Applications of Intertumoural, Intratumoural and Intermolecular Heterogeneity for Personalised Medicine in Colorectal Cancer

Katherine Eason

A thesis submitted to the University of London for the degree of
Doctor of Philosophy

Supervised by
Dr Anguraj Sadanandam

The Institute of Cancer Research
University of London
United Kingdom
September 2019

I confirm that the work presented in this thesis is my own, except where information has been derived from other sources or collaborations as listed in the statement overleaf.

Katherine Eason

Statement on collaborative work and jointly-authored publications

Chapter 2

Results from this chapter have been published jointly with co-authors, who made contributions listed below.

> Ragulan, C.*, Eason, K.*, Fontana, E.*, Nyamundanda, G.*, Poudel, P., Lawlor, R. T., Rio, M. D., Si-Lin, K., Siew, T. W., Sclafani, F., Begum, R., Mendes, L. S. T., Martineau, P., Tan, I. B., Cunningham, D. & Sadanandam, A. Analytical Validation of Multiplex Biomarker Assay to Stratify Colorectal Cancer into Molecular Subtypes. *Scientific Reports* **9**, 7665 (2019). * indicates joint first authors.

Sample material from the Montpellier cohort was kindly provided by Dr Maguy Del Rio and Dr Pierre Martineau of Université Montpellier, France; from the Singapore cohort by Dr Koo Si-Lin, Dr Tan Wah Siew, and Dr Iain Beehuat Tan of the National Cancer Centre Singapore and Singapore General Hospital, Singapore; from the RETRO-C cohort by Dr Francesco Sclafani, Ms Ruwaida Begum, Dr Larissa Sena Teixeira Mendes, and Prof. David Cunningham of the Royal Marsden NHS Foundation Trust, UK; and from the INCLIVA-Valencia cohort by Dr Noelia Tarazona and Prof. Andrés Cervantes of the University of Valencia, Spain.

NanoString nCounter hybridisation and data collection was kindly performed by Ms Chanthirika Ragulan and Dr Elisa Fontana of the Institute of Cancer Research, UK.

OriGene microarray data was kindly generated by Ms Chanthirika Ragulan of the Institute of Cancer Research, UK.

RNA-seq data for the Singapore cohort was kindly provided by Dr Iain Beehuat Tan of the Singapore General Hospital, Singapore, with additional advice provided by Mr Yatish Patil of the Institute of Cancer Research, UK.

The gene selection pipeline was designed by Dr Gift Nyamundanda of the Institute of Cancer Research, UK.

Publicly-available data has been used as cited.

All other work is my own.

Chapter 3

FACS and gene expression data from co-cultured cell lines were kindly performed/-collected by Ms Chanthirika Ragulan of the Institute of Cancer Research, UK.

Publicly-available data has been used as cited.

All other work is my own.

Chapter 4

The isBFAC model was designed, derived and implemented by Dr Gift Nyamundanda, Postdoctoral Fellow at the Institute of Cancer Research, London. I made alterations to the implementation for speed and usability.

Publicly-available data has been used as cited.

All other work is my own.

Other publications

Fontana, E., Eason, K., Cervantes, A., Salazar, R. & Sadanandam, A. Context Matters — Consensus Molecular Subtypes of Colorectal Cancer as Biomarkers for Clinical Trials. *Annals of Oncology* **30**, 520–527 (2019)

Cremolini, C., Benelli, M., Fontana, E., [and 19 others, including Eason, K.]. Benefit From Anti-EGFRs in RAS and BRAF Wild-Type Metastatic Transverse Colon Cancer: A Clinical and Molecular Proof of Concept Study. *ESMO Open* **4**, e000489 (2019)

Wagner, S., Vlachogiannis, G., De Haven Brandon, A., [and 20 others, including Eason, K.]. Suppression of Interferon Gene Expression Overcomes Resistance to MEK Inhibition in KRAS-Mutant Colorectal Cancer. *Oncogene* **38**, 1717–1733 (2019)

Heindl, A., Khan, AM., Rodrigues, DN., Eason, K., [and 7 others]. Microenvironmental Niche Divergence Shapes BRCA1-Dysregulated Ovarian Cancer Morphological Plasticity. *Nature Communications* **9**, 3917 (2018)

Vlachogiannis, G., Hedayat, S., Vatsiou, A., [and 40 others, including Eason, K.]. Patient-Derived Organoids Model Treatment Response of Metastatic Gastrointestinal Cancers. *Science* **359**, 920-926 (2018)

