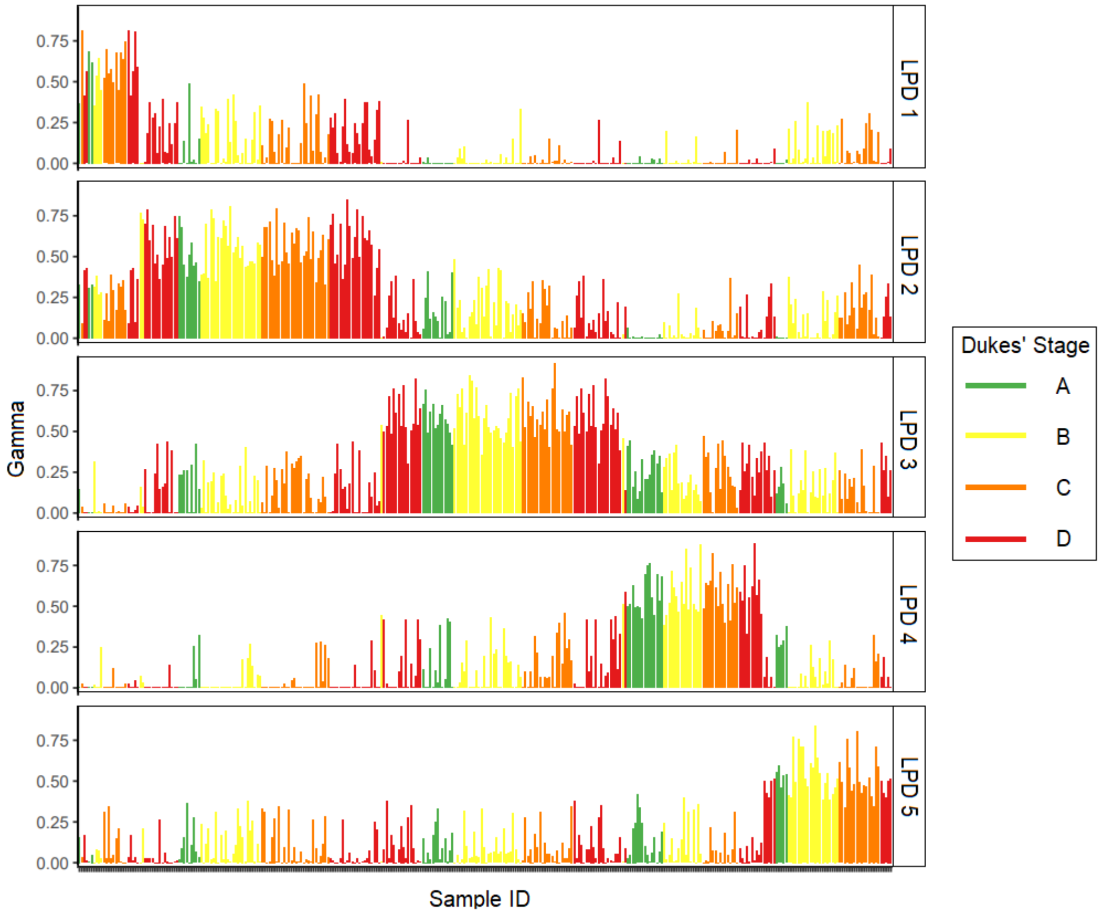


Fig. B.2 Figure depicting the LPD γ values (association between a sample and a process) for each LPD process in the GSE39582 representative run when normalising the data without TCGA samples. Samples have been coloured by their Dukes' stage assignment.

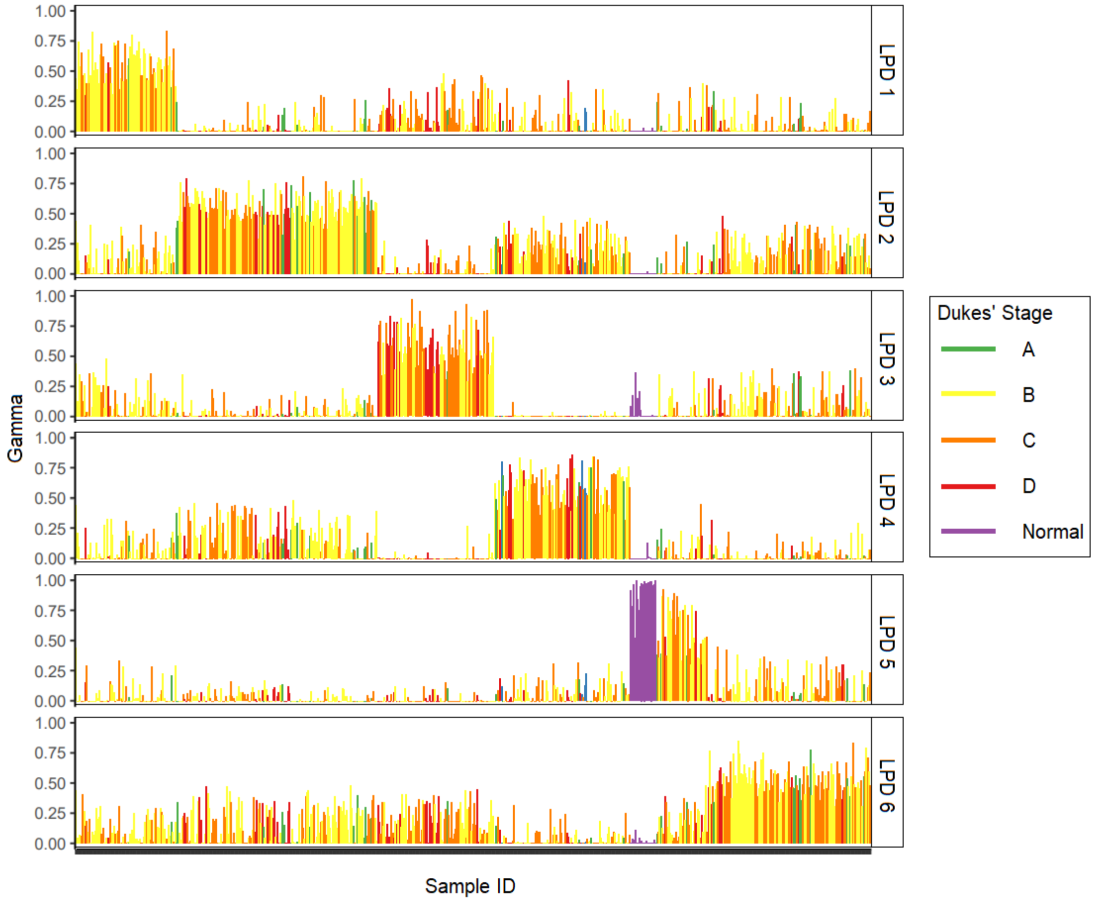


Fig. B.3 Figure depicting the LPD γ values (association between a sample and a process) for each LPD process in the GSE41258 representative run when normalising the data without TCGA samples. Samples have been coloured by their Dukes' stage assignment.

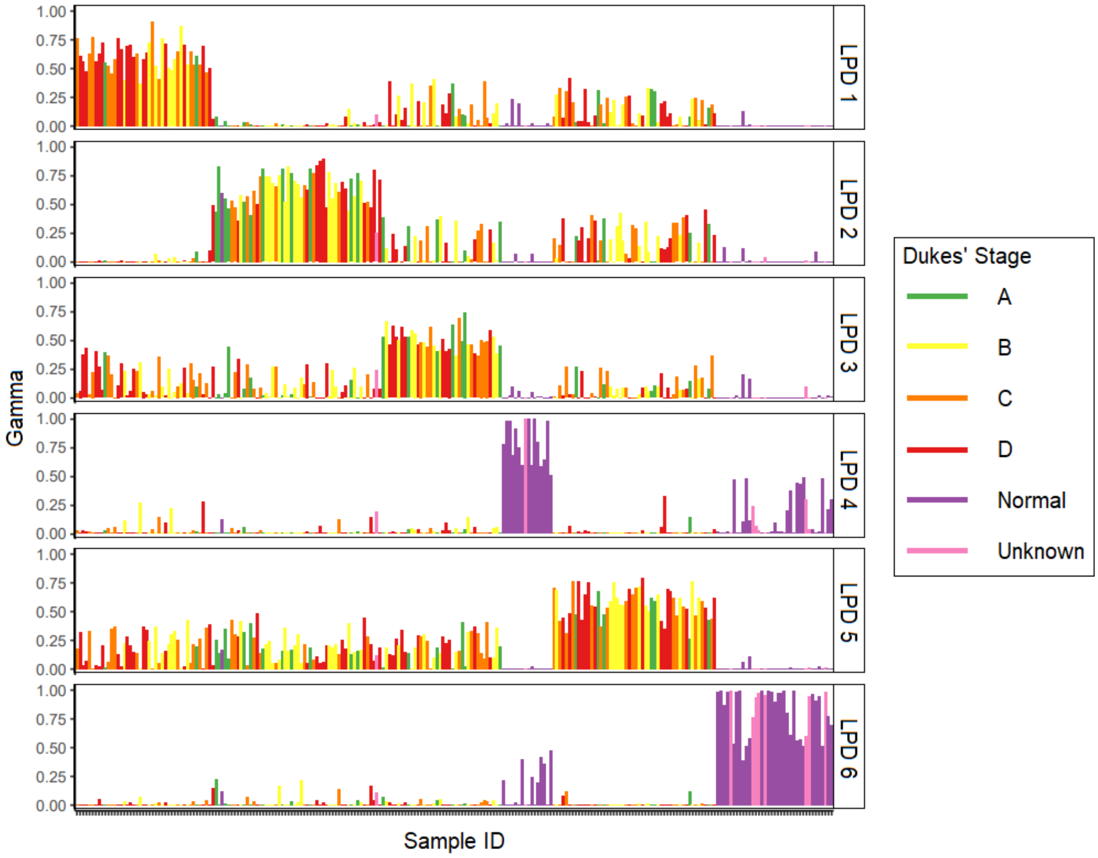
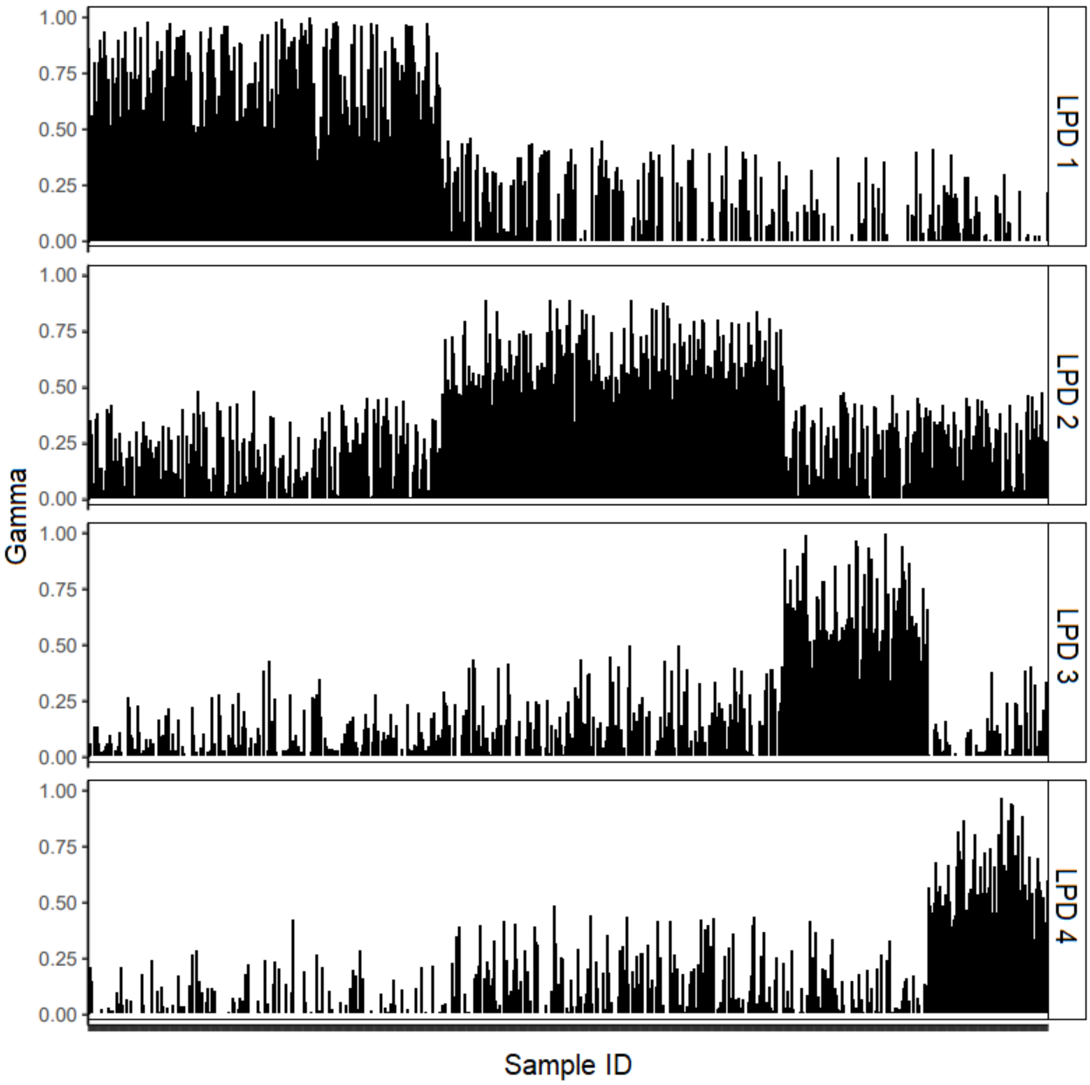


Fig. B.4 Figure depicting the LPD γ values (association between a sample and a process) for each LPD process in the GSE81653 representative run when normalising the data without TCGA samples.



B.2 Colorectal Cancer LPD Densities

Fig. B.5 Figure depicting the identification of a representative LPD run, based on the density of p-values from a set of 100 log-rank tests, each performed on an individual LPD model using the GSE14333plus dataset. The model with the shortest p-value distance to the modal density was selected as the representative LPD run.

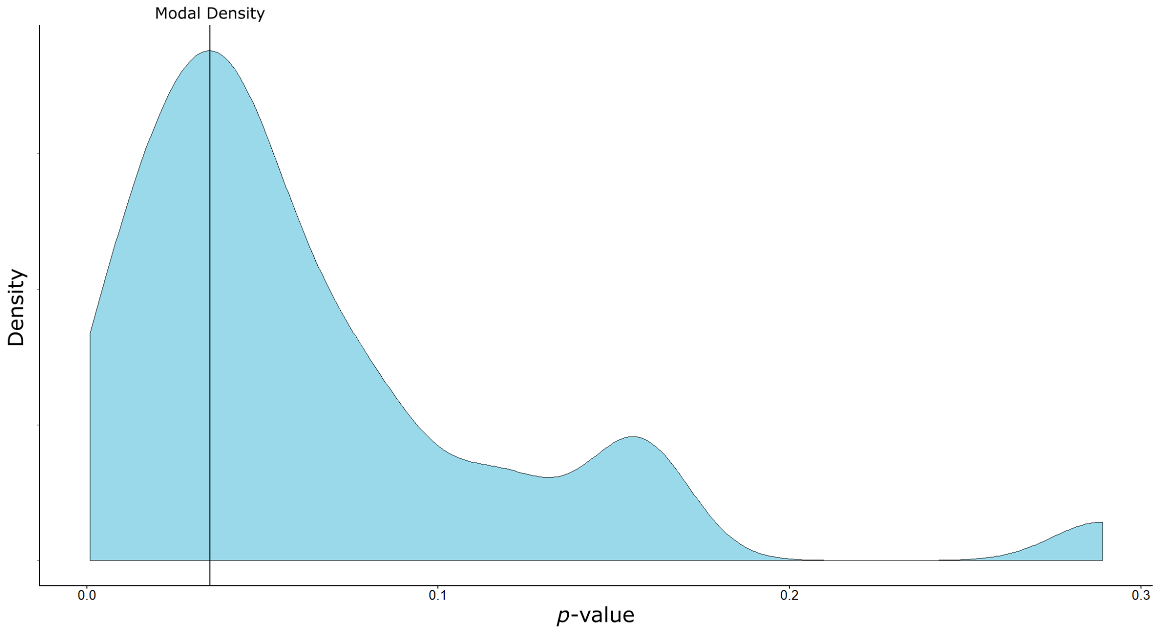


Fig. B.6 Figure depicting the identification of a representative LPD run, based on the density of p-values from a set of 100 log-rank tests, each performed on an individual LPD model using the GSE41258 dataset. The model with the shortest p-value distance to the modal density was selected as the representative LPD run.

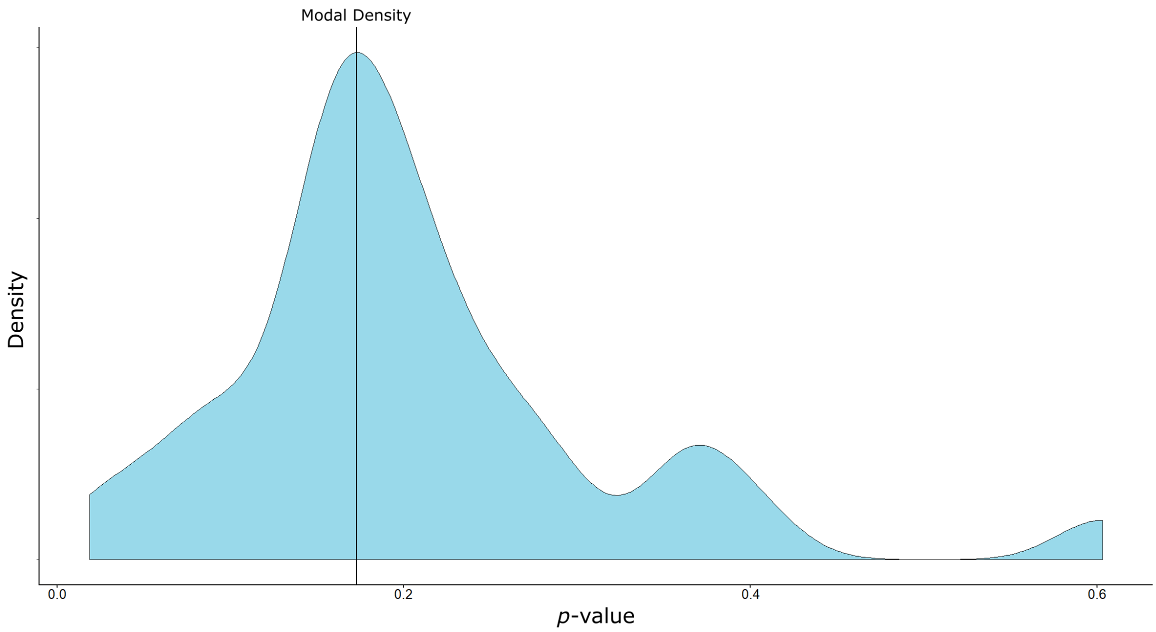
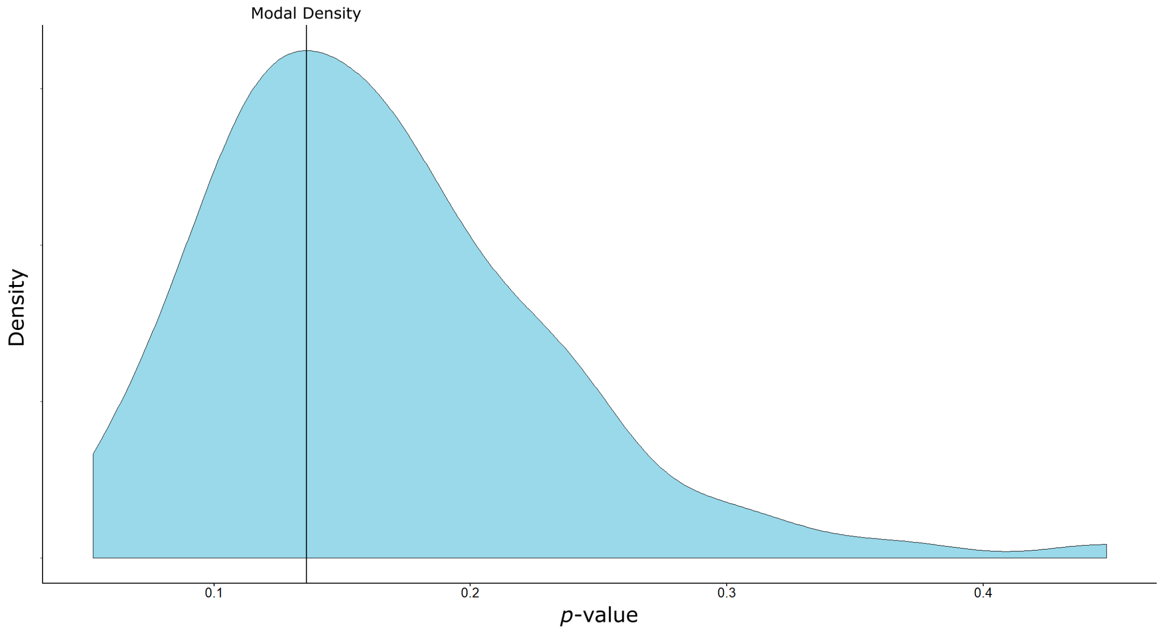


Fig. B.7 Figure depicting the identification of a representative LPD run, based on the density of *p*-values from a set of 100 log-rank tests, each performed on an individual LPD model using the GSE81653 dataset. The model with the shortest *p*-value distance to the modal density was selected as the representative LPD run.



B.3 Colorectal Cancer Representative LPD Models: Gamma Barplots

Fig. B.8 Barplot showing the γ values (Bayesian association) for each sample with each LPD process, in the LPD model built using the GSE14333plus dataset.

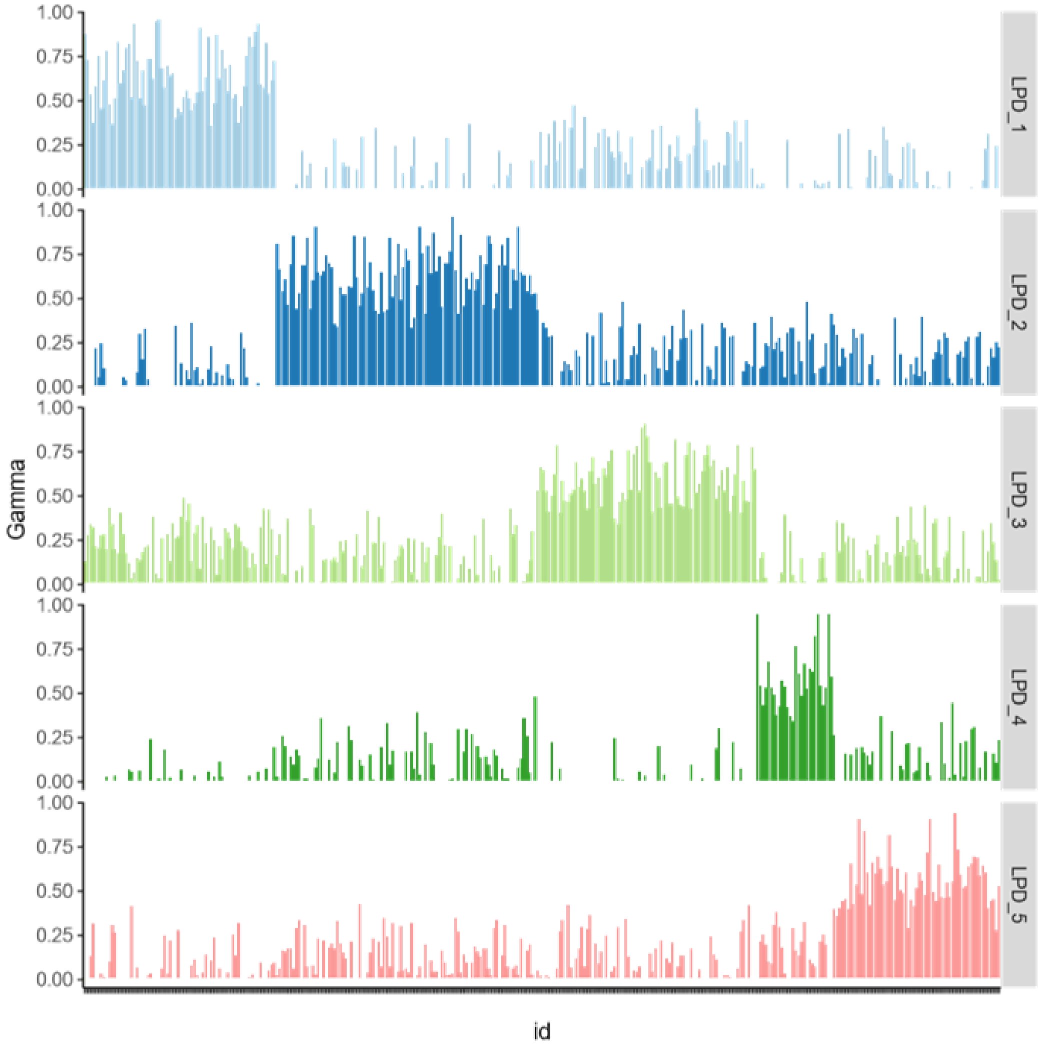


Fig. B.9 Barplot showing the γ values (Bayesian association) for each sample with each LPD process, in the LPD model built using the GSE39582 dataset.

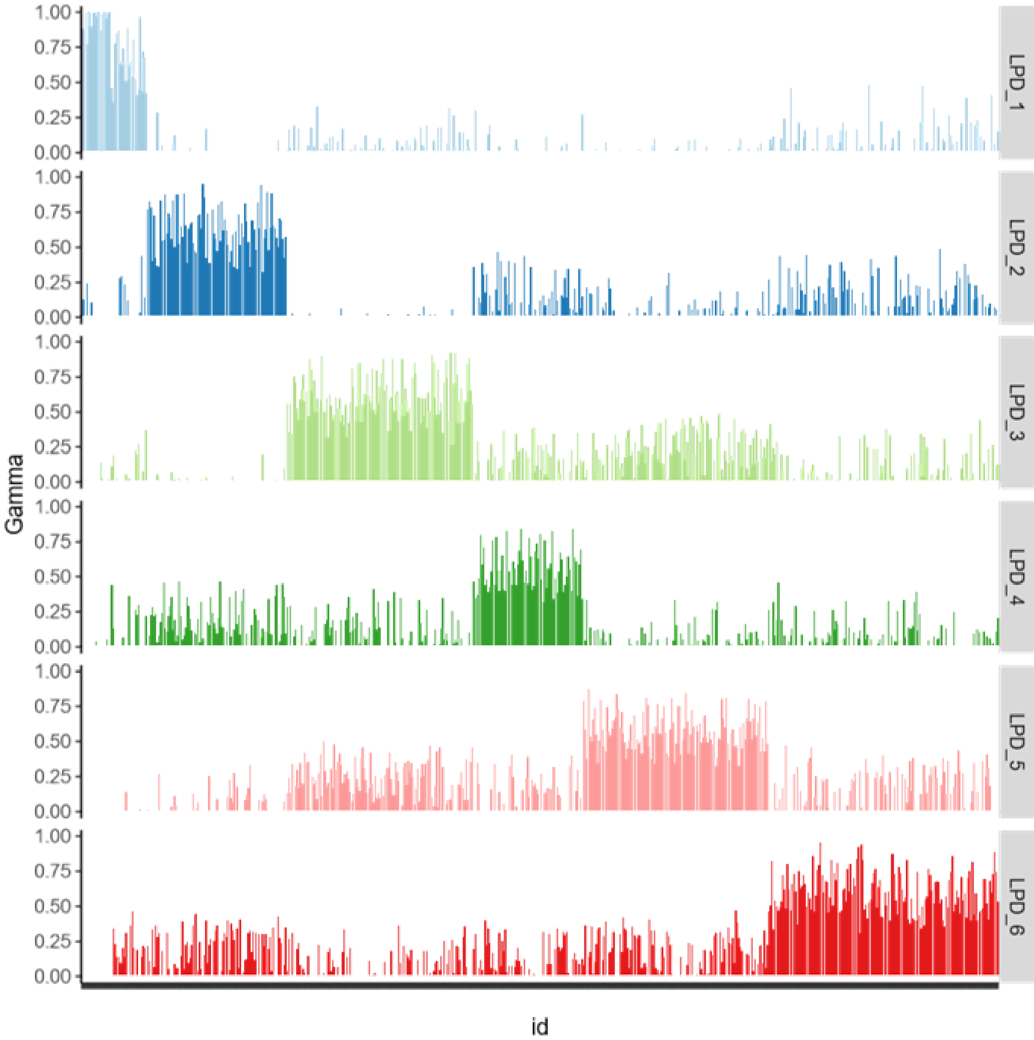


Fig. B.10 Barplot showing the γ values (Bayesian association) for each sample with each LPD process, in the LPD model built using the GSE41258 dataset.

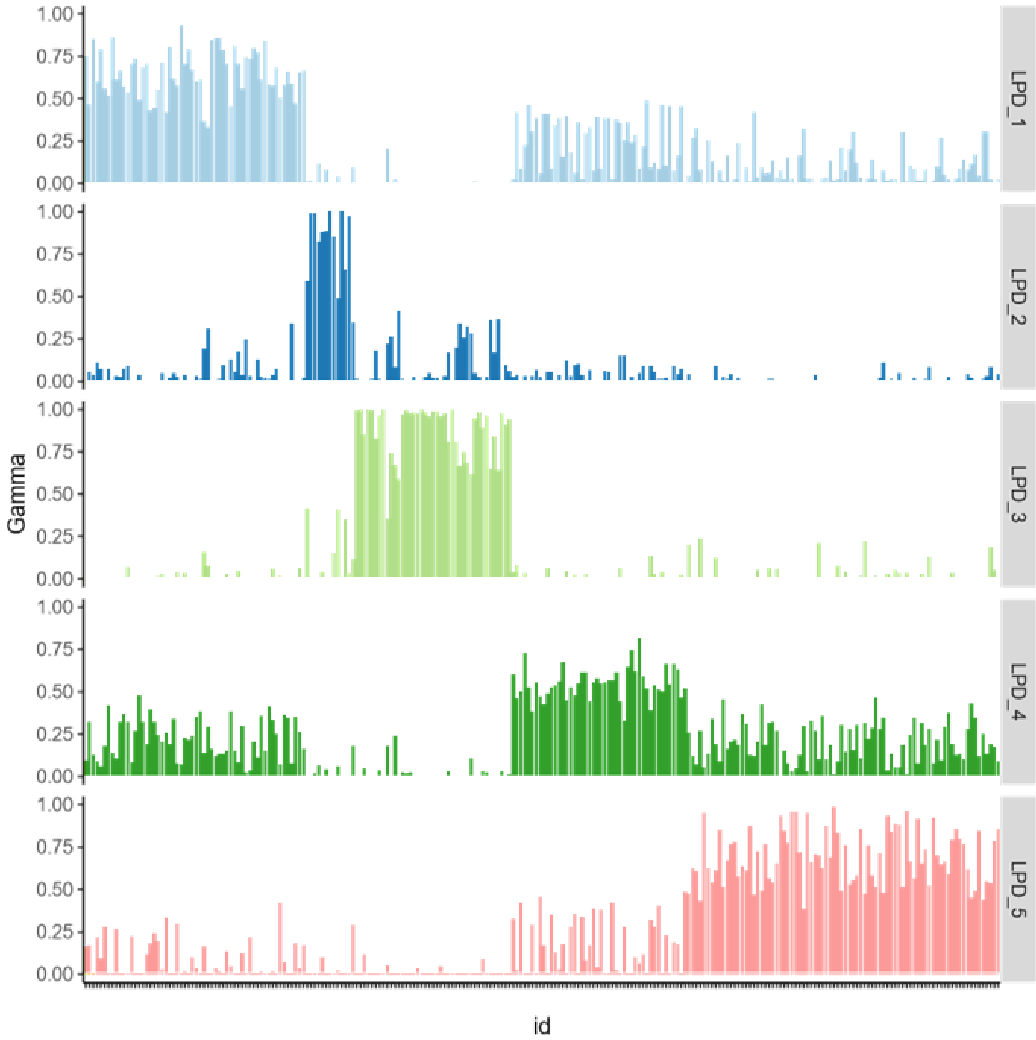
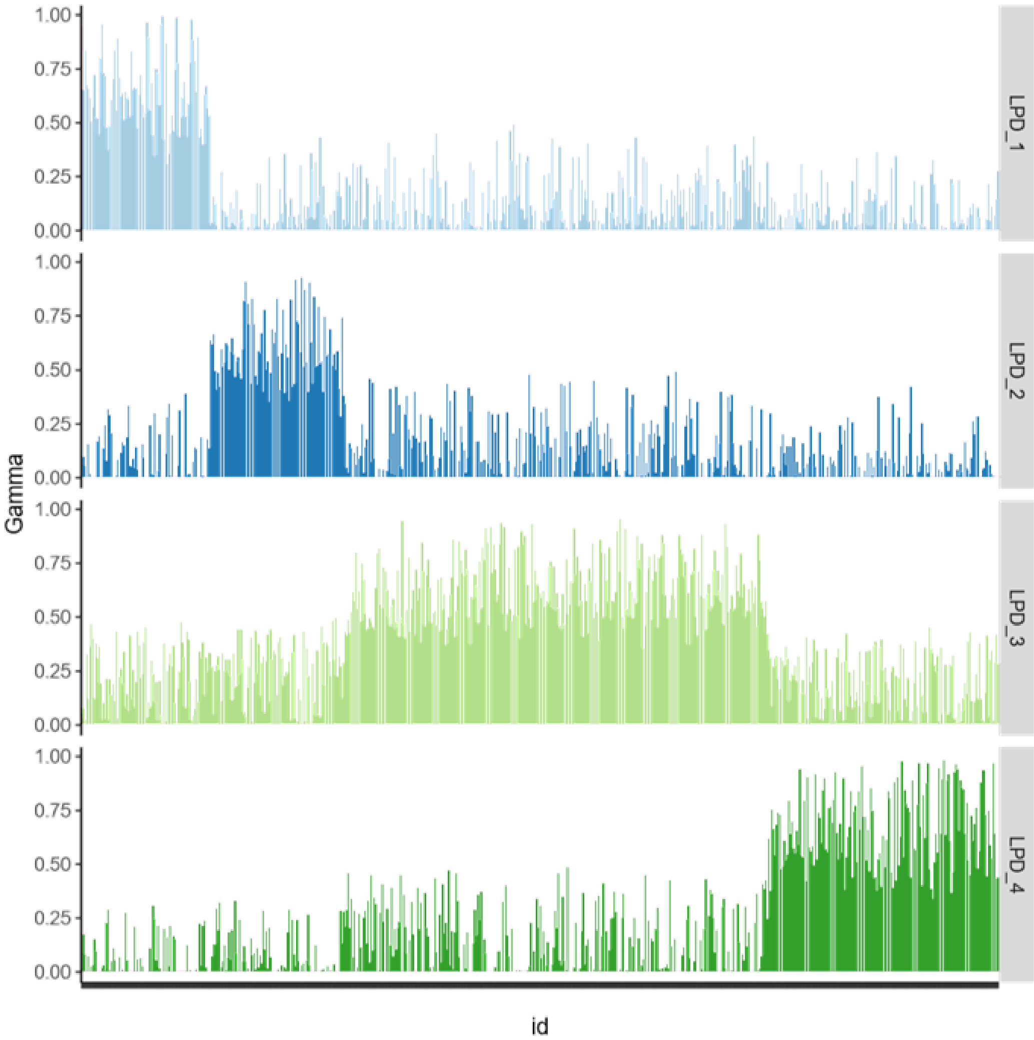


Fig. B.11 Barplot showing the γ values (Bayesian association) for each sample with each LPD process, in the LPD model built using the GSE81653 dataset.



B.4 Colorectal Cancer Representative LPD Models: Kaplain Meier Plots

Fig. B.12 Kaplan-Meier survival curves showing the disease-free survival of the the six LPD processes from the representative run for the GSE14333plus dataset. The total number of samples (with DFS information) for each process is shown in the bottom right, with the number of DFS events displayed in brackets.

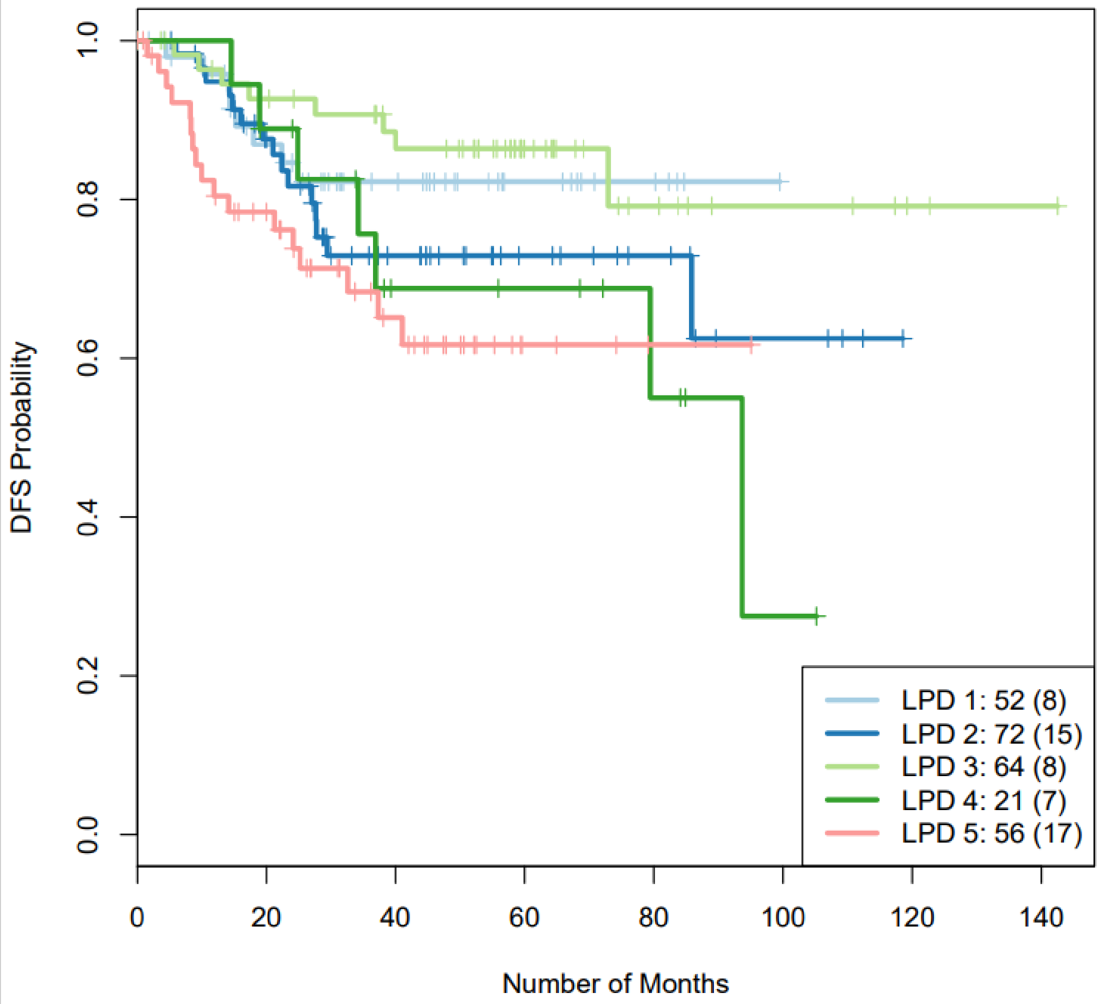


Fig. B.13 Kaplan-Meier survival curves showing the disease-free survival of the the six LPD processes from the representative run for the GSE39582 dataset. The total number of samples (with DFS information) for each process is shown in the bottom right, with the number of DFS events displayed in brackets.

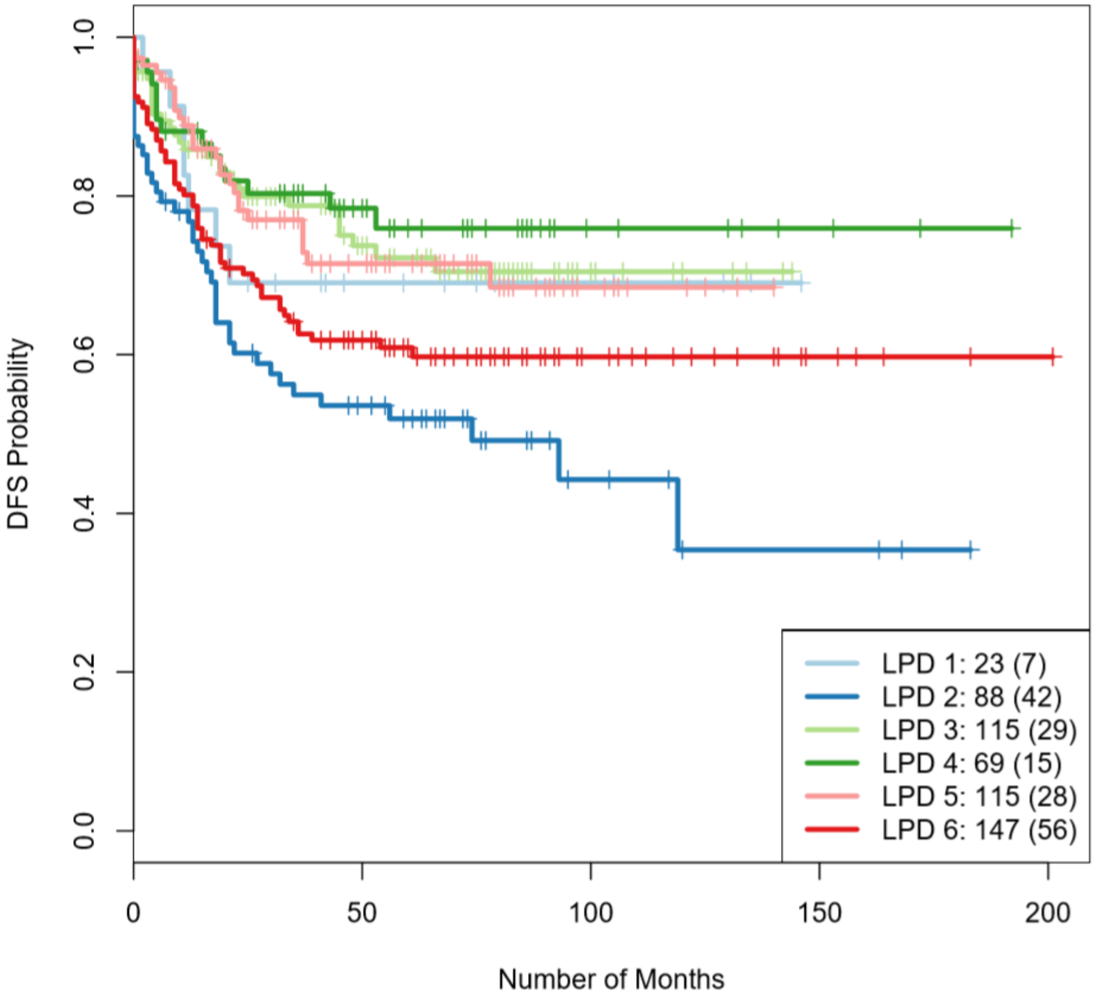
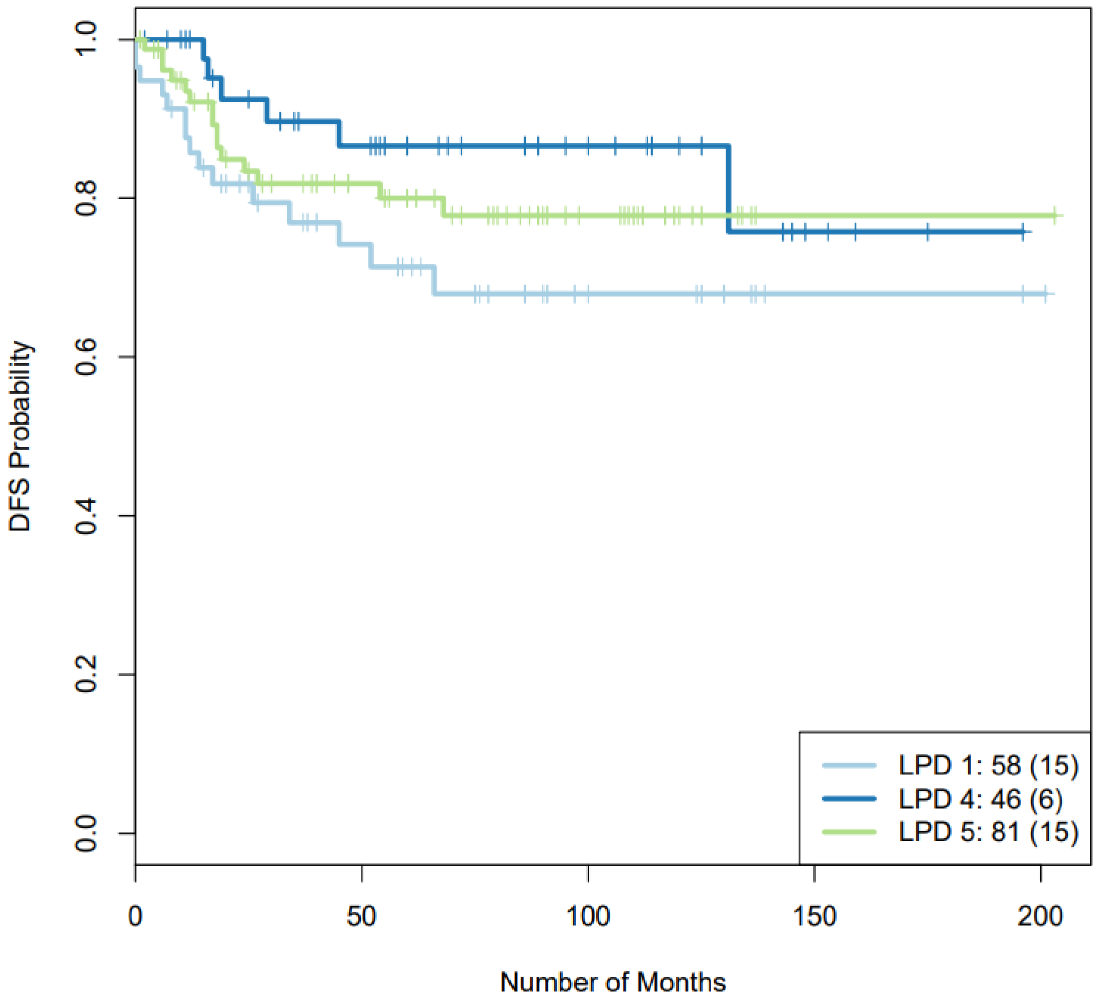


Fig. B.14 Kaplan-Meier survival curves showing the disease-free survival of the the six LPD processes from the representative run for the GSE41258 dataset. The total number of samples (with DFS information) for each process is shown in the bottom right, with the number of DFS events displayed in brackets. LPD2 and LPD 3 are not shown on the figure as they only contain normal samples.