Eason, K., Nyamundanda, G. & Sadanandam, A. polyClustR: Defining Communities of Reconciled Cancer Subtypes with Biological and Prognostic Significance. *BMC Bioinformatics* **19**, 182 (2018)

Fontana, E., Homicsko, K., Eason, K. & Sadanandam, A. Molecular Classification of Colon Cancer: Perspectives for Personalized Adjuvant Therapy. *Current Colorectal Cancer Reports* **12**, 296–302 (2016)

Eason, K. & Sadanandam, A. Molecular or Metabolic Reprogramming: What Triggers Tumor Subtypes? *Cancer Research* **78**, 5195-5200 (2016)

Acknowledgements

My first thanks must go to my supervisor Dr Anguraj Sadanandam for his guidance and expertise, and to the ICR who funded my studies. I have met many kind and intelligent people in Dr Sadanandam's team, but particular thanks should go to Gift, Chanthirika and Erica who were all patient enough to lend their time and knowledge to a naive undergraduate back in 2014, and to Elisa, Pawan, Krisha and Varun for asking me difficult but important questions. My gratitude also goes to our international collaborators who shared their precious samples and data.

In the wider ICR, people from the CTU and beyond, particularly Ellen, Catherine, Jonny, both Joshes, both Hannahs, Steve, Suzie, and François, plus many others, and everyone who has danced with me at the pHbar.

My brother Matt, sister-in-law Meg, and niblings Noah and Winnie have kept me grounded by reminding me that even when I'm stressed, at least I've never had two under-fives, and my gran who has kept me humble by reminding me I'm not a "proper doctor".

I couldn't ever give enough words to repay Andy, who never once believed me when I told him I was a brainless imposter, and whose unwavering love has been the safety net that has saved me more than once.

Finally, to my parents Kit and Val, whose unconditional encouragement and support in all aspects of my life has been the underlying reason for any and every success I could ever claim.

Abstract

Colorectal cancer (CRC) is a heterogeneous disease, both at the molecular level and in the context of patients' responses to treatment. Few biomarkers are currently in place that can help to stratify patients in the clinical setting.

This thesis begins by describing inter- and intratumoural transcriptomic heterogeneity in CRC, before extending to the integration of multiomics data for a system-wide view of the pathways active in this disease.

Initially, taking previously described gene expression subtypes of CRC that have prognostic indications and potential associations with patient outcomes/drug responses, I redefined their gene expression signatures to a smaller gene set (measurable on a platform that has previously been approved for clinical use) using a consensus of statistical gene selection and class prediction methods, thus enabling future subtype-based prospective clinical trials. Subtyping with this new gene set and platform was highly accurate against the previous standard, and has the additional benefit that it can also be applied to large archives of formalin-fixed paraffin-embedded tissues for retrospective analyses.

Furthermore, I explored the intratumoural heterogeneity of these subtypes using machine learning techniques and single-cell data, concluding that they co-exist in the vast majority of tumours. Using these subtype sub-populations, I was able to significantly improve prognostic power in survival models versus traditional "bulk" subtyping, and identify subsets of patients that respond best to already-available therapies (as well as those who could be spared unnecessary toxicities). For example, early-stage patients whose tumours were deemed to have a high stem-like subpopulation by computational deconvolution had significantly poorer prognosis than those with a low subpopulation, while no prognostic difference was observed between patients with bulk stem-like verus other bulk subtype tumours. In addition, TA subtype sub-

populations were significantly higher in patients and pre-clinical models of CRC who responded/were sensitive to cetuximab.

Finally, I have used a Bayesian latent variable machine learning framework to integrate multi-omics data (including gene, miRNA and protein expression, methylation, copy number and mutations) and clinicopathological variables from the TCGA CRC database. In this way, I found patterns of co-expression across molecular levels that relate to complex interactions between clinically interpretable covariates. The results from this analysis included novel biomarkers that had significant and context-specific prognostic implications.

Overall, in this thesis, I present several characterisations of CRC's multi-faceted heterogeneity. I demonstrate how the existing transcriptomic CRCAssigner inter-tumoural subtypes can be profiled in a clinically-practicable manner, expand on our understanding of these subtypes by quantifying their co-existence within individual tumours, and move beyond transcriptomics to delineate CRC heterogeneity on a pan-molecular scale.