B.5 Colorectal Cancer Differentially Expressed Genes

B.5.1 LPD A DEGs

Table B.1 Differentially expressed genes within the colorectal cancer LPD A subtype

ACSM3	CAPN5	EIF2AK3	MOGAT2	SLC35A3

ADAMDEC1	CASP5	EPHX2	MS4A12	SLC44A4
ADAMTS2	CASP7	ETHE1	MUC13	SLC4A4
ADH1C	CEACAM7	FABP2	MUC2	SLC9A2
ADTRP	CES3	FCGBP	MUC4	SUCLG2
AKR1B10	CHPF	GALNT12	NEDD4L	TIMP1
AMPD1	CHST5	GALNT7	NOX4	TJP3
ATP2A3	CLCA1	GBA3	NR3C2	TMPRSS2
BCAS1	CLCA4	GUCA2B	PADI2	TPSG1
BGN	CLDN7	HADH	PARM1	TRPM4
BMP5	CLIC5	HHLA2	PIGR	TSPAN1
BTNL8	CLINT1	HSD17B2	PKP2	UNC13B
C4orf19	CLMN	INHBA	PTGER4	XDH
CA12	CNNM4	ITM2C	RETSAT	ZG16
CA2	COL10A1	KDELC1	SCNN1B	
CA4	COL11A1	KIF26B	SCP2	
CA7	COMP	LIMA1	SERPINH1	
CAMK1D	CPT2	LRRC15	SI	

B.5.2 LPD B DEGs

Table B.2 Differentially expressed genes within the colorectal cancer LPD B subtype

AATF	CSF1R	HLA.DPA1	MORF4L1	SH3BP5
ACKR1	CSF2RB	HLA.DPB1	MRC1	SIGLEC1
ACP2	CSRP1	HLA.DQB1	MS4A4A	SIRT5
ACR	CTSL	HMOX1	MTPAP	SLA
ACTR3B	CXCL12	HMX1	MYO1F	SLC15A3
ADAMTS1	CYB5R3	HNRNPU	NAGK	SLC2A5

ADH1B	CYBRD1	HOOK1	NAP1L3	SLC31A2
AGMAT	CYP1B1	HOXB2	NCKAP1L	SLC7A7
AHNAK	CYR61	HSD11B1	NDN	SLCO2B1
ANK2	DARS2	HTR2B	NNMT	SMAD6
ANKMY1	DCLK1	IFFO1	NR6A1	SMARCC1
APOE	DCLRE1A	IGFBP6	NRP2	SOX9
ARHGDIB	DDN	IGHG1	OAZ2	SPAG5
ARMC6	DNA2	IL10RA	OLFM1	SPON1
ASPM	DOCK6	IL6	ORC6	SPRYD7
ATAD5	DOK5	IL7R	OXLD1	SRPX
ATAT1	DPT	INE1	P2RY13	STAB1
ATP2B4	DPYSL3	INO80D	PABPC3	SUSD6
ATP6V1B2	DUSP1	INTS7	PALB2	SUV39H1
ATXN2L	DUSP5	ITGA5	PALLD	SYNE1
AXL	EFR3B	ITGB2	PDSS1	SYT12
BCL6	EHD2	ITM2A	PDZRN4	TADA2A
BEX4	EIF5B	ITPR1	PEX10	TAF1B
BLVRA	ELK3	ITPR3	PIWIL2	TCF21
BNC2	EMILIN2	JAM2	PLD3	TFDP2
BRSK2	EMP3	JRK	PLEKHO1	TGFB1
BYSL	ENO2	KCNC3	PLEKHO2	THBD
C14orf132	ENPP2	KCNMB1	PMFBP1	THEMIS2
C1orf105	EPB41L3	KIF20B	PMP22	THSD7A
C1orf109	EPHB2	KLF2	PNMA1	TINF2
C5AR1	EVI2A	KNOP1	POLA2	TMEM255A
C7	EVL	KYAT1	POLD1	TMEM45A

CACNA1D	EZH2	L3MBTL1	POLR1C	TNFAIP8
CALHM2	F13A1	LAIR1	PPP4R3B	TPBG
CAPZB	FAM129A	LARP1	PRELP	TRAC
CAV1	FBLN5	LARP4B	PRKCH	TRAF2
CAV2	FCGR2B	LHFPL2	PRLR	TRAP1
CBFA2T2	FCGR3B	LIG3	PTCD3	TRIB2
CCDC69	FEZ1	LILRB2	PTPN3	TRIM22
CCL18	FHL1	LIPE	PUS7L	TRIM24
CCR2	FLI1	LMO2	RAB13	TRPV2
CD14	FLVCR2	LMOD1	RAC2	TSC22D3
CD37	FMO2	LMTK2	RAMP3	TSPAN4
CD4	GABARAPL1	LRRC20	RASSF2	TSPYL5
CD63	GADD45B	LSP1	RCAN1	TUSC3
CD69	GAS1	LST1	RCAN2	UBAP2
CD74	GAS7	MAF	RGS2	UPP1
CD93	GBX2	MAFB	RHOG	URB1
CENPJ	GFPT2	MAOB	ROR1	USP27X
CFH	GGT5	MAP2K7	RPS6	VAMP2
CHAF1A	GIMAP6	MAP4K2	RSAD2	VCAM1
CHD7	GMFG	MAP7D1	RTN1	VEGFC
CHML	GNL3L	MAPK8IP2	S1PR1	VGLL3
CHRDL1	GPR183	MCM3AP.AS1	SAFB	VIM
CHST3	GPR21	MCOLN1	SASH1	VNN2
CILP	GPR68	MDFIC	SCARF1	VPS33A
CLC	GPX3	MEF2C	SCG2	WBP1L
CLDN5	GTF3C2	MEOX1	SELL	WDR3

CMA1	GYPC	MFAP4	SEPT-11	WIPF1
COA1	HCK	MFAP5	SETBP1	WWTR1
COL16A1	HCLS1	MGAT1	SF3B3	XPO6
COLEC12	HGH1	MITF	SFRP1	ZNF142
CPA3	HHEX	MKI67	SFTPC	ZNF443
CRCP	HIST3H2A	MLLT11	SGCE	ZNF473
CRIP2	HLA.DMA	MMRN1	SGK1	ZNF629
CRISPLD2	HLA.DMB	MNX1	SH2B3	ZSWIM1

B.5.3 LPD C DEGs

Table B.3 Differentially expressed genes within the colorectal cancer LPD C subtype

ANO10	ENTPD5	KCNK5	SLC26A2	WARS
BTN3A3	ERO1A	PLEKHG6	SLC39A6	WDR41
CHP2	GBP1	PLXNA2	SMAP1	
CKB	GREM2	RARRES3	SNAPC1	
CLCN2	IHH	RPS6KA6	TFCP2L1	
CXCL10	JAK2	SGK2	UBE2L6	

B.5.4 Pericol DEGs

Table B.4 Differentially expressed genes within the colorectal cancer Pericol subtype

A1CF	COL4A1	GLIPR1	MSN	SFRP4
ABCA1	COL4A2	GLT8D2	MSR1	SGK2
ACOT11	COL5A1	GLUL	MXRA5	SH3BP5
ACSM3	COL5A2	GNA11	MXRA7	SIRPA
ADAM12	COL6A1	GNS	MXRA8	SLA

ADAMTS2	COL6A2	GOT2	MYH10	SLAMF8
ADAP2	COL6A3	GPA33	MYO1A	SLC1A3
ADGRF5	COL8A1	GPD1L	MYO1F	SLC22A18AS
ADGRL4	COLEC12	GPR137B	MYO5A	SLC22A5
ADTRP	COMP	GPR65	MYOF	SLC26A3
AGFG2	COPZ2	GPX7	NAGK	SLC27A2
AGMAT	COQ9	GREM1	NAT2	SLC2A3
AHR	COX4I1	GULP1	NCF2	SLC37A4
AIF1	COX5B	HADH	NDUFAF4	SLC38A2
AKAP12	CPT1A	HCK	NID2	SLC38A6
AKT3	CPTP	HCLS1	NNMT	SLC39A6
ALOX5AP	CREM	HDHD3	NOL12	SLC44A4
ANGPTL2	CRYM	HEG1	NOX4	SLC9A2
ANOS1	CSF1R	HEXA	NPL	SLFN12
ANTXR1	CSGALNACT2	HHLA2	NREP	SMARCA1
ANXA1	CTGF	HIF1A	NRP1	SMPD3
ANXA5	CTSB	HIP1	NXPE4	SNAI2
ANXA6	CTSD	HIVEP2	OLFML2B	SPARC
AP1M2	CTSK	HLA.DMA	OLR1	SPHK2
APLNR	CTSL	HLA.DPA1	OSMR	SPOCK1
APOC1	CTSO	HLA.DPB1	OVOL2	SPP1
ARHGDIB	CWH43	HLA.DRA	PAK4	SRGN
ARL4C	CXCR4	HNRNPAB	PALLD	SSH1
ASPN	CYB5R3	HOXB2	PAM	ST6GALNAC5
ASTE1	CYBB	HSD11B2	PARM1	STAP2
ATP10D	CYP1B1	HTRA1	PBLD	STC1

ATP2C2	CYP2J2	ICAM1	PCK2	STOM
ATP6V1B2	CYP4F12	ID1	PCOLCE	SUCLG1
AXL	CYR61	IFI16	PDE10A	SUCLG2
BASP1	DACT1	IGFBP3	PDE4DIP	SULF1
BCAT1	DBN1	IGFBP4	PDGFC	SULT1B1
BCL6	DCN	IGFBP5	PDGFRB	TAF6L
BDH1	DEGS1	IGFBP7	PDLIM5	TCF4
BGN	DENND5A	IL1R1	PDLIM7	TENM3
BICC1	DGAT1	ILVBL	PDSS1	TFEC
BICD1	DHRS11	IMPA2	PDZD3	TGFB1
BNIP3L	DOCK4	INHBA	PEA15	TGFB3
BTNL3	DOK4	IRAK3	PECAM1	THBS1
C1orf105	DPYSL2	ITGA5	PEX11A	THBS2
C1orf109	DPYSL3	ITGAM	PFKFB3	THEMIS2
C1orf123	DRAM1	ITGAV	PHF21A	THY1
C1orf174	DSE	ITGBL1	PIGR	TIMP1
C1QTNF1	DUSP10	KBTBD11	PILRA	TIMP2
C3	ECM2	KIF26B	PIP4K2A	TIMP3
C3AR1	EDNRA	LAMA4	PIP5K1B	TJP3
C5AR1	EFEMP1	LAMB1	PKD2	TLR1
C5orf30	EFEMP2	LAMB2	PLA2G7	TLR2
CALCRL	ELK3	LAMC1	PLEKHA6	TM6SF1
CALU	EMP3	LAMP5	PLS1	TMEM106C
CAMSAP2	ENG	LAPTM5	PLXDC2	TMEM45A
CAPN5	ENTPD1	LCP2	PLXNC1	TMPRSS2
CASP5	EOGT	LDB2	PLXND1	TNC

CC2D1A	EPB41L4B	LDLRAP1	PMP22	TNFAIP3
CCDC102B	EPN3	LGALS1	POSTN	TNFAIP6
CCDC88A	EPS8L2	LGALS4	PPFIA3	TNFSF4
CD14	EPS8L3	LILRB1	PPFIBP1	TNK1
CD163	EPYC	LMCD1	PRCP	TPST1
CD248	ERG	LOX	PRKCZ	TRIB2
CD53	ESRRA	LOXL1	PRKD1	TRIM22
CD59	ETHE1	LRRC15	PRR16	TST
CD74	EVC	LRRC19	PRRG2	TTC38
CD86	EVI2A	LRRC31	PRRX1	TTC39A
CD99	F2R	LRRC32	PSAP	TTLL12
CDH11	FAAH	LTBP1	PTGIS	TWIST1
CDH5	FAM168A	LUM	PTPRC	TWSG1
CDHR5	FAM198B	LY96	PXDN	TXNDC15
CDK14	FAM83E	MACF1	PXMP2	TYROBP
CDK17	FAP	MAFB	QKI	UGCG
CDS1	FBN1	MAN2B1	RAB31	UNC13B
CDX1	FCER1G	MAP3K8	RAB8B	UQCRC1
CEACAM7	FCGR2A	MAP4K4	RAI14	VAMP5
CEBPG	FCGR2B	MAP7	RARRES2	VCAM1
CES3	FCHSD2	MAR-1	RASGRP3	VCAN
CFI	FGFR3	MEG3	RBMS1	VIM
CHI3L1	FN1	MFAP2	RECQL	VIPR1
CHST11	FOXD2	MFGE8	RGCC	VSIG4
CHST15	FPR3	MGP	RGS1	WBP1L
CHSY1	FRAT2	MITF	RGS2	WIPF1

CLCN2	FSTL1	MLYCD	RHOQ	WISP1
CLDN7	FXYD3	MMP15	ROBO1	WSB1
CLEC2B	FYN	MMP2	SCPEP1	XDH
CLEC7A	FZD1	MN1	SDC2	ZFAND5
CLIC4	GAS1	MNDA	SEC31A	ZG16
CNN3	GBP2	MOGAT2	SELENBP1	ZNF532
CNNM4	GCDH	MOXD1	SEMA4G	ZNF576
COL11A1	GDPD2	MPST	Sept-11	ZYX
COL15A1	GEM	MRC1	SERPINE1	
COL16A1	GFPT2	MRC2	SERPINF1	
COL1A1	GIPC2	MS4A4A	SERPING1	
COL1A2	GJA1	MS4A6A	SERPINH1	

B.6 Colorectal Cancer Subtype Pathways

B.6.1 LPD A Enriched Pathways

Table B.5 GO pathways over-represented in LPD A.

Pathway ID	Description	*p*-value	Fold Enrichment
0015701	bicarbonate transport	4.83×10^{-5}	24.46
0016266	O-glycan processing	1.64×10^{-4}	17.94
0006730	one-carbon metabolic process	3.44×10^{-4}	28.70
0007588	excretion	6.44×10^{-4}	23.27
0030277	maintenance of gastrointestinal epithelium	1.33×10^{-3}	53.82
0006508	proteolysis	7.95×10^{-3}	3.44

2001225	regulation of chloride transport	9.15×10^{-3}	215.28
0051453	regulation of intracellular pH	1.18×10^{-2}	17.94
0030199	collagen fibril organization	1.38×10^{-2}	16.56
0034220	ion transmembrane transport	1.58×10^{-2}	5.13
0032849	positive regulation of cellular pH reduction	1.82×10^{-2}	107.64
0006810	transport	2.18×10^{-2}	3.71
0001501	skeletal system development	2.51×10^{-2}	6.29
0051216	cartilage development	3.00×10^{-2}	10.95
0008104	protein localization	3.19×10^{-2}	10.59
0030574	collagen catabolic process	3.48×10^{-2}	10.09
0008152	metabolic process	4.20×10^{-2}	5.13
0035725	sodium ion transmembrane transport	4.42×10^{-2}	8.85
0005975	carbohydrate metabolic process	4.58×10^{-2}	4.95
0098911	regulation of ventricular cardiac muscle cell action potential	4.93×10^{-2}	39.14

Table B.6 KEGG pathways over-represented in LPD A.

Pathway ID	**Description**	***p*-value**	**Fold Enrichment**
00910	Nitrogen metabolism	2.29×10^{-4}	31.74
04972	Pancreatic secretion	4.48×10^{-3}	7.25
04964	Proximal tubule bicarbonate reclamation	1.19×10^{-2}	17.59
04960	Aldosterone-regulated sodium reabsorption	3.23×10^{-2}	10.38

00071	Fatty acid degradation	3.71×10^{-2}	9.63
01100	Metabolic pathways	4.80×10^{-2}	1.66

Table B.7 Reactome pathways over-represented in LPD A.

Pathway ID	**Description**	***p*-value**	**Fold Enrichment**
1475029	Reversible hydration of carbon dioxide	4.22×10^{-5}	55.00
913709	O-linked glycosylation of mucins	4.00×10^{-4}	13.98
1650814	Collagen biosynthesis and modifying enzymes	7.29×10^{-3}	9.85
977068	Termination of O-glycan biosynthesis	1.03×10^{-2}	19.04
2672351	Stimuli-sensing channels	1.27×10^{-2}	8.05

Fig. B.15 Barplot of the top 20 (ordered by *p*-value) GO pathways over-represented in LPD A.

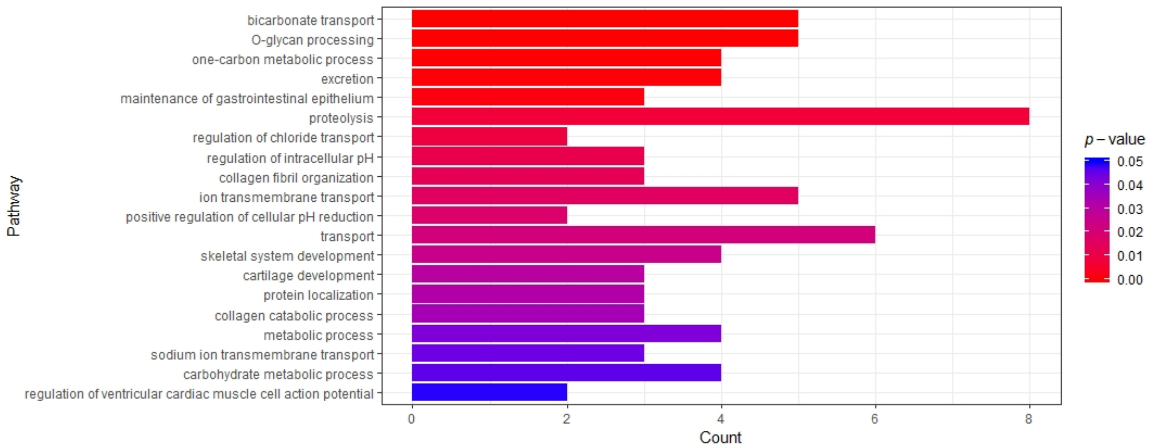


Fig. B.16 Barplot of the KEGG pathways over-represented in LPD A.

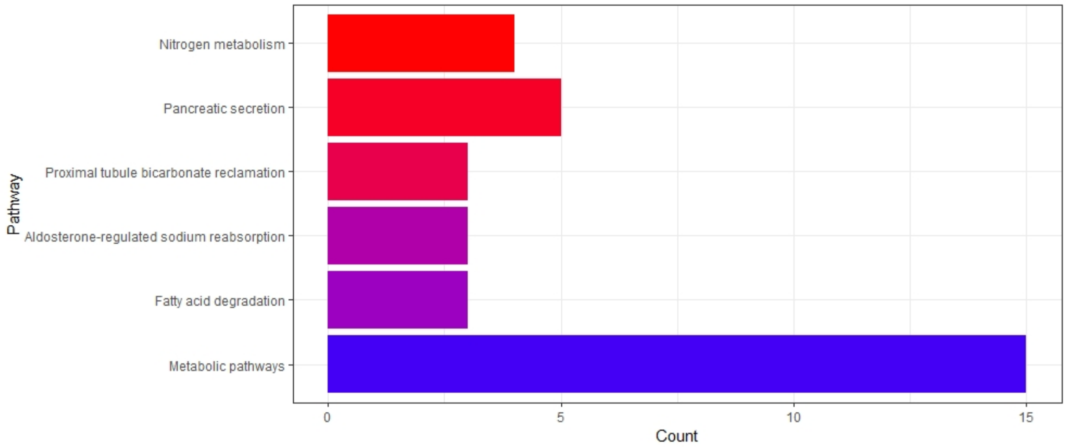
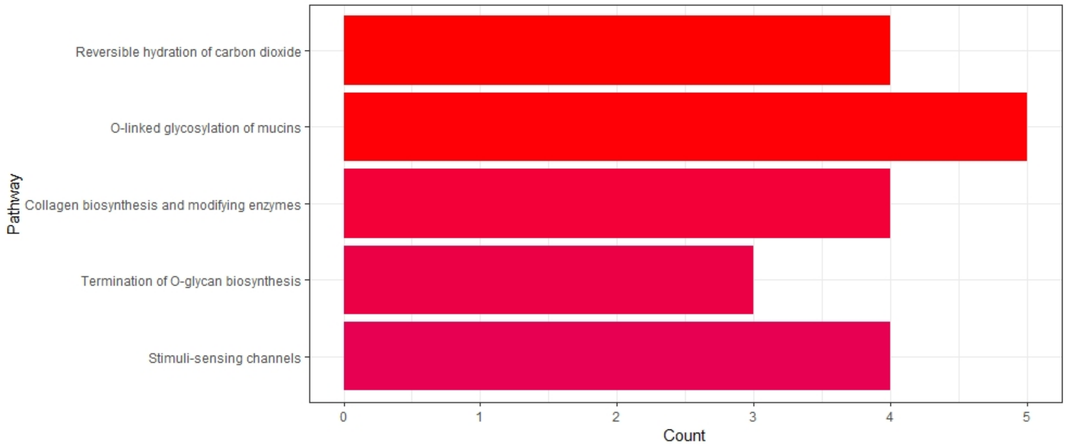


Fig. B.17 Barplot of the Reactome pathways over-represented in LPD A.