Table of contents

List of figures

List of tables

List of abbreviations

(m)CRC	(Metastatic) colorectal cancer
(m)RNA	(Messenger) ribonucleic acid
5-FU	Fluorouracil
APC	Adenomatous Polyposis Coli
AXIN2	Axin 2
ARD	Automatic relevance determination
BRAF	v-Raf Murine Sarcoma Viral Oncogene Homolog B
C-index	Concordance index
CAF	Cancer-associated fibroblast
CD	Cluster of differentiation
CEA	Carcinoembryonic antigen
CIMP(-0/L/H)	CpG island methylator phenotype (-negative/low/high)
CIN	Chromosomal instability
CMS	Consensus molecular subtypes
CNA	Copy number aberration
CT	Computed tomography
DFS	Disease-free survival
DNA	Deoxyribonucleic acid
EGA	European Genome-Phenome Archive
EMT	Epithelial-mesenchymal transition
ER	(O)Estrogen receptor
FACS	Fluorescence-activated cell sorting
FAP	Familial adenomatous polyposis
FDA	(United States) Food and Drug Administration
FDR	False discovery rate
FFPE	Formalin-fixed paraffin-embedded

FF	Fresh frozen
FGA	Fraction of genome altered
FOLFIRI	Folinic acid (leucovorin), fluorouracil (5FU) and irinotecan
FOLFOX	Folinic acid (leucovorin), fluorouracil (5FU) and oxaliplatin
FPKM	Fragments per kilobase of transcript per million mapped reads
GI	Gastrointestinal
H&E	Hematoxylin and eosin
HGNC	HUGO Gene Nomenclature Committee
HNPCC	Hereditary non-polyposis colon cancer
HR	Hazard ratio
IHC	Immunohistochemistry
IGFR2	Insulin-like Growth Factor 2 Receptor
isBFAC	Integrated sparse Bayesian factor analysis with covariates
KEGG	Kyoto Encyclopedia of Genes and Genomes
KRAS	Kirsten Rat Sarcoma Viral Oncogene Homolog
LDA	Linear discriminant analysis
MCMC	Markov chain Monte Carlo
MCR	Misclassification error rate
MLH1	MutL Homolog 1
MMR	Mismatch repair
MRI	Magnetic resonance imaging
MSI(-L/H)	Microsatellite instable (-low/high)
MSS	Microsatellite stable
NGS	Next generation sequencing
NICE	National Institute for Health and Care Excellence
NK	Natural killer
NMF	Non-negative matrix factorisation
OS	Overall survival
PAM	Prediction analysis of microarrays
PCA	Principal components analysis
PCR	Polymerase chain reaction
PD1	Programmed cell death protein 1
PDX	Patient-derived xenograft
PD	Progressive disease
PFS	Progression-free survival

PR	Partial response
RECIST	Response Evaluation Criteria in Solid Tumors
RFS	Regression-free survival
RMA	Robust multiarray averaging
RNA	Ribonucleic acid
RPM	Reads per million
RSEM	RNA-seq by expectation maximization
RPPA	Reverse phase protein array
SCNA	Somatic copy number aberration
SD	Stable disease
SD	Standard deviation
SMAD2,4	Mothers Against Decapentaplegic Homolog 2 and 4
SNV	Single-nucleotide variant
ssGSEA	Single-sample gene set enrichment analysis
SVM	Support vector machine
SVR	Support vector regression
t-SNE	t-distributed stochastic neighbour embedding
TA	Transit-amplifying
TCGA	The Cancer Genome Atlas
TGFβ	Transforming Growth Factor Beta
TCPA	The Cancer Proteome Atlas
TP53	Tumour Protein P53
UCSC	University of California Santa Cruz
WHO	World Health Organization
WT	Wild-type
XELOX	Capecitabine and oxaliplatin

Chapter 1

Introduction

1.1 Colorectal cancer and its heterogeneity

1.1.1 Anatomy, structure and function of the human large intestine and colon

The gastrointestinal (GI) tract is a series of connected organs that runs from the mouth to the anus, at the lower end of which is the large intestine (Betts *et al.*, 2016) (Figure 1.1a). The large intestine's prime function is to absorb nutrients and water not taken up by the small intestine, host a commensal microbiome that synthesises key vitamins (including vitamins K and B12, folic acid, and thiamine) (Gorbach, 1996), and expel faeces from the body (Betts *et al.*, 2016).

The caecum (Figure 1.1b) is the first section of the large intestine to receive food residue from the small intestine, and is also connected to the appendix and colon (Betts *et al.*, 2016). From the caecum, the residue travels up the right side of the abdomen through the ascending colon, which then bends sharply to the left at the hepatic flexure to be passed across the body through the transverse colon (Betts *et al.*, 2016). Another sharp bend at the splenic flexure redirects the colon back down the left side of the abdomen through the descending colon (Betts *et al.*, 2016). A final curve of the sigmoid colon through the pelvis back towards the midline of the body leads to the rectum and anus (Betts *et al.*, 2016) (Figure 1.1b).

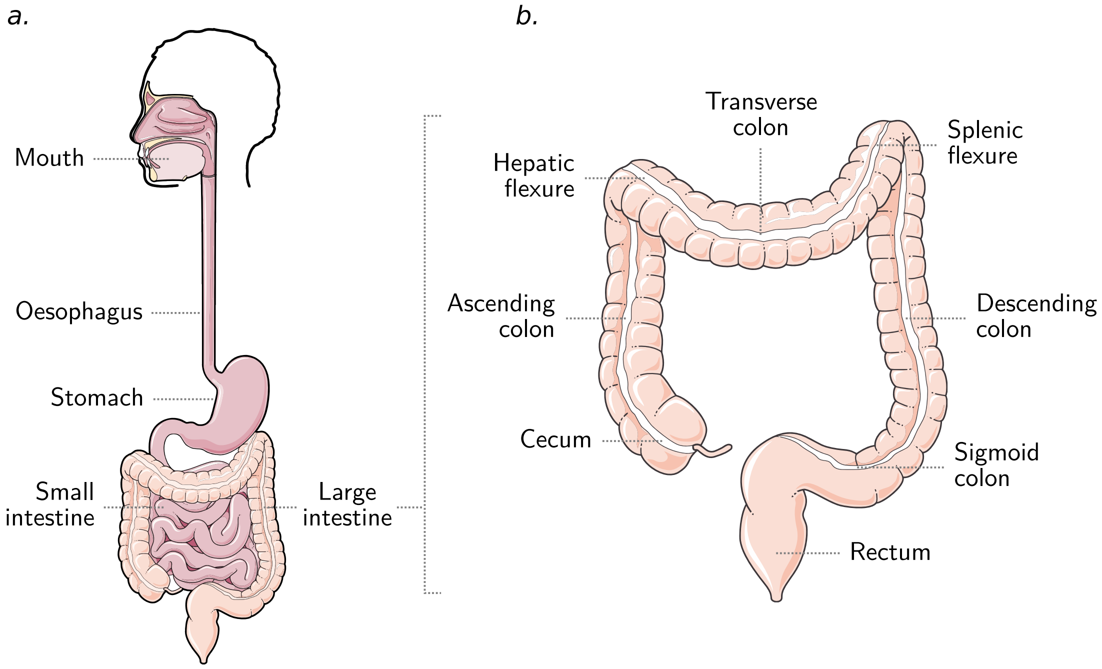


Figure 1.1: Anatomy of the human gastrointestinal system. The anatomy of the human *(a)* gastrointestinal tract and *(b)* large intestine. Adapted from "Complete digestive apparatus" and "Colon" by Servier Medical Art, used under CC BY 3.0. Labels added to original.

The inner mucosal wall of the colon is densely scattered with deep crypts that penetrate perpendicular to the colonic lumen (Betts *et al.*, 2016) (Figure 1.2a-b). At the superficial ends of the crypts lie the differentiated absorptive (enterocyte) and secretive (goblet) cells (Figure 1.2c) that absorb water, salts and microbially-synthesised vitamins, and secrete mucus to lubricate the movement of faeces and protect the intestine (Betts *et al.*, 2016). These cells are maintained by a indefinitely-dividing population of colon stem cells at the deepest ends of the crypts, whose progeny — transit-amplifying cells — will divide a finite number of times before differentiating into enterocyte and goblet cells (Rao & Wang, 2010).