B.6.2 LPD B Enriched Pathways

Table B.8 GO pathways over-represented in LPD B.

Pathway ID	Description	*p*-value	Fold Enrichment
0007155	cell adhesion	2.08×10^{-6}	3.09
0006461	protein complex assembly	3.80×10^{-5}	5.38
0006954	inflammatory response	1.36×10^{-4}	2.84
0030279	negative regulation of ossification	1.44×10^{-4}	17.73

0006935	chemotaxis	3.04×10^{-4}	4.65
0006955	immune response	4.87×10^{-4}	2.56
0030890	positive regulation of B cell proliferation	5.71×10^{-4}	8.73
0050731	positive regulation of peptidyl-tyrosine phosphorylation	5.94×10^{-4}	5.53
0008015	blood circulation	1.11×10^{-3}	7.56
0006874	cellular calcium ion homeostasis	1.26×10^{-3}	4.88
0050900	leukocyte migration	1.41×10^{-3}	4.18
0007399	nervous system development	1.76×10^{-3}	2.77
2000249	regulation of actin cytoskeleton reorganization	3.60×10^{-3}	12.61
0001525	angiogenesis	6.20×10^{-3}	2.80
0090023	positive regulation of neutrophil chemotaxis	6.45×10^{-3}	10.31
0043065	positive regulation of apoptotic process	7.00×10^{-3}	2.46
0001938	positive regulation of endothelial cell proliferation	7.27×10^{-3}	4.93
0030198	extracellular matrix organization	7.94×10^{-3}	2.89
0007088	regulation of mitotic nuclear division	8.27×10^{-3}	9.45
0030866	cortical actin cytoskeleton organization	8.27×10^{-3}	9.45
0031100	organ regeneration	9.20×10^{-3}	6.04
0009408	response to heat	9.90×10^{-3}	5.91

0090073	positive regulation of protein homodimerization activity	1.02×10^{-2}	18.91
0008285	negative regulation of cell proliferation	1.04×10^{-2}	2.15
0071560	cellular response to transforming growth factor beta stimulus	1.06×10^{-2}	5.79
0060021	palate development	1.08×10^{-2}	4.48
0008360	regulation of cell shape	1.19×10^{-2}	3.24
0002377	immunoglobulin production	1.26×10^{-2}	17.02
0045727	positive regulation of translation	1.39×10^{-2}	5.35
0045944	positive regulation of transcription from RNA polymerase II promoter	1.40×10^{-2}	1.62
0001937	negative regulation of endothelial cell proliferation	1.40×10^{-2}	7.82
0097421	liver regeneration	1.40×10^{-2}	7.82
0043066	negative regulation of apoptotic process	1.49×10^{-2}	1.99
0045766	positive regulation of angiogenesis	1.60×10^{-2}	3.45
0030168	platelet activation	1.60×10^{-2}	3.45
0050680	negative regulation of epithelial cell proliferation	1.68×10^{-2}	5.07
0045765	regulation of angiogenesis	1.68×10^{-2}	7.32
0010628	positive regulation of gene expression	1.77×10^{-2}	2.38

0006928	movement of cell or subcellular component	1.77×10^{-2}	3.96
0070301	cellular response to hydrogen peroxide	1.78×10^{-2}	4.98
0032211	negative regulation of telomere maintenance via telomerase	1.81×10^{-2}	14.18
0001974	blood vessel remodeling	1.83×10^{-2}	7.09
0018108	peptidyl-tyrosine phosphorylation	1.86×10^{-2}	2.97
0010718	positive regulation of epithelial to mesenchymal transition	1.98×10^{-2}	6.88
0050679	positive regulation of epithelial cell proliferation	2.11×10^{-2}	4.73
0042102	positive regulation of T cell proliferation	2.11×10^{-2}	4.73
0007160	cell-matrix adhesion	2.12×10^{-2}	3.78
0009887	organ morphogenesis	2.30×10^{-2}	3.70
0048821	erythrocyte development	2.78×10^{-2}	11.35
0050930	induction of positive chemotaxis	2.78×10^{-2}	11.35
0001657	ureteric bud development	2.88×10^{-2}	5.97
0006469	negative regulation of protein kinase activity	3.04×10^{-2}	3.44
0008219	cell death	3.08×10^{-2}	5.82
0033138	positive regulation of peptidyl-serine phosphorylation	3.47×10^{-2}	4.05

0035984	cellular response to trichostatin A	3.48×10^{-2}	56.73
0060390	regulation of SMAD protein import into nucleus	3.48×10^{-2}	56.73
0032764	negative regulation of mast cell cytokine production	3.48×10^{-2}	56.73
0042981	regulation of apoptotic process	3.49×10^{-2}	2.40
0051209	release of sequestered calcium ion into cytosol	3.50×10^{-2}	5.53
0070374	positive regulation of ERK1 and ERK2 cascade	3.52×10^{-2}	2.59
0007165	signal transduction	3.58×10^{-2}	1.47
0071347	cellular response to interleukin-1	3.62×10^{-2}	4.00
0008284	positive regulation of cell proliferation	3.67×10^{-2}	1.83
0045599	negative regulation of fat cell differentiation	3.72×10^{-2}	5.40
0001764	neuron migration	3.78×10^{-2}	3.24
0050776	regulation of immune response	3.80×10^{-2}	2.55
0006325	chromatin organization	3.95×10^{-2}	5.28
0010977	negative regulation of neuron projection development	4.19×10^{-2}	5.16
0030514	negative regulation of BMP signaling pathway	4.43×10^{-2}	5.04
0030838	positive regulation of actin filament polymerization	4.43×10^{-2}	5.04

0048666	neuron development	4.68×10^{-2}	4.93
0071222	cellular response to lipopolysaccharide	4.91×10^{-2}	3.01

Table B.9 KEGG pathways over-represented in LPD B.

Pathway ID	**Description**	***p*-value**	**Fold Enrichment**
05144	Malaria	3.73×10^{-3}	5.69
04068	FoxO signaling pathway	7.77×10^{-3}	3.12
05205	Proteoglycans in cancer	1.03×10^{-2}	2.56
04640	Hematopoietic cell lineage	1.05×10^{-2}	3.74
05152	Tuberculosis	1.33×10^{-2}	2.63
04145	Phagosome	1.48×10^{-2}	2.79
04010	MAPK signaling pathway	1.93×10^{-2}	2.20
04380	Osteoclast differentiation	2.18×10^{-2}	2.84
05202	Transcriptional misregulation in cancer	2.63×10^{-2}	2.50
05150	Staphylococcus aureus infection	2.77×10^{-2}	4.30
05323	Rheumatoid arthritis	3.96×10^{-2}	3.17

Table B.10 Reactome pathways over-represented in LPD B.

Pathway ID	**Description**	***p*-value**	**Fold Enrichment**
202733	Cell surface interactions at the vascular wall	3.71×10^{-3}	5.73
2129379	Molecules associated with elastic fibres	5.92×10^{-3}	6.78

422356	Regulation of insulin secretion	2.52×10^{-2}	6.25
69166	Removal of the Flap Intermediate	2.89×10^{-2}	11.05
1566948	Elastic fibre formation	4.62×10^{-2}	8.59

Fig. B.18 Barplot of the top 20 (ordered by *p*-value) GO pathways over-represented in LPD B.

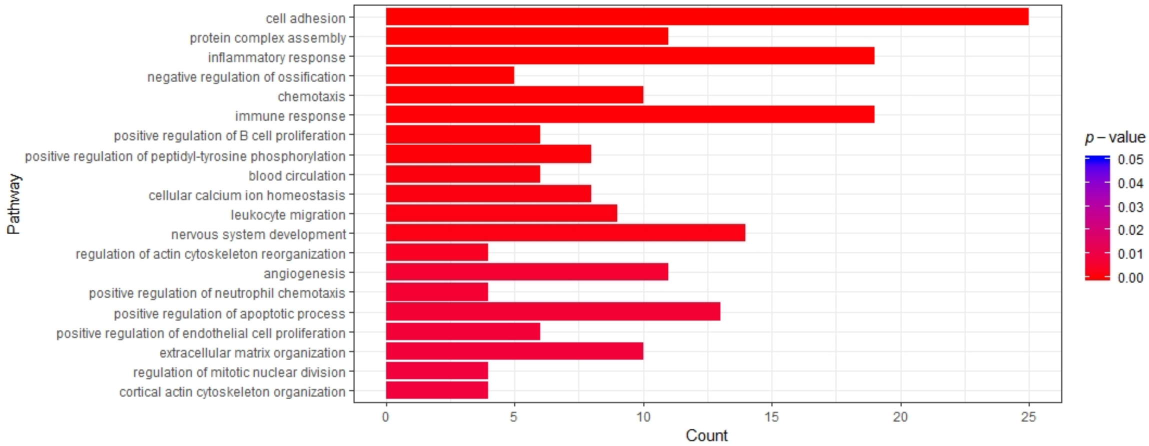


Fig. B.19 Barplot of the KEGG pathways over-represented in LPD B.

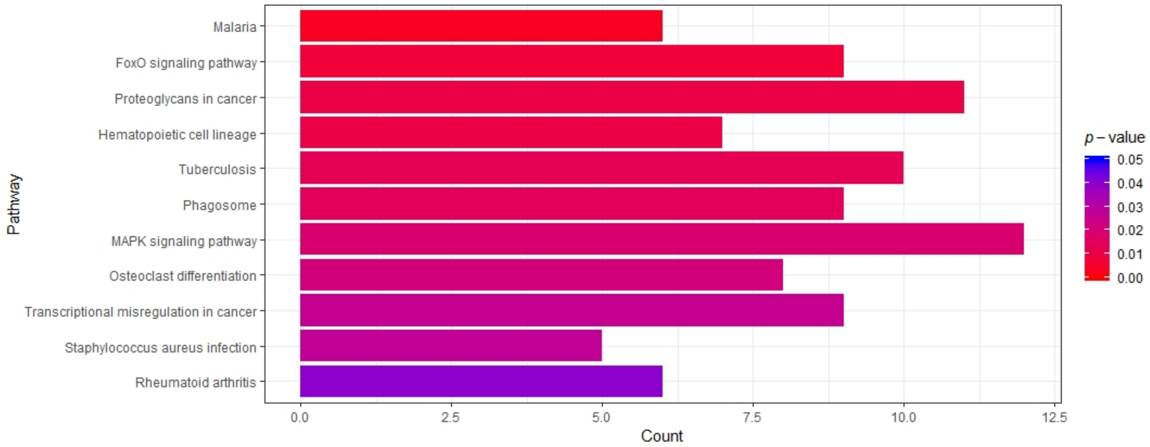
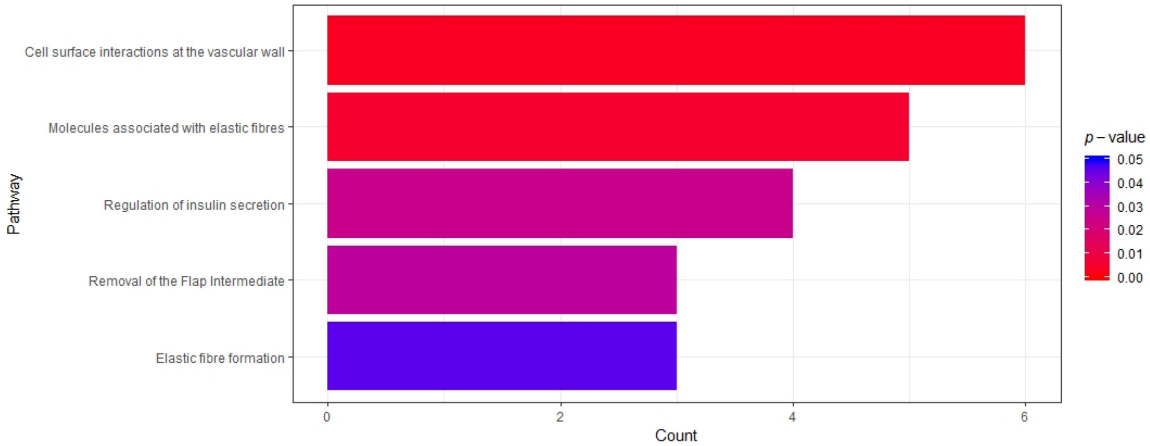


Fig. B.20 Barplot of the Reactome pathways over-represented in LPD B.



B.6.3 Pericol Enriched Pathways

Table B.11 GO pathways over-represented in Pericol.

Pathway ID	**Description**	***p*-value**	**Fold Enrichment**
0030198	extracellular matrix organization	2.51×10^{-27}	8.73
0007155	cell adhesion	1.97×10^{-22}	4.85
0030574	collagen catabolic process	7.51×10^{-16}	12.43
0035987	endodermal cell differentiation	5.08×10^{-10}	16.21
0006954	inflammatory response	2.09×10^{-9}	3.46

0030199	collagen fibril organization	2.93×10^{-8}	11.22
0001525	angiogenesis	4.76×10^{-8}	4.10
0050900	leukocyte migration	6.65×10^{-8}	5.54
0022617	extracellular matrix disassembly	3.84×10^{-7}	6.81
0042060	wound healing	6.81×10^{-7}	6.47
0007165	signal transduction	7.94×10^{-7}	1.99
0030206	chondroitin sulfate biosynthetic process	1.93×10^{-6}	12.73
0045766	positive regulation of angiogenesis	6.18×10^{-6}	4.84
0002576	platelet degranulation	1.03×10^{-5}	5.02
0010628	positive regulation of gene expression	1.06×10^{-5}	3.19
0032967	positive regulation of collagen biosynthetic process	1.68×10^{-5}	12.11
0016525	negative regulation of angiogenesis	2.25×10^{-5}	6.42
0016477	cell migration	2.92×10^{-5}	3.70
0001503	ossification	3.03×10^{-5}	5.47
0007229	integrin-mediated signaling pathway	3.76×10^{-5}	4.82
0001568	blood vessel development	3.82×10^{-5}	8.38
0010575	positive regulation of vascular endothelial growth factor production	4.54×10^{-5}	10.32
0042476	odontogenesis	4.54×10^{-5}	10.32

0001649	osteoblast differentiation	5.94×10^{-5}	4.59
0007507	heart development	6.00×10^{-5}	3.48
0035904	aorta development	6.33×10^{-5}	13.26
0030335	positive regulation of cell migration	6.38×10^{-5}	3.46
0001937	negative regulation of endothelial cell proliferation	6.98×10^{-5}	9.60
0042493	response to drug	8.79×10^{-5}	2.75
0001886	endothelial cell morphogenesis	1.12×10^{-4}	18.09
0071230	cellular response to amino acid stimulus	1.57×10^{-4}	6.77
0001558	regulation of cell growth	1.73×10^{-4}	4.97
0001957	intramembranous ossification	2.96×10^{-4}	26.53
0070208	protein heterotrimerization	3.19×10^{-4}	14.21
0001666	response to hypoxia	4.07×10^{-4}	3.24
0033627	cell adhesion mediated by integrin	4.27×10^{-4}	13.26
0070374	positive regulation of ERK1 and ERK2 cascade	4.80×10^{-4}	3.18
0030334	regulation of cell migration	5.18×10^{-4}	4.84
0030336	negative regulation of cell migration	6.29×10^{-4}	4.19
0051216	cartilage development	6.57×10^{-4}	5.40
0001501	skeletal system development	6.76×10^{-4}	3.49
0050679	positive regulation of epithelial cell proliferation	7.28×10^{-4}	5.31

0050727	regulation of inflammatory response	9.78×10^{-4}	5.05
0051897	positive regulation of protein kinase B signaling	1.21×10^{-3}	4.26
0071560	cellular response to transforming growth factor beta stimulus	1.35×10^{-3}	5.68
0038123	toll-like receptor TLR1:TLR2 signaling pathway	1.85×10^{-3}	39.79
0071727	cellular response to triacyl bacterial lipopeptide	1.85×10^{-3}	39.79
0010906	regulation of glucose metabolic process	1.99×10^{-3}	9.04
0071345	cellular response to cytokine stimulus	1.99×10^{-3}	9.04
0032355	response to estradiol	2.02×10^{-3}	3.94
0048010	vascular endothelial growth factor receptor signaling pathway	2.15×10^{-3}	4.42
0071222	cellular response to lipopolysaccharide	2.16×10^{-3}	3.52
0010759	positive regulation of macrophage chemotaxis	2.22×10^{-3}	14.47
0034446	substrate adhesion-dependent cell spreading	2.46×10^{-3}	6.28
0042981	regulation of apoptotic process	2.85×10^{-3}	2.62
0030208	dermatan sulfate biosynthetic process	2.91×10^{-3}	13.26

0050710	negative regulation of cytokine secretion	2.91×10^{-3}	13.26
0050921	positive regulation of chemotaxis	2.91×10^{-3}	13.26
0006469	negative regulation of protein kinase activity	3.43×10^{-3}	3.62
0010951	negative regulation of endopeptidase activity	3.43×10^{-3}	3.29
0008284	positive regulation of cell proliferation	3.52×10^{-3}	1.96
0048050	post-embryonic eye morphogenesis	3.64×10^{-3}	29.84
0002248	connective tissue replacement involved in inflammatory response wound healing	3.64×10^{-3}	29.84
0030207	chondroitin sulfate catabolic process	4.64×10^{-3}	11.37
0043434	response to peptide hormone	4.70×10^{-3}	5.43
0009611	response to wounding	4.88×10^{-3}	4.42
0001569	patterning of blood vessels	4.95×10^{-3}	7.11
0006687	glycosphingolipid metabolic process	5.18×10^{-3}	5.31
0008015	blood circulation	5.18×10^{-3}	5.31
0030512	negative regulation of transforming growth factor beta receptor signaling pathway	5.27×10^{-3}	4.35

0030203	glycosaminoglycan metabolic process	5.63×10^{-3}	6.86
0002063	chondrocyte development	5.69×10^{-3}	10.61
0060326	cell chemotaxis	5.69×10^{-3}	4.29
0010936	negative regulation of macrophage cytokine production	5.96×10^{-3}	23.87
0022614	membrane to membrane docking	5.96×10^{-3}	23.87
0030449	regulation of complement activation	6.37×10^{-3}	6.63
0051603	proteolysis involved in cellular protein catabolic process	6.83×10^{-3}	4.97
0007565	female pregnancy	7.00×10^{-3}	3.58
0009749	response to glucose	7.08×10^{-3}	4.10
0042127	regulation of cell proliferation	7.09×10^{-3}	2.58
0031663	lipopolysaccharide-mediated signaling pathway	8.04×10^{-3}	6.22
0030855	epithelial cell differentiation	8.14×10^{-3}	3.98
0048514	blood vessel morphogenesis	8.19×10^{-3}	9.36
0007179	transforming growth factor beta receptor signaling pathway	8.35×10^{-3}	3.46
0032496	response to lipopolysaccharide	8.47×10^{-3}	2.67
0051045	negative regulation of membrane protein ectodomain proteolysis	8.80×10^{-3}	19.90
0032964	collagen biosynthetic process	8.80×10^{-3}	19.90
0008360	regulation of cell shape	8.80×10^{-3}	2.84
0007568	aging	8.82×10^{-3}	2.65

0002755	MyD88-dependent toll-like receptor signaling pathway	8.97×10^{-3}	6.03
0010718	positive regulation of epithelial to mesenchymal transition	8.97×10^{-3}	6.03
0071333	cellular response to glucose stimulus	9.56×10^{-3}	4.59
0010596	negative regulation of endothelial cell migration	9.65×10^{-3}	8.84
0006911	phagocytosis, engulfment	1.10×10^{-2}	5.68
0043066	negative regulation of apoptotic process	1.11×10^{-2}	1.84
0048260	positive regulation of receptor-mediated endocytosis	1.12×10^{-2}	8.38
0008285	negative regulation of cell proliferation	1.13×10^{-2}	1.91
0007435	salivary gland morphogenesis	1.21×10^{-2}	17.05
0034616	response to laminar fluid shear stress	1.21×10^{-2}	17.05
0046697	decidualization	1.30×10^{-2}	7.96
0014911	positive regulation of smooth muscle cell migration	1.30×10^{-2}	7.96
0001676	long-chain fatty acid metabolic process	1.49×10^{-2}	7.58
0046718	viral entry into host cell	1.52×10^{-2}	3.48
0033629	negative regulation of cell adhesion mediated by integrin	1.59×10^{-2}	14.92

0010837	regulation of keratinocyte proliferation	1.59×10^{-2}	14.92
0072075	metanephric mesenchyme development	1.59×10^{-2}	14.92
0045630	positive regulation of T-helper 2 cell differentiation	1.59×10^{-2}	14.92
0071407	cellular response to organic cyclic compound	1.60×10^{-2}	4.05
0043410	positive regulation of MAPK cascade	1.60×10^{-2}	3.44
0006915	apoptotic process	1.64×10^{-2}	1.68
0042102	positive regulation of T cell proliferation	1.71×10^{-2}	3.98
0046426	negative regulation of JAK-STAT cascade	1.75×10^{-2}	4.97
0007411	axon guidance	1.90×10^{-2}	2.50
0035924	cellular response to vascular endothelial growth factor stimulus	1.91×10^{-2}	6.92
0006898	receptor-mediated endocytosis	1.91×10^{-2}	2.35
0006955	immune response	1.94×10^{-2}	1.80
0002544	chronic inflammatory response	2.01×10^{-2}	13.26
0044342	type B pancreatic cell proliferation	2.01×10^{-2}	13.26
0021785	branchiomotor neuron axon guidance	2.01×10^{-2}	13.26
0000733	DNA strand renaturation	2.01×10^{-2}	13.26

0035988	chondrocyte proliferation	2.01×10^{-2}	13.26
0001960	negative regulation of cytokine-mediated signaling pathway	2.01×10^{-2}	13.26
0007566	embryo implantation	2.06×10^{-2}	4.74
0060348	bone development	2.06×10^{-2}	4.74
0051496	positive regulation of stress fiber assembly	2.06×10^{-2}	4.74
0031623	receptor internalization	2.23×10^{-2}	4.63
0014068	positive regulation of phosphatidylinositol 3-kinase signaling	2.34×10^{-2}	3.67
0071158	positive regulation of cell cycle arrest	2.39×10^{-2}	6.37
0007159	leukocyte cell-cell adhesion	2.39×10^{-2}	6.37
0045087	innate immune response	2.40×10^{-2}	1.76
0010977	negative regulation of neuron projection development	2.40×10^{-2}	4.52
0048048	embryonic eye morphogenesis	2.47×10^{-2}	11.94
0030168	platelet activation	2.58×10^{-2}	2.77
0035025	positive regulation of Rho protein signal transduction	2.65×10^{-2}	6.12
0002224	toll-like receptor signaling pathway	2.93×10^{-2}	5.90
1903364	positive regulation of cellular protein catabolic process	2.97×10^{-2}	10.85

0045824	negative regulation of innate immune response	2.97×10^{-2}	10.85
0042535	positive regulation of tumor necrosis factor biosynthetic process	2.97×10^{-2}	10.85
0032760	positive regulation of tumor necrosis factor production	2.98×10^{-2}	4.23
0046854	phosphatidylinositol phosphorylation	3.08×10^{-2}	2.96
0007528	neuromuscular junction development	3.22×10^{-2}	5.68
0014047	glutamate secretion	3.22×10^{-2}	5.68
0071260	cellular response to mechanical stimulus	3.27×10^{-2}	3.36
0071456	cellular response to hypoxia	3.37×10^{-2}	2.90
0007169	transmembrane receptor protein tyrosine kinase signaling pathway	3.37×10^{-2}	2.90
0042340	keratan sulfate catabolic process	3.51×10^{-2}	9.95
1902287	semaphorin-plexin signaling pathway involved in axon guidance	3.51×10^{-2}	9.95
0060394	negative regulation of pathway-restricted SMAD protein phosphorylation	3.51×10^{-2}	9.95

0048662	negative regulation of smooth muscle cell proliferation	3.53×10^{-2}	5.49
0048008	platelet-derived growth factor receptor signaling pathway	3.53×10^{-2}	5.49
0043542	endothelial cell migration	3.53×10^{-2}	5.49
0050776	regulation of immune response	3.57×10^{-2}	2.24
0060325	face morphogenesis	3.85×10^{-2}	5.31
2000379	positive regulation of reactive oxygen species metabolic process	3.85×10^{-2}	5.31
2000573	positive regulation of DNA biosynthetic process	4.08×10^{-2}	9.18
0051926	negative regulation of calcium ion transport	4.08×10^{-2}	9.18
0043568	positive regulation of insulin-like growth factor receptor signaling pathway	4.08×10^{-2}	9.18
0045730	respiratory burst	4.08×10^{-2}	9.18
0043518	negative regulation of DNA damage response, signal transduction by p53 class mediator	4.08×10^{-2}	9.18
0001934	positive regulation of protein phosphorylation	4.09×10^{-2}	2.51
0030513	positive regulation of BMP signaling pathway	4.19×10^{-2}	5.13
0030324	lung development	4.20×10^{-2}	3.14

0007596	blood coagulation	4.27×10^{-2}	2.16
0055114	oxidation-reduction process	4.33×10^{-2}	1.55
0030154	cell differentiation	4.37×10^{-2}	1.64
0007166	cell surface receptor signaling pathway	4.38×10^{-2}	1.89
0050873	brown fat cell differentiation	4.54×10^{-2}	4.97
0014066	regulation of phosphatidylinositol 3-kinase signaling	4.61×10^{-2}	3.06
0048146	positive regulation of fibroblast proliferation	4.61×10^{-2}	3.68
0042554	superoxide anion generation	4.68×10^{-2}	8.53
2000353	positive regulation of endothelial cell apoptotic process	4.68×10^{-2}	8.53
0010812	negative regulation of cell-substrate adhesion	4.68×10^{-2}	8.53
0043537	negative regulation of blood vessel endothelial cell migration	4.68×10^{-2}	8.53
2000147	positive regulation of cell motility	4.68×10^{-2}	8.53
0014912	negative regulation of smooth muscle cell migration	4.68×10^{-2}	8.53
0030194	positive regulation of blood coagulation	4.68×10^{-2}	8.53
0019221	cytokine-mediated signaling pathway	4.70×10^{-2}	2.43
0045471	response to ethanol	4.86×10^{-2}	2.65

0001764	neuron migration	4.86×10^{-2}	2.65
0060548	negative regulation of cell death	4.88×10^{-2}	3.62
0070555	response to interleukin-1	4.90×10^{-2}	4.82
1902042	negative regulation of extrinsic apoptotic signaling pathway via death domain receptors	4.90×10^{-2}	4.82
0090291	negative regulation of osteoclast proliferation	4.95×10^{-2}	39.79
0042495	detection of triacyl bacterial lipopeptide	4.95×10^{-2}	39.79
1905049	negative regulation of metallopeptidase activity	4.95×10^{-2}	39.79
0001300	chronological cell aging	4.95×10^{-2}	39.79
0001798	positive regulation of type IIa hypersensitivity	4.95×10^{-2}	39.79
1903225	negative regulation of endodermal cell differentiation	4.95×10^{-2}	39.79
0070483	detection of hypoxia	4.95×10^{-2}	39.79
1905005	regulation of epithelial to mesenchymal transition involved in endocardial cushion formation	4.95×10^{-2}	39.79
0009756	carbohydrate mediated signaling	4.95×10^{-2}	39.79
0009440	cyanate catabolic process	4.95×10^{-2}	39.79
0061441	renal artery morphogenesis	4.95×10^{-2}	39.79

Table B.12 KEGG pathways over-represented in Pericol.

Pathway ID	**Description**	***p*-value**	**Fold Enrichment**
04512	ECM-receptor interaction	2.39×10^{-13}	7.77
04510	Focal adhesion	3.72×10^{-10}	4.17
05146	Amoebiasis	6.06×10^{-9}	5.50
04151	PI3K-Akt signaling pathway	1.76×10^{-7}	2.85
04145	Phagosome	3.02×10^{-7}	4.09
05144	Malaria	2.23×10^{-5}	6.27
04974	Protein digestion and absorption	2.33×10^{-5}	4.54
05152	Tuberculosis	5.45×10^{-5}	3.12
04380	Osteoclast differentiation	7.96×10^{-5}	3.52
05205	Proteoglycans in cancer	2.46×10^{-4}	2.76
05150	Staphylococcus aureus infection	3.09×10^{-4}	5.12
04670	Leukocyte transendothelial migration	1.19×10^{-3}	3.20
04611	Platelet activation	3.18×10^{-3}	2.83
00532	Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate	3.46×10^{-3}	7.68
04142	Lysosome	5.72×10^{-3}	2.79
05222	Small cell lung cancer	6.05×10^{-3}	3.25
04514	Cell adhesion molecules (CAMs)	6.22×10^{-3}	2.60
04610	Complement and coagulation cascades	6.70×10^{-3}	3.56
05200	Pathways in cancer	8.29×10^{-3}	1.80
04015	Rap1 signaling pathway	1.83×10^{-2}	2.05

04620	Toll-like receptor signaling pathway	2.13×10^{-2}	2.61
05323	Rheumatoid arthritis	2.36×10^{-2}	2.79
05145	Toxoplasmosis	2.60×10^{-2}	2.51
05140	Leishmaniasis	2.70×10^{-2}	3.03
00920	Sulfur metabolism	3.24×10^{-2}	10.24
04810	Regulation of actin cytoskeleton	3.90×10^{-2}	1.90

Table B.13 Reactome pathways over-represented in Pericol.

Pathway ID	**Description**	***p*-value**	**Fold Enrichment**
3000178	ECM proteoglycans	2.74×10^{-20}	11.65
216083	Integrin cell surface interactions	2.31×10^{-16}	9.38
1442490	Collagen degradation	1.88×10^{-13}	9.98
2022090	Assembly of collagen fibrils and other multimeric structures	4.34×10^{-10}	10.23
1650814	Collagen biosynthesis and modifying enzymes	7.05×10^{-10}	8.03
186797	Signaling by PDGF	9.41×10^{-10}	12.60
3000171	Non-integrin membrane-ECM interactions	1.36×10^{-8}	10.08
3000170	Syndecan interactions	4.54×10^{-8}	12.45
1474244	Extracellular matrix organization	2.56×10^{-6}	15.69
202733	Cell surface interactions at the vascular wall	3.35×10^{-6}	6.85
2129379	Molecules associated with elastic fibres	1.21×10^{-5}	7.96

1474228	Degradation of the extracellular matrix	1.69×10^{-5}	5.17
2022870	Chondroitin sulfate biosynthesis	1.75×10^{-5}	11.76
3000157	Laminin interactions	2.11×10^{-5}	8.96
114608	Platelet degranulation	2.90×10^{-5}	3.88
1566948	Elastic fibre formation	1.38×10^{-4}	11.20
3000480	Scavenging by Class A Receptors	1.82×10^{-4}	10.61
2243919	Crosslinking of collagen fibrils	2.11×10^{-4}	15.28
3595177	Defective CHSY1 causes TPBS	1.29×10^{-3}	16.81
419037	NCAM1 interactions	1.37×10^{-3}	5.60
166058	MyD88:MAL(TIRAP) cascade initiated on plasma membrane	1.66×10^{-3}	9.34
5602498	MyD88 deficiency (TLR2/4)	3.56×10^{-3}	12.22
2022923	Dermatan sulfate biosynthesis	3.56×10^{-3}	12.22
5603041	IRAK4 deficiency (TLR2/4)	3.56×10^{-3}	12.22
114604	GPVI-mediated activation cascade	6.01×10^{-3}	4.20
2024101	CS/DS degradation	7.36×10^{-3}	9.60
977606	Regulation of Complement cascade	8.82×10^{-3}	6.00
75892	Platelet Adhesion to exposed collagen	9.00×10^{-3}	8.96
2214320	Anchoring fibril formation	9.00×10^{-3}	8.96
399956	CRMPs in Sema3A signaling	1.08×10^{-2}	8.40
168179	Toll Like Receptor TLR1:TLR2 Cascade	1.21×10^{-2}	16.81

1592389	Activation of Matrix Metalloproteinases	1.57×10^{-2}	5.09
416700	Other semaphorin interactions	1.76×10^{-2}	7.08
3560783	Defective B4GALT7 causes EDS, progeroid type	2.02×10^{-2}	6.72
4420332	Defective B3GALT6 causes EDSP2 and SEMDJL1	2.02×10^{-2}	6.72
3560801	Defective B3GAT3 causes JDSS-DHD	2.02×10^{-2}	6.72
1236973	Cross-presentation of particulate exogenous antigens (phagosomes)	2.18×10^{-2}	12.60
3595172	Defective CHST3 causes SED-CJD	2.18×10^{-2}	12.60
3595174	Defective CHST14 causes EDS, musculocontractural type	2.18×10^{-2}	12.60
381426	Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs)	2.31×10^{-2}	6.40
389357	CD28 dependent PI3K/Akt signaling	2.62×10^{-2}	6.11
210500	Glutamate Neurotransmitter Release Cycle	3.29×10^{-2}	5.60

1971475	A tetrasaccharide linker sequence is required for GAG synthesis	4.05×10^{-2}	5.17
1660662	Glycosphingolipid metabolism	4.35×10^{-2}	3.73
210990	PECAM1 interactions	4.75×10^{-2}	8.40

Fig. B.21 Barplot of the top 20 (ordered by p-value) GO pathways over-represented in Pericol.

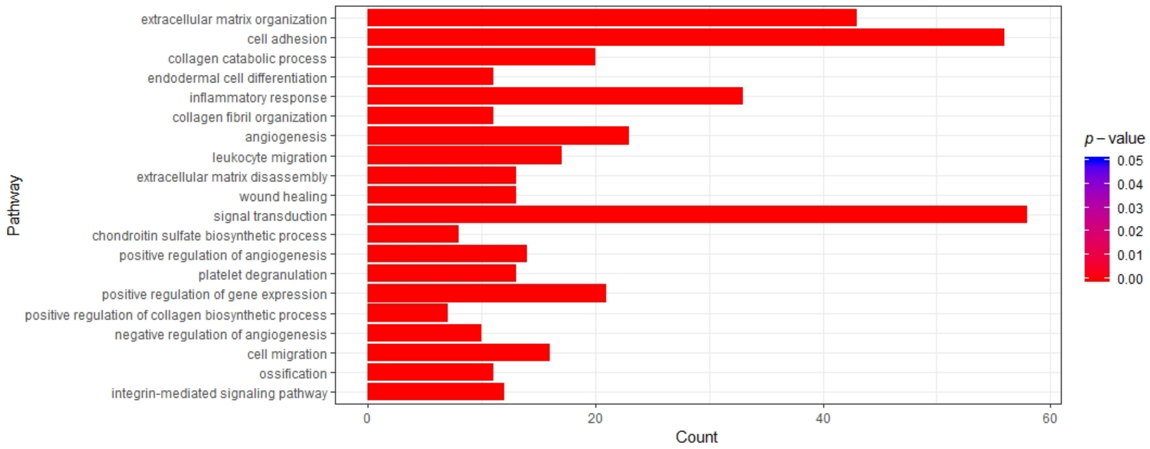


Fig. B.22 Barplot of the top 20 (ordered by p-value) KEGG pathways over-represented in Pericol.

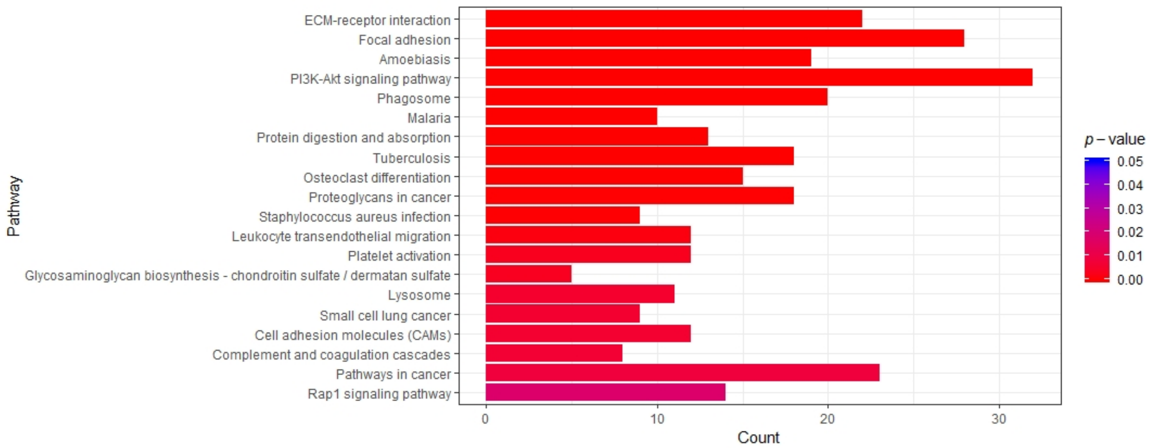
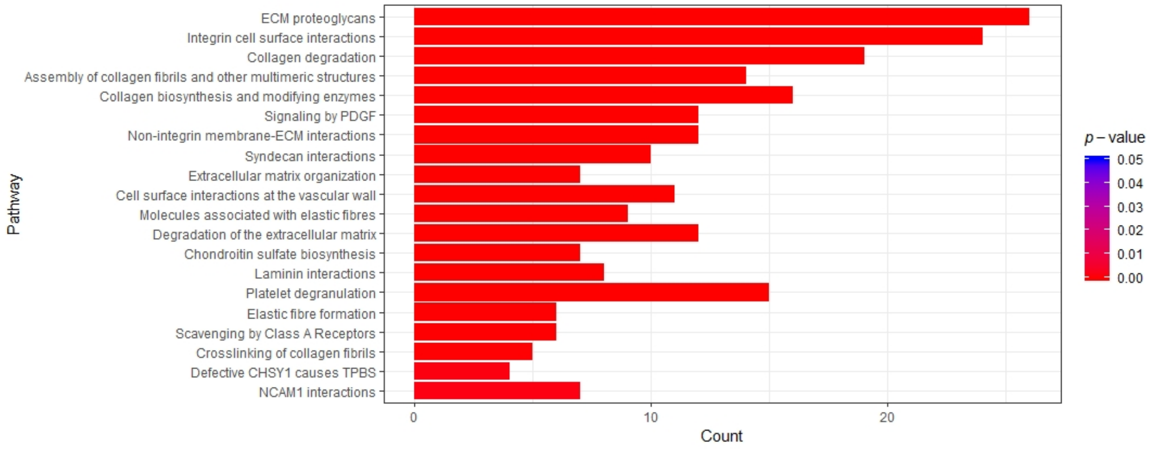


Fig. B.23 Barplot of the top 20 (ordered by *p*-value) Reactome pathways over-represented in Pericol.



B.7 CRC Differential Methylation

Table B.14 A differential methylation analysis performed on the TCGA-COAD methylation dataset using Limma [65] and methylGSA [251]. Results restricted to only include genes where the majority of CpG sites were significantly (adj *p*-value ≤ 0.05) hyper-methylated and the genes were under expressed, or genes where the majority of CpG sites were significantly (adj *p*-value ≤ 0.05) hypo-methylated and the genes were over expressed.

CG ID	logFC	P.Value	adj.P.Val	Gene Symbol
cg03817621	0.184648306	1.15e-07	5.6e-06	A1CF
cg16531903	0.11884799	6.09e-05	0.000909273	A1CF
cg24411946	0.22016005	2.13e-07	9.11e-06	A1CF
cg04919581	0.14541615	5.29e-08	3.08e-06	ACOT11
cg11177235	0.113410513	2.52e-05	0.000438634	ACOT11
cg13458781	0.142263087	3.27e-08	2.14e-06	ACOT11
cg04699460	0.044267614	0.000358867	0.003866349	ACSM3
cg10265472	0.012608956	0.008158775	0.045052703	ACSM3
cg00096810	-0.110286892	0.002060027	0.015616298	ADAMTS2
cg10208897	-0.061067412	0.004117343	0.026749801	ADAMTS2
cg20422099	-0.13978107	0.002715654	0.019368919	ADAMTS2

cg02262923	-0.078592923	0.000365816	0.003927325	ADAP2
cg01129238	0.101639739	0.006460857	0.037749673	ADTRP
cg01699293	0.17530894	9.05e-09	8.17e-07	ADTRP
cg26761744	0.164732205	0.001991218	0.015215569	ADTRP
cg03431524	0.086159277	0.000680705	0.006485845	AGFG2
cg18991321	0.145084721	2.23e-08	1.6e-06	AGFG2
cg19543017	0.070686181	0.00056802	0.005606519	AGFG2
cg23606385	0.015390539	0.003513402	0.023672262	AGFG2
cg01540571	0.01050801	0.007379017	0.041743255	AGMAT
cg17385448	0.07926638	2.03e-07	8.77e-06	AGMAT
cg17759086	0.003439294	0.001073172	0.00932609	AGMAT
cg04812347	-0.096099184	4.85e-05	0.000754616	AIF1
cg18113826	-0.087750362	3.4e-07	1.32e-05	AIF1
cg19563932	-0.077175948	0.001821208	0.014197668	AIF1
cg21440587	-0.08380856	0.007305971	0.041406914	AIF1
cg25403205	-0.096843247	1.59e-08	1.25e-06	AIF1
cg17520539	-0.090004548	0.000307769	0.003414371	AKAP12
cg11496569	-0.075398438	0.004422985	0.028252391	AKT3
cg23166773	-0.081990816	4.45e-09	4.89e-07	AKT3
cg21054703	-0.05237527	0.005832398	0.034898937	ALOX5AP
cg08076018	-0.114920336	1.58e-05	0.000298933	ANGPTL2
cg11213150	-0.102155663	2.77e-07	1.12e-05	ANGPTL2
cg13508369	-0.062833282	2.52e-05	0.000439641	ANGPTL2
cg13662634	-0.102798218	3.58e-05	0.000586622	ANGPTL2
cg14281592	-0.111205077	3.35e-07	1.31e-05	ANGPTL2
cg09983301	-0.188427534	1.86e-08	1.4e-06	ANTXR1

cg19130824	-0.074983986	9.13e-05	0.001268339	ANXA5
cg07434260	0.055145766	1.06e-05	0.000216256	AP1M2
cg08148553	0.021460037	0.003095218	0.021459338	AP1M2
cg15555932	0.051900916	2.85e-06	7.37e-05	AP1M2
cg22614759	-0.119737575	1.18e-06	3.61e-05	ARHGDIB
cg09935994	-0.040991561	0.008231202	0.04534706	ARL4C
cg24441922	-0.099206275	0.003754298	0.02492116	ARL4C
cg10062193	-0.224968933	1.44e-14	6.01e-11	ATP10D
cg03085712	0.013467253	0.007091268	0.040487287	ATP2C2
cg03548384	0.144232345	7.23e-08	3.91e-06	ATP2C2
cg06786050	0.121821902	1.09e-08	9.39e-07	ATP2C2
cg07277633	0.025477896	0.000292359	0.003277385	ATP2C2
cg27459353	0.026184642	7.9e-05	0.001125208	ATP2C2
cg15310871	-0.037876848	5.38e-10	1.09e-07	ATP6V1B2
cg16479633	-0.122554968	1.94e-06	5.38e-05	ATP6V1B2
cg22479161	-0.047728454	0.000199731	0.002401092	BASP1
cg02585702	-0.10585109	0.000917499	0.008234603	BCAT1
cg04011247	-0.127849014	0.005818576	0.034834323	BCAT1
cg04543413	-0.11182852	0.004469585	0.028476118	BCAT1
cg08724310	-0.13530639	0.001713177	0.013527503	BCAT1
cg09800500	-0.077653089	0.009103718	0.048953418	BCAT1
cg10764357	-0.127374332	0.002241639	0.016682348	BCAT1
cg20399616	-0.08627581	0.00033553	0.003657901	BCAT1
cg22229906	-0.158767998	0.000154382	0.001950156	BCAT1
cg23792314	-0.146503179	1.75e-07	7.8e-06	BCAT1
cg23930313	-0.109244935	0.00132155	0.011017504	BCAT1