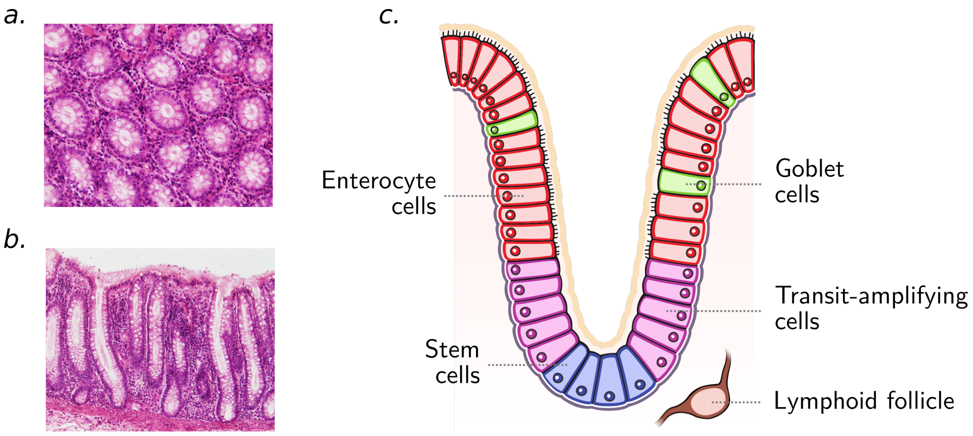


Figure 1.2: The colon crypts contain multiple distinct cell types along their depth. Haemotoxylin and eosin (H&E) light microscopy images of colon crypts in the *(a)* transverse and *(b)* longitudinal planes. "Crypt of Lieberkühn, transverse section" and "Crypt of Lieberkühn, muscularis mucosae" (cropped) by Lutz Slomianka. *(c)* An illustration of the cellular structure of a crypt in the colon. Derived from "Intestinal villi" by Servier Medical Art, used under CC BY 3.0.

The passage of foreign substances, including microbes, in the colonic lumen also warrants the presence of immune cells in the deeper layers of the colon wall to react to pathogens, as well as to the invasion of commensal microorganisms. Underneath the epithelium, the lamina propria layer of connective tissue hosts leukocyte cells in solitary lymphoid follicles (Mowat & Agace, 2014).

1.1.2 Colorectal cancer biology and progression

Colorectal cancer (CRC) can sometimes be caused by mutations in either *APC* (Adenomatous Polyposis Coli) or DNA mismatch repair (MMR) genes, leading to Familial Adenomatous Polyposis (FAP) or Hereditary Non-Polyposis Colon Cancer (HNPCC) syndrome respectively, and an extremely high probability of developing CRC at an early age (Sturrock *et al.*, 2015). However, these familial cases account for a small minority of CRCs (Sturrock *et al.*, 2015).

The majority of CRCs are sporadic, and originate from the mucosal epithelial lining of the large intestine, making them adenocarcinomas (Sturrock *et al.*, 2015). The canonical pathological sequence of CRC follows a progression from adenoma to carcinoma (Christie & Sieber, 2011), illustrated in Figure 1.3.

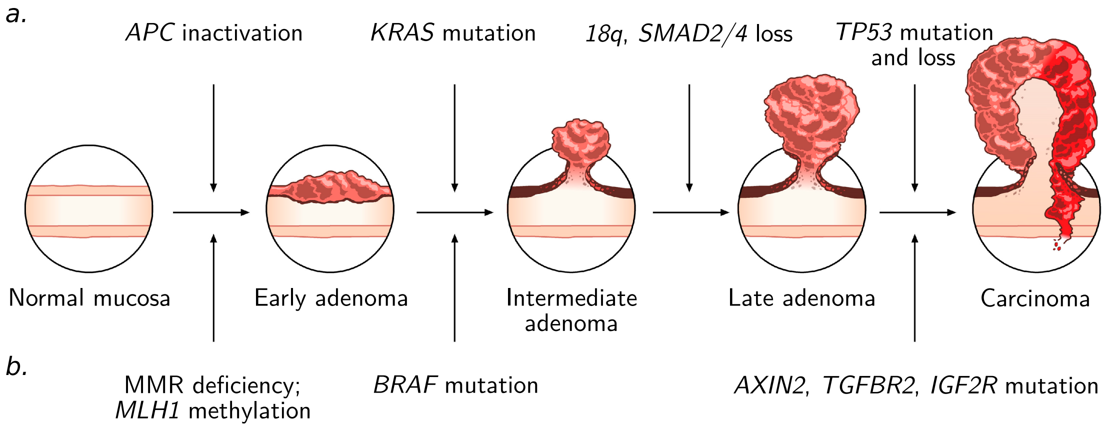


Figure 1.3: Colorectal cancer can develop along a canonical adenoma-carcinoma sequence. *(a)* The chromosome instability (CIN) and *(b)* microsatellite instability (MSI) adenoma-carcinoma pathways. Derived from *Figure 1* of (De Palma *et al.*, 2019), used under CC BY 3.0 license.