cg05663031	-0.074646508	3.51e-08	2.26e-06	BCL6
cg06070445	-0.051042125	0.000137304	0.001768387	BCL6
cg06164260	-0.113684483	2.27e-09	3e-07	BCL6
cg09643398	-0.027996284	0.000628828	0.006085838	BCL6
cg17394304	-0.070120704	1.36e-05	0.000264856	BCL6
cg23655939	-0.079771426	2.41e-07	1e-05	BCL6
cg00177496	0.015437558	0.000258542	0.002966641	BDH1
cg00456086	0.029721007	1.18e-05	0.000235277	BDH1
cg03792768	0.009801446	0.003047623	0.021198377	BDH1
cg07155478	0.039461261	2.39e-05	0.000420404	BDH1
cg09363194	0.002263322	0.006269644	0.036898919	BDH1
cg09610644	0.138932766	2.92e-07	1.17e-05	BDH1
cg16775939	0.176085299	9.58e-06	0.000198449	BDH1
cg19798702	0.199916986	1.16e-08	9.81e-07	BDH1
cg00820740	-0.039822452	0.003680517	0.024544529	BICC1
cg07690768	-0.083336769	0.002995867	0.020921047	BICD1
cg03280108	-0.112932168	1.4e-05	0.000270826	BNIP3L
cg18341905	-0.073406972	0.003990982	0.026121854	BNIP3L
cg02720155	0.166470964	4.73e-05	0.000737866	BTNL3
cg26081900	0.145516387	0.000229078	0.002686941	BTNL3
cg02288969	0.202110673	1.55e-07	7.06e-06	C1orf105
cg03748243	0.146747909	0.000271516	0.00308565	C1orf105
cg09989886	0.007099679	1.6e-05	0.000302159	C1orf105
cg11584111	0.008099298	0.002566075	0.018541165	C1orf105
cg11841239	0.036239636	0.003673737	0.024509215	C1orf105
cg03714163	0.012123344	4.71e-06	0.00011092	C1orf109

cg06917450	0.083831513	8.7e-05	0.001220389	C1orf109
cg07997689	0.013043978	0.003251765	0.022295298	C1orf109
cg12339802	0.023352367	0.0024382	0.017822999	C1orf109
cg14170840	0.133859161	0.000774139	0.007186369	C1orf109
cg18172330	0.007265984	0.004230518	0.027307935	C1orf109
cg21010197	0.005191234	0.000921723	0.008261715	C1orf109
cg21933021	0.025668598	0.006728665	0.038900967	C1orf109
cg22449745	0.095895105	5.22e-07	1.86e-05	C1orf109
cg24088508	0.08489184	0.000428143	0.004462431	C1orf109
cg27170260	0.012614132	0.005946594	0.035413862	C1orf109
cg27170383	0.009617234	0.003940491	0.02586781	C1orf109
cg20314620	0.012132897	0.007491529	0.042221802	C1orf174
cg17612991	-0.072395537	0.00654993	0.038132287	C3
cg08890828	0.004634842	0.004039813	0.026367435	C5orf30
cg09671005	0.007061006	0.004073155	0.026534331	C5orf30
cg03449125	0.13772522	0.000192037	0.002325112	CAPN5
cg03547523	0.011219859	0.000558252	0.005527191	CAPN5
cg10548968	0.125656739	5.86e-09	5.95e-07	CAPN5
cg14918391	0.00596063	0.002911764	0.020468345	CAPN5
cg16641915	0.009978135	0.009223811	0.049438776	CAPN5
cg00350296	-0.121175447	5.62e-07	1.98e-05	CD248
cg07145284	-0.101339343	0.001598104	0.012811983	CD248
cg13860849	-0.079823146	0.000332922	0.003635481	CD248
cg18935353	-0.083467487	0.000356139	0.003840028	CD248
cg07440264	-0.091771543	0.001089405	0.009440235	CD59
cg08608126	-0.086343828	2.92e-06	7.51e-05	CD59

cg09864245	-0.109420152	0.005737077	0.034480354	CD59
cg15382933	-0.087817492	1.43e-08	1.16e-06	CD59
cg16578387	-0.032030633	0.000904675	0.008142859	CD59
cg23903301	-0.119673075	2.44e-07	1.01e-05	CD59
cg06288788	-0.112779926	0.006645642	0.038554927	CDH11
cg00456593	0.155503637	6.4e-08	3.56e-06	CDHR5
cg03875235	0.15596109	7.74e-08	4.13e-06	CDHR5
cg05198900	0.123161829	6.19e-07	2.14e-05	CDHR5
cg07336987	0.124840134	2.64e-07	1.07e-05	CDHR5
cg08057038	0.09895496	3.04e-11	1.47e-08	CDHR5
cg11464053	0.101432852	0.00507154	0.03139266	CDHR5
cg13937462	0.143434368	1.62e-06	4.64e-05	CDHR5
cg14149701	0.098806389	2.54e-06	6.69e-05	CDHR5
cg15806880	0.138028767	1.01e-08	8.9e-07	CDHR5
cg16753939	0.144839798	3.36e-10	7.85e-08	CDHR5
cg22289115	0.114369376	3.76e-07	1.44e-05	CDHR5
cg27261397	0.135688561	1.87e-07	8.2e-06	CDHR5
cg03834767	-0.151861623	3.65e-05	0.000595565	CDK14
cg04331802	0.150180329	0.000102388	0.001390828	CDS1
cg04673465	0.138014312	1.98e-09	2.7e-07	CDS1
cg17084361	0.017995571	5.12e-05	0.000789476	CDS1
cg22884714	0.228713877	7.26e-08	3.92e-06	CDS1
cg00919055	0.132654046	4.63e-07	1.69e-05	CDX1
cg03545404	0.09067197	0.004642515	0.029331572	CDX1
cg09690765	0.295201947	1.17e-09	1.87e-07	CDX1
cg11117637	0.142456238	1.03e-07	5.12e-06	CDX1

cg11503274	0.101046815	4.71e-06	0.000111061	CDX1
cg11524248	0.160599531	1.38e-06	4.1e-05	CDX1
cg15452204	0.157552555	3.96e-08	2.47e-06	CDX1
cg17378342	0.152668608	3.06e-08	2.03e-06	CDX1
cg17512474	0.117766528	1.09e-08	9.41e-07	CDX1
cg18424208	0.139737054	1.57e-07	7.14e-06	CDX1
cg23266594	0.119106484	5.64e-06	0.000128601	CDX1
cg24216701	0.138813432	1.44e-07	6.68e-06	CDX1
cg25132276	0.151766875	1.08e-07	5.31e-06	CDX1
cg26531174	0.190940469	3.76e-08	2.37e-06	CDX1
cg07922062	0.120426117	8.03e-05	0.001141028	CEBPG
cg15046693	0.090371507	1.21e-05	0.000240008	CEBPG
cg25876406	0.002362408	0.004219917	0.027256702	CEBPG
cg02256576	0.086043686	0.000170259	0.002110726	CES3
cg03447083	0.061689679	0.000438096	0.00454345	CES3
cg06799321	0.076921957	0.001051175	0.009173776	CES3
cg09407859	0.226602391	5.72e-05	0.000864371	CES3
cg26538442	0.139858448	0.002335779	0.017230773	CES3
cg19081101	-0.113995718	0.007217134	0.041016706	CHI3L1
cg01940855	-0.137227706	0.007887153	0.043916509	CHST11
cg06647068	-0.084929359	0.002951423	0.02068362	CHST11
cg07911905	-0.064799382	0.002501188	0.018179752	CHST11
cg11425280	-0.145475059	0.002228972	0.016607116	CHST11
cg05228404	-0.134349003	0.00383083	0.025301122	CHST15
cg09341154	-0.128623144	0.006890051	0.039620252	CHST15
cg04861869	-0.039848002	0.004169811	0.027007624	CHSY1

cg07891271	-0.105015054	0.000836204	0.007644699	CHSY1
cg08165960	-0.110776267	3.92e-07	1.48e-05	CHSY1
cg14232082	-0.087364192	2.07e-06	5.68e-05	CHSY1
cg14397171	-0.068934205	0.000530782	0.005301579	CHSY1
cg15754630	-0.053719678	0.003895383	0.025641951	CHSY1
cg19589317	-0.088692387	0.000512164	0.005150314	CHSY1
cg20357538	-0.094690609	8.89e-07	2.86e-05	CHSY1
cg24312730	-0.021250976	0.002298188	0.017011453	CHSY1
cg08808811	0.007688369	0.000378283	0.00403612	CLCN2
cg22613010	0.007595998	0.004895525	0.030545168	CLCN2
cg00072720	0.034648724	5.92e-05	0.000888798	CLDN7
cg03186999	0.047292485	7.41e-06	0.000160885	CLDN7
cg05490983	0.060500025	2.95e-06	7.57e-05	CLDN7
cg13724311	0.171821652	1.27e-07	6.05e-06	CLDN7
cg15298719	0.095249664	9.09e-08	4.68e-06	CLDN7
cg17265693	0.038266049	0.000230988	0.002705655	CLDN7
cg24944395	0.029602163	3.83e-05	0.000620325	CLDN7
cg02913511	0.137770402	1.91e-07	8.36e-06	CNNM4
cg07224438	0.003052915	0.00156688	0.012602275	CNNM4
cg08313757	0.073339912	0.004122599	0.026774957	CNNM4
cg11158729	0.029257608	0.005706002	0.034333709	CNNM4
cg11464842	0.236148875	9.28e-06	0.000193314	CNNM4
cg12942038	0.228246503	4.27e-06	0.00010236	CNNM4
cg14228484	0.167335541	0.000239995	0.002790672	CNNM4
cg15009198	0.167641842	0.000906455	0.008155484	CNNM4
cg17383207	0.018359255	0.000174438	0.002152498	CNNM4

cg19953799	0.275507855	2.25e-06	6.06e-05	CNNM4
cg21238414	0.008925947	0.000106641	0.0014381	CNNM4
cg00172849	-0.110547322	0.001013319	0.008911143	COL11A1
cg03520644	-0.163513669	0.000586213	0.005748702	COL11A1
cg20847625	-0.053014258	0.006116968	0.036212379	COL11A1
cg26436330	-0.195154503	0.001341379	0.011150196	COL11A1
cg19461644	-0.128642045	0.00283143	0.020031266	COL15A1
cg00160583	-0.058114244	0.000312839	0.003461021	COL16A1
cg00439089	-0.07185919	2.06e-05	0.000372258	COL1A1
cg02134839	-0.073187044	0.005672158	0.034182543	COL1A1
cg02827061	-0.064894259	0.001028916	0.009017077	COL1A1
cg03743861	-0.043119421	1.95e-07	8.48e-06	COL1A1
cg03799835	-0.058893599	0.001307671	0.010925715	COL1A1
cg11615029	-0.074988444	7.99e-10	1.44e-07	COL1A1
cg11993636	-0.082964658	1.37e-07	6.42e-06	COL1A1
cg16514513	-0.075506207	2.7e-08	1.86e-06	COL1A1
cg18405262	-0.055306379	2.76e-06	7.16e-05	COL1A1
cg24540710	-0.050277984	3.67e-05	0.000599104	COL1A1
cg25735490	-0.073271292	1.69e-08	1.31e-06	COL1A1
cg00765737	-0.14916473	2.49e-07	1.03e-05	COL4A2
cg04771838	-0.127286462	5.91e-07	2.06e-05	COL4A2
cg16872841	-0.122111096	2.47e-06	6.53e-05	COL4A2
cg13618741	-0.083147356	0.000689065	0.006548014	COL5A1
cg13969662	-0.080127426	0.006246275	0.036791313	COL5A1
cg24784350	-0.136596556	0.004882765	0.030487164	COL6A2
cg20185461	-0.142297073	0.000810098	0.007450748	COMP

cg02308245	0.017323415	0.003262526	0.022349733	COX4I1
cg03303025	0.002993548	0.004261472	0.027461191	COX4I1
cg04399085	0.114412228	3.97e-09	4.49e-07	COX4I1
cg05744264	0.160011173	8.33e-06	0.000176666	COX4I1
cg15819333	0.156638699	1.42e-12	1.73e-09	COX4I1
cg21963318	0.027952284	0.000688862	0.006546712	COX4I1
cg19255191	0.008011924	0.000139931	0.001796561	COX5B
cg00574958	0.11044616	0.000441103	0.004568294	CPT1A
cg00941258	0.04897125	0.00131323	0.010963937	CPT1A
cg01260103	0.007900073	0.000720477	0.00678874	CPT1A
cg09737197	0.111436266	0.008187051	0.045166652	CPT1A
cg14073497	0.078051371	0.000210559	0.002507898	CPT1A
cg16296442	0.051365177	0.004461062	0.028437369	CPT1A
cg17058475	0.104474556	0.000322485	0.003544473	CPT1A
cg19081843	0.047314248	0.000422132	0.004412327	CPT1A
cg20562447	0.066516926	0.002064087	0.015641305	CPT1A
cg22911054	0.203425177	2.43e-08	1.71e-06	CPT1A
cg26192826	0.224317645	1.37e-07	6.4e-06	CPT1A
cg17679427	-0.086151782	0.000694183	0.006587723	CREM
cg00687714	0.219981355	9.49e-13	1.29e-09	CRYM
cg07666035	0.010796743	4.33e-05	0.000686794	CRYM
cg10757684	0.005852212	0.000212277	0.002524482	CRYM
cg23014871	-0.191911489	9.13e-15	4.11e-11	CSGALNACT2
cg24376955	-0.062151343	0.000407285	0.004281995	CSGALNACT2
cg00261832	-0.083321299	0.000309099	0.003426345	CTGF
cg18222609	-0.064748793	0.007416752	0.041906442	CTGF

cg21919729	-0.044133015	0.001492655	0.012132414	CTSB
cg10980495	-0.02837531	0.001457828	0.011912658	CTSD
cg11946165	-0.051970785	2.28e-09	3e-07	CTSK
cg20802392	-0.075528249	0.000156608	0.001972803	CTSK
cg20817941	-0.067731169	2.07e-05	0.000373391	CTSK
cg24060908	0.176554272	8.8e-06	0.000185208	CWH43
cg00565882	-0.097337802	0.000272234	0.003092742	CYP1B1
cg01410359	-0.070432773	0.001901795	0.014682647	CYP1B1
cg02162897	-0.029479026	0.001664364	0.013222273	CYP1B1
cg02486145	-0.090835667	3.4e-06	8.5e-05	CYP1B1
cg03890222	-0.103133591	0.003815176	0.025221289	CYP1B1
cg06264984	-0.087491263	0.001337832	0.011127719	CYP1B1
cg09799983	-0.114275306	0.000300834	0.003352928	CYP1B1
cg16439198	-0.150278098	0.002461755	0.017956076	CYP1B1
cg20254225	-0.095746429	0.006270534	0.036903609	CYP1B1
cg11540204	0.231629411	1.5e-05	0.000287124	CYP2J2
cg26815229	0.024122411	0.00731849	0.04146848	CYP2J2
cg02824980	0.086557871	1.18e-09	1.88e-07	CYP4F12
cg05722906	0.060425749	3.38e-08	2.19e-06	CYP4F12
cg14711976	0.099118506	0.003397732	0.023058201	CYP4F12
cg23080427	0.092271194	1.2e-07	5.77e-06	CYP4F12
cg07602008	-0.030733803	1.48e-06	4.32e-05	CYR61
cg15648041	-0.062494147	0.004206213	0.027189594	CYR61
cg21091766	-0.065619952	0.001639088	0.013069182	DBN1
cg17239057	-0.094789785	1.97e-05	0.00035889	DEGS1
cg11217654	-0.055692851	1.48e-07	6.8e-06	DENND5A

cg05080966	0.045363715	0.003268255	0.022379306	DGAT1
cg08079052	0.088388238	1.79e-05	0.000332189	DGAT1
cg11976048	0.009286064	0.004141081	0.026868116	DHRS11
cg13261938	0.10450984	2.9e-05	0.000493195	DHRS11
cg14076390	0.141301003	6e-06	0.000135282	DHRS11
cg19847411	0.005644689	0.003710842	0.024699706	DHRS11
cg24338748	0.0582815	0.001415241	0.01163137	DHRS11
cg01540102	0.185121572	2.2e-05	0.000392267	DOK4
cg02993352	-0.141376347	0.005891452	0.035173974	DPYSL2
cg10399402	-0.020787628	0.00429656	0.027634228	DPYSL2
cg10789956	-0.094553885	0.000859562	0.007814468	DPYSL2
cg19610383	-0.051922191	0.000384864	0.004094094	DPYSL2
cg15366353	-0.116888765	0.001488318	0.012105615	DSE
cg07946977	-0.152957903	2.48e-10	6.39e-08	DUSP10
cg00974629	-0.066476725	0.002397408	0.017589092	EDNRA
cg04045079	-0.147386236	0.001805774	0.014101578	EDNRA
cg05102394	-0.161445466	0.000377978	0.004033294	EDNRA
cg05618426	-0.096536432	0.004045487	0.02639707	EDNRA
cg14948448	-0.043538714	0.000118272	0.001564893	EDNRA
cg17073859	-0.053943078	0.00237362	0.017450772	EDNRA
cg20557687	-0.142466031	0.002936187	0.020600175	EDNRA
cg05385513	-0.155579386	0.000758659	0.007073076	EFEMP1
cg16100120	-0.173242736	0.000156009	0.001966692	EFEMP1
cg20786074	-0.130178994	0.000125965	0.001648391	EFEMP1
cg24719005	-0.132812096	0.000224772	0.00264623	EFEMP1
cg02586730	-0.05433554	0.002454857	0.017918301	EFEMP2

cg11265839	-0.089393592	4.55e-07	1.67e-05	ELK3
cg14265043	-0.025583141	0.005894364	0.035187304	ELK3
cg05050341	-0.034151123	0.003913509	0.025733225	ENG
cg13531977	0.154965653	3.13e-05	0.000525473	EPB41L4B
cg14071612	0.045806128	0.003782067	0.025059246	EPB41L4B
cg14235574	0.175659986	5.21e-05	0.000800935	EPB41L4B
cg14306058	0.134579529	2.08e-06	5.68e-05	EPB41L4B
cg01687402	0.156926149	1.57e-06	4.55e-05	EPN3
cg04267101	0.062775567	0.00168957	0.013382085	EPN3
cg08096854	0.215720924	1.1e-09	1.79e-07	EPN3
cg08449531	0.013756541	0.000288548	0.003242182	EPN3
cg08842032	0.116219763	0.000458671	0.004711064	EPN3
cg09827751	0.060602093	8.58e-05	0.001206625	EPN3
cg10567637	0.022453754	0.000560965	0.005549192	EPN3
cg12791192	0.007260404	0.009214852	0.049399443	EPN3
cg16010178	0.062817868	3.25e-05	0.000541644	EPN3
cg24006361	0.011841436	0.006880961	0.039580854	EPN3
cg24236903	0.010972409	3.21e-05	0.000535806	EPN3
cg00491404	0.152557454	7.17e-10	1.34e-07	EPS8L3
cg00515905	0.177137038	2.25e-09	2.97e-07	EPS8L3
cg03956353	0.125180016	5.34e-08	3.1e-06	EPS8L3
cg23957643	0.131738654	3.73e-08	2.36e-06	EPS8L3
cg01613817	-0.118647713	0.00019861	0.002389719	ERG
cg04163967	-0.132850576	3.73e-08	2.36e-06	ERG
cg17228105	-0.071168012	0.004222695	0.027268317	ERG
cg23340514	-0.085309155	0.002461232	0.017953258	ERG

cg03195230	0.00378182	0.000172199	0.002130179	ESRRA
cg03527086	0.070082439	1.32e-09	2.03e-07	ESRRA
cg03764506	0.005791764	0.000984268	0.008702823	ESRRA
cg01330762	0.007983178	9.37e-05	0.001294436	ETHE1
cg12984635	0.133647176	7.69e-09	7.24e-07	ETHE1
cg15012607	0.127187003	1.67e-05	0.000313601	ETHE1
cg16664233	0.007226988	0.000804258	0.007406318	ETHE1
cg25261059	0.040212417	0.004096048	0.026651026	ETHE1
cg14178794	-0.139717645	1.58e-07	7.18e-06	EVC
cg16418810	-0.115434524	0.001668494	0.013248169	EVC
cg17460447	-0.151332738	0.00894431	0.048299311	EVC
cg22473770	-0.104371814	0.000316067	0.003490006	EVI2A
cg23352695	-0.181916445	6.42e-14	1.82e-10	EVI2A
cg01134183	0.022596514	0.003470694	0.023452885	FAAH
cg06911238	0.198922922	2.11e-06	5.75e-05	FAAH
cg07168328	0.119371444	2e-05	0.000363549	FAAH
cg12671744	0.11143201	1.13e-05	0.000226754	FAAH
cg16267850	0.063387337	4.27e-06	0.000102423	FAAH
cg18261491	0.085739889	3.14e-06	7.98e-05	FAAH
cg25706281	0.068452233	2.62e-05	0.000453829	FAAH
cg04177684	0.127614793	3.55e-07	1.37e-05	FAM83E
cg10772322	0.189046734	1.56e-05	0.000296681	FAM83E
cg20082196	0.17434632	4.59e-05	0.000720693	FAM83E
cg27530053	0.115300686	0.008858816	0.04794984	FAM83E
cg25406989	-0.0888068	0.000199907	0.002402479	FBN1
cg07356342	-0.083890778	4.21e-06	0.000101306	FCER1G

cg20609803	-0.094473946	3.66e-09	4.25e-07	FCER1G
cg26394055	-0.065544302	0.000261937	0.002999199	FCER1G
cg12643083	-0.113650241	0.001460219	0.011927523	FCGR2A
cg01335180	-0.059768502	0.003063648	0.021286661	FCGR2B
cg03105929	-0.125822265	0.000175028	0.002158498	FCGR2B
cg04094791	-0.033310686	0.006943982	0.039860691	FCGR2B
cg10815343	-0.068434002	2.33e-06	6.23e-05	FCGR2B
cg13139730	-0.034200644	0.004668508	0.029458205	FCGR2B
cg17508302	-0.134607074	1.03e-06	3.23e-05	FCGR2B
cg23270415	-0.042416796	8.91e-06	0.000187039	FCGR2B
cg13912027	-0.167186584	2.14e-05	0.00038297	FCHSD2
cg06883949	0.006993046	2.76e-06	7.16e-05	FGFR3
cg19870628	0.019875332	5.92e-06	0.000133859	FGFR3
cg21311834	0.231673987	3.6e-08	2.3e-06	FGFR3
cg02294302	0.116843876	1.37e-05	0.000267108	FOXD2
cg06611075	0.008677944	0.001534647	0.012401202	FOXD2
cg16657448	0.011196552	0.000539297	0.005369988	FOXD2
cg23659056	0.130945582	0.006246363	0.036791313	FOXD2
cg26518431	0.124165844	0.005014186	0.031125205	FOXD2
cg15800907	-0.122113167	4.37e-06	0.000104369	FPR3
cg08421900	0.005146575	2.3e-05	0.000407569	FRAT2
cg13680696	0.041871448	0.000972758	0.008624821	FRAT2
cg19105245	0.008478691	0.001393327	0.011486063	FRAT2
cg19649259	0.002448071	0.003133202	0.021665277	FRAT2
cg00091633	-0.082088615	3.72e-08	2.35e-06	FSTL1
cg13408152	-0.094537438	1.54e-05	0.00029276	FSTL1

cg20114394	-0.040780199	0.002706183	0.019315807	FSTL1
cg00480115	0.196346909	9.42e-09	8.41e-07	FXYD3
cg01408817	0.125052892	2.22e-09	2.95e-07	FXYD3
cg02633817	0.1869328	1.95e-09	2.68e-07	FXYD3
cg02704949	0.123342716	8.48e-10	1.5e-07	FXYD3
cg03322974	0.008406406	9.9e-06	0.000204031	FXYD3
cg21122474	0.101912048	4.57e-08	2.75e-06	FXYD3
cg21304163	0.171410948	2.87e-09	3.53e-07	FXYD3
cg02816367	-0.138922148	0.003707223	0.024682313	FYN
cg05517541	-0.079904626	0.000298877	0.003336849	FYN
cg08130572	-0.200058011	1.14e-11	7.35e-09	FYN
cg14482998	-0.102865383	6.05e-09	6.08e-07	FYN
cg01480180	-0.072285067	0.000111392	0.001490182	FZD1
cg08714590	-0.092262137	5.61e-05	0.000849722	FZD1
cg17497608	-0.096832366	3.77e-05	0.000611898	FZD1
cg26447413	-0.107587705	0.004838525	0.030269548	GAS1
cg03619083	0.135761512	8.23e-07	2.68e-05	GCDH
cg07556193	0.111295535	0.006985432	0.040033063	GDPD2
cg25685838	0.106061012	0.007326822	0.041504894	GDPD2
cg00008544	-0.053651892	9.79e-06	0.000202092	GFPT2
cg09838217	-0.101925591	0.002394819	0.017574656	GFPT2
cg01074657	0.069819468	0.000326275	0.003577098	GIPC2
cg04912843	0.091184472	4.71e-07	1.72e-05	GIPC2
cg09107315	0.083432883	8.38e-07	2.72e-05	GIPC2
cg09662920	0.113862381	2.87e-08	1.94e-06	GIPC2
cg09826056	0.119326032	8.31e-07	2.71e-05	GIPC2

cg19766489	0.149080155	1.48e-06	4.33e-05	GIPC2
cg24496666	0.060077152	6.99e-06	0.000153367	GIPC2
cg25288420	0.075479994	1.55e-05	0.000295201	GIPC2
cg03038418	0.105282202	5.73e-07	2.01e-05	GNA11
cg13960192	0.052460123	2.36e-05	0.000416216	GNA11
cg00028829	0.010279274	0.005902452	0.03522249	GOT2
cg06665322	0.194722472	1.26e-06	3.79e-05	GPA33
cg00620452	0.066487919	9.87e-05	0.00134982	GPD1L
cg19143336	0.120782421	7.61e-08	4.07e-06	GPD1L
cg19409588	0.003243707	0.000180401	0.002211263	GPD1L
cg21145686	0.080371178	0.000658748	0.006318589	GPD1L
cg19755435	-0.108503551	4.65e-07	1.7e-05	GPR65
cg09161043	-0.073569965	0.003866563	0.025490364	GPX7
cg18087326	-0.050567926	0.003155061	0.02178576	GPX7
cg18755653	0.003683892	0.006027006	0.035787327	HADH
cg02311725	-0.075228269	0.002233547	0.016635259	HCK
cg00141162	-0.114684365	9e-07	2.88e-05	HCLS1
cg02167021	-0.148642926	3.98e-06	9.66e-05	HCLS1
cg06577710	-0.116364293	3.61e-06	8.9e-05	HCLS1
cg01378515	-0.070575445	3.4e-05	0.000561696	HEG1
cg03440673	-0.094801136	0.005742908	0.034507543	HEG1
cg16143049	-0.069645783	0.000774135	0.007186369	HEG1
cg23174662	-0.096525322	0.000543876	0.005409196	HIF1A
cg02261294	0.066014691	0.007945177	0.044161988	HNRNPAB
cg02370807	0.009864433	0.000670948	0.006414167	HNRNPAB
cg06538757	0.012262158	4.92e-05	0.000763416	HNRNPAB

cg07868885	0.00853327	0.003669136	0.024486356	HNRNPAB
cg08583763	0.016308761	1.79e-06	5.03e-05	HNRNPAB
cg10038259	0.006591194	0.004506142	0.028660152	HNRNPAB
cg22125717	0.120193709	0.000715412	0.006749315	HNRNPAB
cg01443318	0.161464138	3.1e-06	7.89e-05	HSD11B2
cg07545640	0.004179551	0.001546887	0.012479487	HSD11B2
cg07724674	0.011175522	0.000352685	0.003810425	HSD11B2
cg20981893	0.01829476	0.000134929	0.001743298	HSD11B2
cg27130954	0.031035632	3.15e-08	2.08e-06	HSD11B2
cg00413099	0.00310086	0.00070281	0.006654837	ID1
cg00494337	0.149926454	3.1e-07	1.23e-05	ID1
cg03154513	0.012883586	0.000288912	0.00324581	ID1
cg09923107	0.088225616	3e-05	0.00050686	ID1
cg21626886	0.023121014	7.63e-06	0.000164681	ID1
cg04690927	-0.232185307	0.000171704	0.002125384	IGFBP3
cg07910986	-0.130018739	0.000755507	0.007050491	IGFBP3
cg08541297	-0.129655345	0.002924676	0.020539424	IGFBP3
cg08831744	-0.150306319	0.002037323	0.015482411	IGFBP3
cg09619271	-0.170312435	0.000333518	0.003639782	IGFBP3
cg10094651	-0.231604472	0.000177115	0.002179146	IGFBP3
cg16447589	-0.17126944	0.00931167	0.049793113	IGFBP3
cg16460681	-0.161524951	0.005915883	0.035278227	IGFBP3
cg23455440	-0.180137288	0.001051214	0.009173914	IGFBP3
cg24772240	-0.140720821	0.000121997	0.001605439	IGFBP3
cg24942272	-0.173253673	0.001580683	0.012694212	IGFBP3
cg26434048	-0.196605598	2.25e-07	9.47e-06	IGFBP3

cg03635766	-0.047486243	0.000884466	0.007998953	IGFBP4
cg00790071	-0.102639045	0.000204262	0.00244577	IL1R1
cg05886087	-0.127307627	1.36e-06	4.04e-05	IL1R1
cg27598107	-0.086364348	0.000847278	0.007725628	IL1R1
cg11926473	0.041749232	0.001065631	0.009276233	IMPA2
cg07914866	-0.162664832	0.00031685	0.003496214	IRAK3
cg12866960	-0.082293796	1.73e-09	2.47e-07	IRAK3
cg18177616	-0.119669475	0.0027056	0.019312865	IRAK3
cg02419321	-0.136108152	0.00035592	0.00383825	ITGA5
cg23795217	-0.156611568	0.000143425	0.001834036	ITGA5
cg09326409	-0.148878071	0.003747898	0.024886601	ITGAM
cg15337006	-0.077299933	1.92e-05	0.00035104	ITGAM
cg22490695	-0.093284594	0.002703677	0.019299829	ITGAM
cg13538571	-0.054802197	0.002043085	0.015516451	ITGBL1
cg01648999	0.082425418	0.000111497	0.001491284	KBTBD11
cg08867707	0.147249599	0.001310112	0.010942189	KBTBD11
cg09689137	0.095735216	0.000415144	0.004351683	KBTBD11
cg20910202	0.08755285	0.007561748	0.042531854	KBTBD11
cg23426958	0.063098295	0.001684391	0.013348971	KBTBD11
cg25021970	0.067372469	0.000705002	0.006672084	KBTBD11
cg27126872	0.090262092	6.65e-06	0.000147102	KBTBD11
cg07953201	-0.129312387	0.000490897	0.004977649	KIF26B
cg11912591	-0.062418507	0.002073087	0.015694869	KIF26B
cg15561613	-0.137097218	3.71e-07	1.42e-05	KIF26B
cg21301514	-0.113439195	3.44e-06	8.58e-05	KIF26B
cg26072254	-0.055222414	0.002250624	0.016733808	KIF26B

cg23082393	-0.044733829	0.002536148	0.018378706	LAMA4
cg09803764	-0.103734425	1.79e-08	1.36e-06	LAMB1
cg12689670	-0.158148063	1.59e-08	1.25e-06	LAMC1
cg26809372	-0.033006232	0.005600527	0.033845599	LAMC1
cg17227967	-0.09771243	0.004085769	0.026600313	LAMP5
cg08463932	-0.048082039	0.004182441	0.027070391	LAPTM5
cg10001720	-0.096175135	8.17e-09	7.57e-07	LAPTM5
cg12732155	-0.093527794	0.000712902	0.006731315	LAPTM5
cg15459165	-0.071630644	0.000149345	0.001897846	LAPTM5
cg19919590	-0.103098971	0.001710848	0.013513352	LAPTM5
cg24459792	-0.126178091	8.47e-07	2.74e-05	LAPTM5
cg09451413	-0.101435911	0.000502272	0.005070144	LCP2
cg09672233	-0.092931561	8.95e-05	0.001247496	LCP2
cg11528914	-0.11255277	5.04e-05	0.000779172	LCP2
cg17127769	-0.120433863	3.43e-05	0.000566642	LCP2
cg10867751	0.043831315	7.12e-05	0.001033464	LDLRAP1
cg19759804	0.064887249	5.13e-08	3e-06	LDLRAP1
cg04420917	0.123819574	8.46e-08	4.42e-06	LGALS4
cg06394229	0.119779211	3.77e-08	2.38e-06	LGALS4
cg16731016	0.156602503	1.9e-07	8.3e-06	LGALS4
cg19419519	0.150013928	1.43e-09	2.16e-07	LGALS4
cg26510945	0.134749384	0.00030896	0.003425001	LGALS4
cg09719124	-0.073629346	4.98e-06	0.000116103	LMCD1
cg14455403	-0.064019571	0.000741612	0.006945696	LMCD1
cg26083045	-0.049430573	0.005753339	0.034553448	LMCD1
cg12183875	0.19913235	1.02e-07	5.08e-06	LRRC31

cg10489463	-0.109819156	0.00368689	0.024574613	LTBP1
cg13213009	-0.127294252	0.000132711	0.001719513	LY96
cg23732024	-0.065098851	0.007499934	0.042256929	LY96
cg04000234	-0.099670626	0.000641078	0.006179926	MACF1
cg08456420	-0.069532482	4.24e-05	0.000674368	MACF1
cg18647268	-0.054823448	0.000735912	0.006904664	MACF1
cg21808448	-0.021527542	0.001033501	0.009050249	MACF1
cg22367631	-0.101947735	0.003852155	0.025411877	MACF1
cg00965748	-0.121637598	0.000186043	0.00226612	MAFB
cg16844989	-0.113240631	0.006403948	0.03749157	MAFB
cg27493965	-0.024738239	0.000693655	0.006583415	MAN2B1
cg25182066	-0.059204639	0.001320631	0.011010542	MAP3K8
cg01611777	-0.089142734	0.000635042	0.006135179	MAP4K4
cg25248045	-0.042302469	0.000678152	0.006467585	MAP4K4
cg02303324	0.08629019	0.000784937	0.007266409	MAP7
cg11114242	0.262091813	3.2e-09	3.84e-07	MAP7
cg12963560	0.079613105	6.24e-08	3.49e-06	MAP7
cg18872215	0.055403112	0.000584873	0.005737687	MAP7
cg19555986	0.042983053	0.002034939	0.015469824	MAP7
cg20026346	0.141833099	1.04e-06	3.25e-05	MAP7
cg20481343	0.004951951	0.004036346	0.026347843	MAP7
cg21462732	0.010736088	0.000712405	0.006727671	MAP7
cg24401026	0.232514634	6.82e-10	1.29e-07	MAP7
cg24584345	0.167330852	1.81e-06	5.08e-05	MAP7
cg00764369	-0.053240068	2.48e-06	6.57e-05	MEG3
cg04291079	-0.044013666	0.000277302	0.003140054	MEG3

cg04304932	-0.091228297	1.04e-05	0.000212779	MEG3
cg04576764	-0.091237411	0.000218613	0.002586018	MEG3
cg05711886	-0.075218556	0.003612222	0.024183957	MEG3
cg09280976	-0.07907408	0.002628305	0.018886404	MEG3
cg10515315	-0.064985251	0.003376154	0.022949912	MEG3
cg10943497	-0.06401956	6.18e-09	6.18e-07	MEG3
cg11110759	-0.093782344	0.000344066	0.003732369	MEG3
cg12967319	-0.108911489	0.000263642	0.003015202	MEG3
cg14034270	-0.043537157	0.001155275	0.009893271	MEG3
cg14123427	-0.078944042	0.001124929	0.00968471	MEG3
cg14245102	-0.108469834	0.004140347	0.026865113	MEG3
cg15373285	-0.076850336	0.00571532	0.034380376	MEG3
cg15419911	-0.081342143	0.001056214	0.009208627	MEG3
cg23870378	-0.074040611	0.001884826	0.01458258	MEG3
cg26374305	-0.096071746	0.000156846	0.001975305	MEG3
cg04875062	-0.080659221	1.68e-07	7.54e-06	MFAP2
cg00961326	0.012820767	0.006999298	0.040090176	MMP15
cg16181803	0.008949616	7.54e-06	0.000163087	MMP15
cg20566643	0.008117261	0.001264803	0.01064095	MMP15
cg20722590	0.08442776	1.61e-05	0.000303727	MMP15
cg20751926	0.069971357	4.42e-05	0.000698275	MMP15
cg24306779	0.015973781	0.001218733	0.010327954	MMP15
cg27208052	0.054775683	0.000208378	0.002487154	MMP15
cg08133699	-0.02318553	0.00324695	0.022267336	MMP2
cg09530163	-0.12205862	0.003968161	0.026007342	MMP2
cg22950163	-0.079336185	0.007084858	0.040461188	MMP2

cg12531542	0.198645313	1.78e-05	0.000330409	MOGAT2
cg15955277	0.192031784	2.79e-06	7.23e-05	MOGAT2
cg20938170	0.101047811	2.27e-07	9.57e-06	MOGAT2
cg25255988	0.124604655	1.6e-07	7.25e-06	MOGAT2
cg02179764	0.225085024	1.06e-07	5.24e-06	MPST
cg04129736	0.01590704	0.000415115	0.004351496	MPST
cg06230247	0.15035132	0.001874515	0.014520717	MPST
cg07494646	0.023751263	9.2e-06	0.000192029	MPST
cg07819160	0.160616904	0.005071409	0.031392341	MPST
cg08727202	0.097593889	1.8e-06	5.07e-05	MPST
cg12253469	0.157111081	0.00628294	0.036960144	MPST
cg17575915	0.15924199	1.17e-09	1.87e-07	MPST
cg11628739	-0.083647649	0.000344103	0.003732673	MRC2
cg08564601	-0.129923951	9.85e-06	0.000203195	MS4A4A
cg18025430	-0.071873074	0.001164672	0.009956523	MS4A4A
cg03055440	-0.037602474	6.22e-05	0.000925301	MS4A6A
cg24026212	-0.052788181	0.000327771	0.003590418	MS4A6A
cg22771999	-0.186467929	5.67e-06	0.000129235	MSN
cg01375994	-0.062504099	0.00766799	0.042988089	MXRA5
cg09293286	-0.038076341	1.16e-05	0.00023223	MXRA5
cg13581022	-0.051310869	2.06e-06	5.65e-05	MXRA5
cg12472603	-0.087577699	4.01e-07	1.51e-05	MXRA7
cg14042121	-0.065536791	3.31e-05	0.000550279	MXRA7
cg09975715	-0.017925755	0.008873378	0.048012056	MYH10
cg25921609	-0.039775869	7e-05	0.001020035	MYH10
cg01514487	0.147549765	1.54e-06	4.47e-05	MYO1A

cg09541248	0.131750683	2.67e-08	1.84e-06	MYO1A
cg11276093	-0.100487318	5.85e-07	2.05e-05	MYOF
cg13669036	-0.097706051	7.53e-06	0.000162956	MYOF
cg14428166	-0.051989752	1.85e-05	0.000340554	MYOF
cg18991240	-0.152274808	1.1e-08	9.45e-07	NAGK
cg14494313	0.185644126	8.47e-05	0.001193196	NAT2
cg18736775	0.238720503	2.24e-05	0.000398436	NAT2
cg09472600	-0.114008101	7.21e-09	6.91e-07	NCF2
cg00950244	0.153243186	4.16e-05	0.000664541	NDUFAF4
cg11787828	0.024243429	0.005167608	0.031842006	NDUFAF4
cg11087503	-0.044121326	0.005154259	0.031782499	NID2
cg16695483	-0.111336963	0.008542482	0.046665354	NID2
cg25685519	-0.143569774	2.16e-06	5.86e-05	NID2
cg14520913	-0.074655952	0.002042491	0.015513055	NNMT
cg01575652	0.0795892	0.001042524	0.009113136	NOL12
cg14370507	0.207995782	1.27e-08	1.05e-06	NOL12
cg19884546	0.013543036	0.000665329	0.006369522	NOL12
cg03793270	-0.097788376	0.000165954	0.002066393	NOX4
cg17063929	-0.114862299	0.00083004	0.007597112	NOX4
cg01885839	-0.024337631	0.000283506	0.003196141	NREP
cg08651538	-0.049807249	0.008071129	0.044693276	NREP
cg25763127	-0.064417308	0.000148421	0.001887956	NRP1
cg27270412	-0.137956492	5.87e-05	0.000882851	NRP1
cg00557402	0.162952847	0.001251944	0.010554519	NXPE4
cg21833776	0.133025857	0.008261179	0.045479948	NXPE4
cg22223402	0.191486665	5.77e-05	0.000869924	NXPE4

cg05524246	-0.131334811	0.004372619	0.028000042	OLFML2B
cg02390103	-0.115533744	0.001398145	0.011517394	OSMR
cg03138091	-0.14201413	1.2e-05	0.000238972	OSMR
cg05485663	-0.109050166	0.004540672	0.028832549	OSMR
cg15599832	-0.076753988	0.00011896	0.001572473	OSMR
cg17528648	-0.167653489	0.001775926	0.013918989	OSMR
cg19609242	-0.218510994	7.96e-08	4.22e-06	OSMR
cg22473846	-0.104848974	0.004046507	0.026401986	OSMR
cg26475085	-0.09324139	0.003961187	0.025970761	OSMR
cg04865264	0.028906885	5.62e-05	0.000850919	OVOL2
cg17507897	0.129193285	0.002087778	0.01578286	OVOL2
cg08893575	0.111814642	2.52e-07	1.04e-05	PAK4
cg09506385	0.007763597	3.14e-05	0.000526497	PAK4
cg12406027	0.044752942	1.17e-07	5.69e-06	PAK4
cg24710020	0.064919819	0.00729296	0.041346784	PAK4
cg07840446	-0.079687995	1.88e-05	0.000345742	PALLD
cg08947774	-0.099271944	0.000135591	0.001750763	PALLD
cg17044159	-0.099719143	1.86e-05	0.000342827	PAM
cg00035945	0.046306789	2.24e-07	9.44e-06	PARM1
cg04423976	0.099732954	7.56e-07	2.51e-05	PARM1
cg15871647	0.082472948	5.44e-05	0.000828587	PARM1
cg01535080	-0.119899146	0.004935231	0.030740605	PDE10A
cg04249522	-0.16406998	0.005904088	0.03522964	PDE10A
cg13351249	-0.100512368	0.001355086	0.011242187	PDE10A
cg16051195	-0.130337753	9.89e-06	0.000203768	PDE10A
cg17712241	-0.039077748	0.001990319	0.015211306	PDE10A