This sequence is estimated to take 10-40 years, although not all adenomas will become malignant (Christie & Sieber, 2011).

In early sporadic adenomas, *APC* mutation is present in the majority of samples (161/278 adenomas in (Diergaarde *et al.*, 2005); Figure 1.3a), leading to aberrant Wnt activation which disrupts stem cell maintenance, proliferation, and differentiation of large intestine epithelia (Zhan, Rindtorff & Boutros, 2017). Also common are gains of chromosome 7p, occuring in 36% of cases according to meta-analysis of 430 CRC samples (Baudis, 2007). At the intermediate stages, approximately 12% of adenomas harbour mutant *KRAS* (72/622 in(Juárez *et al.*, 2017)) (Kirsten Rat Sarcoma Viral Oncogene Homolog), abrogating its regulation of cell proliferation (Pylayeva-Gupta, Grabocka & Bar-Sagi, 2011). This is followed in approximately 47%% of adenomas by loss of chromosome 18q (Baudis, 2007) (which carries *SMAD2* and *SMAD4* (Mothers Against Decapentaplegic Homolog 2 and 4) of the TGFβ (Transforming Growth Factor Beta) signalling pathway). Additionally, in approximately 27% of cases, 17p is lost (Baudis, 2007) (which carries *TP53* (Tumor Protein P53), a critical regulator of cellular stress and the DNA damage response).

Another precursor sequence can lead to adenocarcinomas from adenomas (Figure 1.3b). Estimated to be the pathway of ~10-15% of sporadic CRCs, it has been described as constituting early methylation of the promoter region of *MLH1* (MutL Homolog 1), leading to impaired MMR (Sandmeier *et al.*, 2009), and often *BRAF* (v-Raf Murine Sarcoma Viral Oncogene Homolog B) mutation. Defective MMR causes genome-wide hypermutation — particularly at sections of short, repeated sequences of 1-6 basepairs called microsatellites — dubbed microsatellite instability (MSI) (Boland & Goel, 2010). Mutations are common in the repeated sections of *AXIN2* (Axin 2), *TGFβR2* (Transforming Growth Factor Beta Receptor 2), and *IGFR2* (Insulin-like Growth Factor 2 Receptor). Patients with MSI cancers tend to be female, older, have disease in the right-hand side of the colon, and can expect better prognosis (Kawakami, Zaanan & Sinicrope, 2015).

While this level of understanding of molecular and morphological aberrations has certainly shed light on the biology and progression of CRC, only a fraction of this has translated into the refinement of clinical practice, as discussed below.

1.1.3 Clinical management of CRC and clinico-pathological markers

1.1.3.1 Staging

Due to the known progression of carcinomas from adenomas, screening and removal of precancerous polyps by colonoscopy can prevent CRC or facilitate early diagnosis (Sturrock *et al.*, 2015). When adenocarcinoma is suspected, diagnosis is confirmed via a biopsy (National Institute for Health and Care Excellence, 2014). The process of staging the cancer then follows (Table 1.1). The primary concern is identifying if distant metastasis has occurred, as this would reduce the benefit to the patient of invasive surgical resection of the primary tumour (Sturrock *et al.*, 2015). A computed tomography (CT) scan of the chest, abdomen and pelvis to evaluate the common sites of metastases (mesenteric lymph nodes, liver and lungs) is performed, alongside blood work including carcinoembryonic antigen (CEA) levels as a baseline for post-treatment surveillance (Sturrock *et al.*, 2015).

Table 1.1: Overview of the staging criteria for colorectal tumours. (Martins *et al.*, 2018)

Stage	Primary tumour state	Regional lymph node metastases	Distant metastases
Stage I	Invades submucosa or muscularis propria	None	None
Stage II	Penetrates muscularis propria, peritoneum or other organs	None	None
Stage III	Any of the above	At least one metastasis	None
Stage IV	Any of the above	Any of the above	Metastasis in at least one distant organ/site

For rectal cancer, a more thorough assessment of the primary tumour is performed due to concerns over local invasion due to the rectum's extraperitoneal location and associated high risk of local invasion and recurrence; endoscopic ultrasound or magnetic resonance imaging (MRI) can be used to precisely evaluate the tumour's location in relation to these (National Institute for Health and Care Excellence, 2014).