cg04117986	-0.056297182	2.55e-05	0.00044358	PDGFRB
cg04173992	-0.056027048	1.11e-06	3.42e-05	PDGFRB
cg12727795	-0.116543653	0.00450249	0.028641528	PDGFRB
cg15924831	-0.026776313	0.003470171	0.023450947	PDGFRB
cg16429070	-0.116012273	0.000187381	0.00227926	PDGFRB
cg25110734	-0.114885134	0.000194886	0.002354218	PDGFRB
cg25440811	-0.124036107	0.008834257	0.047848098	PDGFRB
cg17815886	-0.046358051	1.19e-07	5.76e-06	PDLIM5
cg19674166	-0.029518334	0.007037385	0.040254499	PDLIM5
cg04044188	0.227723744	2.88e-06	7.43e-05	PDSS1
cg25196348	0.064586333	3.15e-05	0.000527884	PDSS1
cg09799714	0.222256295	4.39e-08	2.66e-06	PDZD3
cg01348757	0.14366513	0.003343615	0.022782287	PEX11A
cg08749443	0.159646008	7.49e-05	0.001078173	PEX11A
cg11526413	0.218840434	2.96e-07	1.18e-05	PEX11A
cg14154973	0.0313475	6.34e-05	0.000940306	PEX11A
cg17732044	0.023938603	6.96e-05	0.00101503	PEX11A
cg23328050	0.029771949	4.21e-05	0.00067105	PEX11A
cg24252910	0.006280033	0.006615955	0.038421527	PEX11A
cg00902516	-0.043486562	0.003359401	0.022868208	PFKFB3
cg03261682	-0.03807574	0.000364069	0.003911432	PFKFB3
cg05014727	-0.138083379	2.4e-07	9.99e-06	PFKFB3
cg08994060	-0.11853361	1.94e-07	8.45e-06	PFKFB3
cg12235073	-0.060164757	1.3e-11	7.87e-09	PFKFB3
cg16179674	-0.038628868	0.000213415	0.002535811	PFKFB3
cg17545652	-0.063655116	2.15e-05	0.000384417	PFKFB3

cg26262157	-0.104413948	4.05e-07	1.52e-05	PFKFB3
cg27545615	-0.068477839	1.41e-05	0.000272871	PFKFB3
cg15928480	0.125068502	0.001930149	0.014846279	PIGR
cg03885270	-0.065781258	4.13e-05	0.000661304	PIP4K2A
cg07171687	-0.070001289	1.97e-05	0.000358814	PIP4K2A
cg09216670	-0.101983645	0.002361182	0.01738098	PIP4K2A
cg09273683	-0.108589469	2.54e-08	1.77e-06	PIP4K2A
cg14215711	-0.134425578	3.7e-08	2.34e-06	PIP4K2A
cg17277615	-0.064178307	8.36e-05	0.001179587	PIP4K2A
cg20641026	-0.037528841	0.000779199	0.007223755	PIP4K2A
cg25073089	-0.019958227	0.006499513	0.037921008	PIP4K2A
cg12488447	0.007171379	0.009287287	0.049690907	PIP5K1B
cg13442709	0.103140978	0.000131338	0.001705019	PIP5K1B
cg13634133	0.02012154	4.1e-05	0.000657166	PIP5K1B
cg02736228	0.170297634	4e-08	2.49e-06	PLEKHA6
cg07010486	0.101898478	4.4e-08	2.66e-06	PLEKHA6
cg08784911	0.107448514	2.33e-05	0.000411561	PLEKHA6
cg09150239	0.014092412	0.000958513	0.00852592	PLEKHA6
cg10094886	0.163590439	5.72e-09	5.85e-07	PLEKHA6
cg13056222	0.07103083	0.002383105	0.017506855	PLEKHA6
cg19403103	0.006641156	0.006332488	0.037184541	PLEKHA6
cg24734365	0.094810401	4.83e-06	0.000113156	PLEKHA6
cg26787863	0.099236183	0.003617343	0.024209243	PLEKHA6
cg00063773	0.081704025	6.37e-06	0.000142052	PLS1
cg05652551	0.232757899	9.48e-09	8.45e-07	PLS1
cg20824294	0.095246718	3.07e-05	0.000516326	PLS1

cg05490233	-0.026539101	0.009093583	0.048914192	PLXND1
cg12415479	-0.125206696	0.000919223	0.008246104	PLXND1
cg16850690	-0.109413971	4.87e-07	1.76e-05	PLXND1
cg19727499	0.124146354	4.26e-07	1.59e-05	PPFIA3
cg13625187	-0.020927133	0.000808299	0.00743731	PPFIBP1
cg20912752	-0.024213277	0.004326093	0.027777284	PPFIBP1
cg04314111	0.182753659	1.08e-06	3.35e-05	PRKCZ
cg11227141	0.180629981	1.97e-09	2.69e-07	PRKCZ
cg12639453	0.084278948	0.000822705	0.007543321	PRKCZ
cg16269144	0.126127576	0.008859634	0.047952101	PRKCZ
cg17023856	0.091485809	0.004968313	0.030900016	PRKCZ
cg17815669	0.118694768	0.002562734	0.018523211	PRKCZ
cg22332339	0.255371017	5.63e-09	5.78e-07	PRKCZ
cg22865720	0.003642371	0.003290717	0.02250115	PRKCZ
cg24035370	0.010410715	0.003144151	0.021724059	PRKCZ
cg10077239	-0.103096733	0.000107585	0.001448707	PRKD1
cg10698355	-0.136692469	0.006516313	0.0379916	PRR16
cg22003366	-0.187743346	0.001459351	0.011922081	PRR16
cg25584626	-0.1462167	0.003694709	0.024616704	PRR16
cg26464221	-0.158557631	0.003351043	0.022822689	PRR16
cg00916255	-0.131186811	2.36e-07	9.84e-06	PRRX1
cg07149609	-0.163985176	0.000586658	0.005752783	PRRX1
cg07957294	-0.096257386	0.000261063	0.002991402	PRRX1
cg09010107	-0.124010779	1.35e-05	0.000263765	PRRX1
cg21914290	-0.140487105	0.000496924	0.005026766	PRRX1
cg24376434	-0.102681618	2.61e-08	1.8e-06	PRRX1

cg00031402	-0.04554563	0.000513122	0.005157457	PSAP
cg08788055	-0.068494362	0.001565256	0.012592022	PTGIS
cg10772290	-0.097091219	0.003565289	0.023947467	PTGIS
cg01629007	-0.086861074	4.99e-06	0.000116304	PXDN
cg06599209	-0.163476788	1.05e-05	0.000214007	PXDN
cg08534653	-0.145710913	0.00884718	0.047901063	PXDN
cg09618102	-0.164260033	0.007188827	0.040891036	PXDN
cg12780678	-0.155106864	0.002616358	0.018820683	PXDN
cg19517718	-0.112227129	0.001499266	0.012176408	PXDN
cg21647182	-0.04969819	0.005257278	0.032264529	PXDN
cg25181651	-0.155255158	0.002422701	0.017734577	PXDN
cg26691059	0.216243203	1.24e-09	1.95e-07	PXMP2
cg17982102	-0.136477758	9.01e-06	0.000188857	RAB31
cg18456459	-0.15064537	1.37e-05	0.000266876	RAB31
cg17360854	-0.101836667	0.007899433	0.043960163	RBMS1
cg20472746	-0.115120629	2.35e-06	6.28e-05	RGCC
cg02586212	-0.075820979	8.84e-06	0.000185905	RGS1
cg04562217	-0.128753375	6.02e-05	0.000900619	ROBO1
cg08661007	-0.113978837	0.004340965	0.027849752	ROBO1
cg11980129	-0.122090529	0.000342384	0.003718604	ROBO1
cg15325658	-0.103690315	0.000149402	0.001898511	ROBO1
cg21865845	-0.133359615	8.79e-07	2.83e-05	ROBO1
cg07680533	0.091102395	6.22e-07	2.15e-05	SELENBP1
cg16911672	0.181899172	0.000377562	0.004030372	SELENBP1
cg17759475	0.148671679	1.64e-09	2.37e-07	SELENBP1
cg18515587	0.158272743	4.02e-07	1.51e-05	SELENBP1

cg24480379	0.138075722	1.96e-07	8.51e-06	SELENBP1
cg24486037	0.165375403	9.27e-08	4.75e-06	SELENBP1
cg26065909	0.162902272	2.51e-07	1.03e-05	SELENBP1
cg21542842	-0.103186833	1.48e-07	6.79e-06	Sep-11
cg01975495	-0.079819935	6.88e-05	0.001005905	SERPINE1
cg17968347	-0.059453535	0.000347443	0.00376189	SERPINE1
cg11692409	-0.07892156	3e-06	7.68e-05	SERPINF1
cg24214470	-0.07118616	8.95e-06	0.000187719	SERPINF1
cg27102649	-0.083087473	3.35e-07	1.3e-05	SERPINF1
cg19453665	-0.026273465	0.007931149	0.044104441	SERPINH1
cg26104986	-0.070264853	0.000594207	0.005812704	SERPINH1
cg26679912	-0.135125791	0.000540917	0.005384766	SFRP4
cg12122241	-0.068203195	7.75e-06	0.000166906	SIRPA
cg14613594	-0.094209407	0.000168188	0.002089802	SIRPA
cg02794695	-0.116510982	7.42e-06	0.000161116	SLA
cg04756252	-0.111746257	8.19e-05	0.001159924	SLA
cg21653105	-0.047978957	3.14e-06	7.97e-05	SLA
cg22801799	-0.094855862	0.000171987	0.002128218	SLA
cg04275881	-0.124438496	6.74e-08	3.71e-06	SLAMF8
cg06764092	-0.120100716	9.2e-05	0.001275839	SLAMF8
cg07625783	-0.15481556	2.44e-08	1.71e-06	SLAMF8
cg17972058	-0.065494719	0.00024917	0.002878779	SLAMF8
cg15355952	-0.051139243	2.17e-08	1.57e-06	SLC1A3
cg21050001	-0.034943206	0.006404563	0.037494616	SLC1A3
cg04996020	0.199514525	9.61e-07	3.04e-05	SLC26A3
cg17268483	0.115342588	0.004784184	0.03000836	SLC27A2

cg06567290	0.136081683	9.42e-06	0.000195737	SLC37A4
cg08998953	0.016842072	0.000139891	0.001796264	SLC37A4
cg17791936	0.092907713	0.008664715	0.047171079	SLC37A4
cg21561712	0.081942495	0.000749308	0.007004024	SLC37A4
cg01347702	0.152762715	1.78e-08	1.36e-06	SLC44A4
cg03045620	0.087153561	2.28e-06	6.14e-05	SLC44A4
cg04021562	0.063920997	6.08e-08	3.42e-06	SLC44A4
cg04567302	0.184896387	8.68e-09	7.92e-07	SLC44A4
cg05686323	0.120961022	0.001934579	0.014873704	SLC44A4
cg07185041	0.148221279	2.32e-07	9.73e-06	SLC44A4
cg07357081	0.084718674	2.3e-06	6.18e-05	SLC44A4
cg07363637	0.123175065	2.66e-07	1.08e-05	SLC44A4
cg07546508	0.11122955	6.23e-05	0.000927021	SLC44A4
cg09298971	0.087066376	0.007465083	0.042107444	SLC44A4
cg11726150	0.130910734	9.1e-07	2.91e-05	SLC44A4
cg11943056	0.146805789	1.72e-06	4.87e-05	SLC44A4
cg15821546	0.106310684	0.00421004	0.027206345	SLC44A4
cg16553272	0.117821092	3.08e-06	7.84e-05	SLC44A4
cg18856043	0.108732367	9.23e-07	2.94e-05	SLC44A4
cg19117051	0.117385658	7.67e-06	0.000165341	SLC44A4
cg23431175	0.092294133	5.14e-07	1.84e-05	SLC44A4
cg24529722	0.104423403	4.45e-07	1.64e-05	SLC44A4
cg24707219	0.137120373	6.78e-08	3.72e-06	SLC44A4
cg27003765	0.129644581	8.14e-08	4.29e-06	SLC44A4
cg27005847	0.137380284	1.43e-07	6.62e-06	SLC44A4
cg00292986	0.026336766	0.001396213	0.011504461	SLC9A2

cg01272393	0.142646043	0.002504856	0.018200069	SLC9A2
cg11915641	0.026851992	0.006172237	0.036463271	SLC9A2
cg20050113	0.113390089	1.49e-08	1.19e-06	SLC9A2
cg21697381	-0.095299739	0.002013215	0.015341005	SLFN12
cg24447042	-0.114076716	0.008786387	0.047654645	SMARCA1
cg26010110	-0.188865417	0.009010294	0.048569748	SMARCA1
cg01041405	0.200269441	4.88e-05	0.000758975	SMPD3
cg03412735	0.046854471	0.003755301	0.024926986	SMPD3
cg18497162	0.129572427	7.56e-08	4.05e-06	SPHK2
cg09054633	-0.115810019	0.001916618	0.014770885	SPOCK1
cg18263365	-0.100089929	0.004680251	0.029511613	SPOCK1
cg18603028	-0.117752251	0.001915359	0.014763476	SPOCK1
cg02851793	-0.115628956	4.37e-06	0.000104432	SRGN
cg17342283	-0.108577524	0.000201897	0.002422179	SRGN
cg18278184	-0.097468123	0.007461358	0.042095461	SRGN
cg26522946	-0.141593261	6.25e-05	0.000928952	SRGN
cg27208307	-0.130701601	0.000225493	0.002652899	SRGN
cg01389506	-0.043459562	2.48e-06	6.56e-05	SSH1
cg01791669	-0.046701592	3.98e-06	9.66e-05	SSH1
cg07700680	-0.022496955	0.002106308	0.015898573	SSH1
cg07887608	-0.020278079	2.74e-05	0.000470293	SSH1
cg11114313	-0.040627755	1.2e-08	1.01e-06	SSH1
cg11699334	-0.056962089	0.000473688	0.004833347	SSH1
cg13033858	-0.148484614	6.23e-07	2.15e-05	SSH1
cg14854315	-0.031100585	0.001070424	0.00930806	SSH1
cg17446956	-0.051516053	0.000103496	0.001402847	SSH1

cg19256314	-0.093163963	0.000839577	0.007669692	SSH1
cg21224380	-0.042808291	0.000763169	0.007105254	SSH1
cg21616405	-0.047889219	0.001784879	0.013972976	SSH1
cg22522688	-0.055642091	0.000484045	0.004919638	SSH1
cg23126152	-0.097003395	2.07e-09	2.79e-07	SSH1
cg25270574	-0.030191788	0.0058154	0.034822153	SSH1
cg26508200	-0.032606918	0.007881671	0.043895241	SSH1
cg27553890	-0.098639513	1.65e-12	1.93e-09	SSH1
cg04077662	-0.097295689	0.002823389	0.019985446	ST6GALNAC5
cg06201642	-0.138915009	0.001157056	0.009905351	ST6GALNAC5
cg09511846	-0.173921607	0.00694338	0.039857816	ST6GALNAC5
cg13463054	-0.179224325	0.001368513	0.011328196	ST6GALNAC5
cg13823136	-0.137191698	0.005833378	0.034902365	ST6GALNAC5
cg15100100	-0.158710554	0.007376443	0.04173227	ST6GALNAC5
cg16966815	-0.104741665	0.003003594	0.020964284	ST6GALNAC5
cg14365564	-0.029412132	0.004376269	0.028016344	STOM
cg12158889	0.098586558	1.88e-05	0.000345361	SUCLG1
cg07438401	0.123270175	2.98e-08	1.99e-06	SUCLG2
cg07703372	0.009558228	0.00260688	0.018770455	SUCLG2
cg13668339	0.173648911	8.84e-07	2.85e-05	SUCLG2
cg16414852	0.165896381	6.13e-07	2.12e-05	SULT1B1
cg23824376	-0.131220347	0.000137705	0.001772858	TENM3
cg27540367	-0.07845929	7.52e-07	2.5e-05	TGFB1
cg08470742	-0.173785849	0.008517192	0.046557382	THBS2
cg17608103	-0.176486619	0.001174174	0.010022387	THBS2
cg21652958	-0.194667161	0.002700437	0.01928296	THBS2

cg23691781	-0.097700431	3.68e-08	2.33e-06	THEMIS2
cg00156427	-0.080990786	0.002304022	0.017045531	THY1
cg12508624	-0.077771236	0.003690451	0.024593987	THY1
cg13524082	-0.132017558	0.000638268	0.006158438	THY1
cg16566400	-0.123128896	0.004267495	0.027491511	THY1
cg01263877	0.063279338	0.006904857	0.039680404	TJP3
cg02489438	0.037252714	0.002310855	0.017083821	TJP3
cg10733063	0.073354273	0.008115007	0.044869199	TJP3
cg04120171	-0.163376622	0.007298107	0.041371816	TM6SF1
cg09682213	-0.09222304	0.00389696	0.02564935	TM6SF1
cg01157146	0.126553466	0.000661451	0.006339908	TMPRSS2
cg02613803	0.01340107	0.000843712	0.007700033	TMPRSS2
cg12384236	0.029010936	0.003933156	0.025832418	TMPRSS2
cg13489049	0.103759664	4.72e-08	2.81e-06	TMPRSS2
cg14982276	0.04023119	0.006395364	0.037455709	TMPRSS2
cg16084872	0.038591879	1.82e-05	0.000335516	TMPRSS2
cg24901042	0.044203805	0.000434304	0.004512749	TMPRSS2
cg26309194	0.048505249	0.000327248	0.003585871	TMPRSS2
cg01981433	-0.025119162	0.001689289	0.013380292	TNFAIP3
cg18287768	-0.036720648	0.000794083	0.007331202	TNFAIP3
cg18054943	0.054872648	0.000148777	0.001891598	TNK1
cg18632631	0.115600729	2.66e-05	0.000458892	TNK1
cg25499099	0.037630755	0.000340355	0.003700483	TNK1
cg06041363	0.07463485	0.000142792	0.001827832	TTC38
cg08796741	0.01131577	0.000683033	0.006502231	TTC38
cg16674248	0.17325242	4.15e-09	4.64e-07	TTC38

cg00770085	0.024268442	0.001717828	0.013555646	TTC39A
cg03814321	0.101317082	0.003240371	0.022234881	TTC39A
cg05132999	0.16008967	1.9e-06	5.29e-05	TTC39A
cg07591515	0.136170043	2.72e-05	0.000467718	TTC39A
cg10653240	0.047396092	0.000788607	0.007292375	TTC39A
cg20942910	0.107946565	0.009298771	0.049734882	TTC39A
cg23271269	0.023592361	0.000271375	0.003084629	TTC39A
cg26351104	0.018431638	2.19e-06	5.92e-05	TTC39A
cg09177949	0.003085246	0.005849486	0.034977358	TTLL12
cg01202666	-0.116503661	0.001477033	0.012035077	TWIST1
cg02400740	-0.164499801	0.000485965	0.004936232	TWIST1
cg04917226	-0.123704483	0.001347648	0.011191029	TWIST1
cg05380019	-0.11532784	0.005383943	0.032857077	TWIST1
cg06243400	-0.130769915	0.000341931	0.00371491	TWIST1
cg14515453	-0.125271398	8.53e-05	0.001200064	TWIST1
cg18791205	-0.108632394	0.00686099	0.039490177	TWIST1
cg20121142	-0.152160755	0.000391618	0.004152218	TWIST1
cg23244488	-0.107610989	0.005664874	0.034152129	TWIST1
cg23603376	-0.131671211	0.002944459	0.020645133	TWIST1
cg27013696	-0.108839039	0.000264255	0.003020638	TWIST1
cg14655843	-0.084877935	1.52e-05	0.000289531	UGCG
cg02096633	0.019415573	1.57e-07	7.13e-06	UNC13B
cg06424576	0.016406399	0.001800914	0.01407264	UQCRC1
cg23902361	-0.062251668	0.001143675	0.009813246	VAMP5
cg17771652	-0.128150888	0.008972244	0.048410762	VCAN
cg23991622	-0.18008124	0.004238654	0.027349462	VIM

cg03160740	0.017844679	2.97e-05	0.000501664	VIPR1
cg06783423	0.052963263	9.07e-06	0.000189896	VIPR1
cg09384400	0.113704207	2.75e-05	0.000471194	VIPR1
cg15185458	0.005289861	0.00604943	0.035890812	VIPR1
cg16180367	0.004734128	0.004996059	0.031034081	VIPR1
cg23517013	0.134728157	1.27e-06	3.82e-05	VIPR1
cg25968378	0.004478506	0.00106869	0.009295913	VIPR1
cg12124912	-0.095074612	3.45e-06	8.59e-05	VSIG4
cg00037952	-0.088071411	2.01e-07	8.69e-06	WBP1L
cg03161190	-0.134722	6.7e-07	2.27e-05	WBP1L
cg09038267	-0.043811381	0.000128201	0.001671968	WBP1L
cg14015502	-0.039634747	0.001884039	0.014577958	WBP1L
cg14939082	-0.095631819	0.000305051	0.003390679	WBP1L
cg15227982	-0.117444848	8.97e-07	2.88e-05	WBP1L
cg15615645	-0.109569318	2.87e-08	1.94e-06	WBP1L
cg17740322	-0.107096655	0.001769412	0.013879466	WBP1L
cg17894755	-0.090371482	0.000271849	0.00308873	WBP1L
cg23247968	-0.098453927	8.41e-05	0.001186298	WBP1L
cg25104397	-0.111821122	3.77e-06	9.23e-05	WBP1L
cg26640901	-0.104594274	6.31e-06	0.000140973	WBP1L
cg27517345	-0.100209443	3.07e-05	0.000516419	WBP1L
cg09842053	0.17374387	2.02e-08	1.48e-06	XDH
cg16862361	0.219221188	1.49e-09	2.21e-07	XDH
cg26767897	0.221560511	1.9e-08	1.43e-06	XDH
cg04125273	0.127860208	0.000264552	0.003023081	ZG16
cg06289826	0.085090239	0.000148217	0.001885941	ZG16

cg21710408	-0.071808424	0.006646637	0.038557308	ZNF532
cg20332503	-0.084537156	0.000116924	0.001550258	ZYX

B.8 CRC Over-Represented Hyper/Hypo-Methylated Pathways

Table B.15 GO pathways over-represented in the TCGA-COAD methylation data.

Pathway ID	**Description**	**Count**	**Fold Enrichment**	***p*-adjusted**
GO:0007155	cell adhesion	31	4.647951713	8.32E-09
GO:0030198	extracellular matrix organization	20	7.022415524	5.67E-08
GO:0030574	collagen catabolic process	13	13.9789959	6.03E-08
GO:0001525	angiogenesis	15	4.629125928	0.001972279
GO:0030206	chondroitin sulfate biosynthetic process	6	16.51672131	0.008504295
GO:0035987	endodermal cell differentiation	6	15.29326047	0.010057799
GO:0016477	cell migration	12	4.801372474	0.010057799
GO:0007229	integrin-mediated signaling pathway	9	6.25633383	0.019663004
GO:0030199	collagen fibril organization	6	10.58764187	0.04340959
GO:0045766	positive regulation of angiogenesis	9	5.385887384	0.04458756
GO:0032967	positive regulation of collagen biosynthetic process	5	14.96079829	0.046180941