1.1.3.2 Treatment strategies in early and late stage CRC

The main treatment strategy for CRC is complete surgical resection, wherever possible (National Institute for Health and Care Excellence, 2014). When disease is metastatic or the primary tumour is unresectable, treatment refocuses on limiting progression and palliating symptoms (Sturrock *et al.*, 2015).

1.1.3.2.1 Early stage colon and rectal cancer management. Table 1.2 gives an overview of the main treatment pathways for early-stage (i.e. non-metastatic, stages I-III) CRC. For colon tumours, resection aims to remove the tumour, its surrounding margins, and any associated lymph nodes (Sturrock *et al.*, 2015). Adjuvant chemotherapy is then offered in the case of lymph node involvement or deep invasion of the tumour through the colon wall to reduce the odds of subsequent recurrence or metastasis. (National Institute for Health and Care Excellence, 2014).

For locally advanced rectal cancer, a more multimodal treatment approach is employed (National Institute for Health and Care Excellence, 2014), in view of the fact that neoadjuvant radio- and/or chemotherapy reduce the chance of recurrence(Sturrock *et al.*, 2015; National Institute for Health and Care Excellence, 2014).

1.1.3.2.2 Late stage colon and rectal cancer management. At stage IV metastatic disease, surgery with curative intent may still be feasible (resection of both the primary and metastatic tumours) if metastases are isolated to the liver or lungs (Sturrock *et al.*, 2015; National Institute for Health and Care Excellence, 2014). In this case, perioperative chemotherapy can also be offered to reduce the risk of recurrence.

If the tumours are not immediately resectable, local control of the primary tumour with chemotherapy (and/or radiotherapy for rectal cancer, Table 1.3) may help to control symptoms, control disease and prolong life (Sturrock *et al.*, 2015). Targeted biological agents such as anti-EGFR (Epidermal Growth Factor Receptor; e.g. cetuximab, panitumumab) or anti-angiogenic agents (e.g. bevacizumab) can also be applied. If the disease is widely metastatic, palliative chemotherapy (with biological agents, depending on the patient's fitness) is the remaining option for patients (Sturrock *et al.*, 2015).

Underlying patients' responses to all these therapies is the crucial factor of intertumoural heterogeneity: the spatial, morphological and molecular differences between tumours. This variation – particularly molecular variation – will be introduced next.

Table 1.2: Overview of the recommended treatment pathway of patients with early-stage (stage I–III) colorectal tumours in the UK. Adapted from "NICE Pathways: Managing local colorectal tumours"* and BMJ Best Practice (Stein, 2019) .

Treatment stage	Patient group	Treatment
Preoperative management	Moderate/high-risk resectable primary rectal tumours	Preoperative radio/chemoradiotherapy
	High-risk primary rectal tumours that appear unresectable or borderline resectable	Preoperative chemoradiotherapy
Surgery	All patients	Open or laparoscopic surgery

Treatment stage	Patient group	Treatment
Further treatment	High risk stage II and all stage III tumours	Adjuvant chemotherapy (e.g. capecitabine/5FU and oxaliplatin)
Follow-up after apparently curative resection	All patients	CT scans, CEA levels, colonoscopies

Table 1.3: Overview of the recommended treatment pathway of patients with late-stage (stage IV) colorectal tumours in the UK. Adapted from "NICE Pathways: Managing advanced and metastatic colorectal cancer"* and BMJ Best Practice (Stein, 2019).

Treatment stage	Patient group	Treatment type
Imaging	All patients	CT scans to determine metastases' extent/location
Perioperative management	Resectable rectal tumours	Fluoropyrimidine-based chemotherapy ± oxaliplatin and/or irinotecan ± radiotherapy
	Resectable colon tumours	Fluoropyrimidine-based chemotherapy ± oxaliplatin and/or irinotecan
Surgery	Resectable tumours	
First-line therapy	All patients	Fluoropyrimidine-based chemotherapy ± oxaliplatin and/or irinotecan ± anti-angiogenic or anti-EGFR (*RAS* WT only)
Second-line therapy	Patients who progress after first line	Alternative chemotherapy to first line

1.1.3.3 Biomarkers in clinical use in CRC

Molecular biomarkers are still limited in their clinical application to CRC. RAS family mutations predict primary resistance to anti-EGFR monoclonal antibodies, e.g. cetuximab (Sorich *et al.*, 2015) in metastatic CRC. However, while approximately ~47% of metastatic CRCs harbour wild-type *RAS*, only half of these will respond to these drugs (Sorich *et al.*, 2015), with varying unclear mechanisms of primary and acquired resistance (Martins *et al.*, 2018). In early stage cancers, MSI cases may have better prognosis, are less likely to gain a survival benefit from fluorouracil (5FU; a chemotherapy agent that inhibits DNA replication by blocking the synthesis of thymidine), but may benefit from longer survival under irinotecan treatment (which inhibits topoisomerase I, damaging DNA) (Christie & Sieber, 2011). Overall, only around half of CRC patients will survive 10 years or more after diagnosis (Cancer Research UK, 2016; Yu *et al.*, 2019), although 10-year disease-specific survival is higher at 68%, partially due to the high incidence of mortality from other causes in CRC patients owing to their age at diagnosis (median approximately 62 years(Yu *et al.*, 2019)).

1.1.4 Characterisations of intertumoural molecular CRC heterogeneity

Despite the appealing simplicity of a step-wise accumulation of driver mutations as a model for CRC development, as described above, the vast majority of CRCs do not exhibit mutations in all three key genes in this model (*APC*, *KRAS* and *TP53*) (Joung *et al.*, 2017), and a minority show no detectable mutations in any of these genes. Hence, additional molecular features have gained traction as important discriminators in CRC.

Alongside MSI, described earlier in this chapter, CpG island methylator phenotype (CIMP) and chromosomal instability (CIN) are markers which have been widely investigated in CRC research, although not included in routine clinical practice. CIMP is distinguished by increased methylation of CpG island-enriched promoter regions, silencing tumour suppressor genes (Toyota *et al.*, 1999). CIMP tumours are highly mutated, poorly differentiated, but CIMP is not a reliable independent predictor of prognosis (Toyota *et al.*, 1999). The CIN phenotype (Fearon & Vogelstein, 1990), where aneuploidy causes the loss of e.g. APC, KRAS or TP53 function, is mostly exclusive of MSI (Simons *et al.*, 2013) and has worse prognosis than MSI tumours (Watanabe *et al.*, 2012) in early-stage cancers.

MSI CRCs can be further subdivied into MSI-high (MSI-H) and MSI-low (MSI-L) subgroups, the latter representing tumours where a lower number of repeat markers show evidence of MSI — however, there is not a consensus over whether MSI-L tumours are functionally different enough from MSS tumours to be treated as a distinct group

(Pawlik, Raut & Rodriguez-Bigas, 2004). MSI and CIMP status can be combined into 4 major subtypes, in order of decreasing incidence (Ogino & Goel, 2008):

i. MSI-L/Microsatellite stable (MSS), CIMP-low (L) or negative (0): (~75-80%)
ii. MSI-H, CIMP-high (CIMP-H) (~10%)
iii. MSI-L/MSS, CIMP-H (~5-10%)
iv. MSI-H, CIMP-L/0 (~5%)

These methylation-, genomic- and chromosome-level classifications can be augmented by transcriptomic subgroups, as pioneered in breast cancer (Perou *et al.*, 2000; Dowsett *et al.*, 2013). Previously, our lab published five gene expression subtypes of CRC (Sadanandam *et al.*, 2013), the CRCAssigner subtypes, named for their enrichment of genes associated with normal colon cell types described in Chapter 1.1.1 (Figure 1.4): enterocyte, goblet-like, TA, stem-like and inflammatory. The CRCAssigner subtypes are appealing as an adjunct to MSI, CIMP and CIN as they are conceptually equivalent to these other classifications: their designation is based on transcriptomic cell phenotypes alone, similarly to MSI being based on genomic, CIMP on methylomic, and CIN on chromosomal features.

1.1.4.1 The CRCAssigner subtypes

The five CRCAssigner subtypes can be distinguished by their expression of genes characteristic of normal cell types found in the colon (Sadanandam *et al.*, 2013) (Chapter 1.1.1, Figure 1.4). The enterocyte subtype expresses enterocyte marker genes (*CA1-2*, *KRT20*, *SLC26A3*, *AQP8* and *MS4A12*); the goblet-like subtype is characterised by goblet cell markers (*MUC2* and *TFF3*); the inflammatory subtype has high expression of chemokines and interferon-related genes (*CXCL9-13* and *IFIT3*); the stem-like subtype disproportionately expresses Wnt signalling targets and stem cell, myoepithelial and mesenchymal markers (*SFRP2*, *SFRP4*, *FN1*, *TAGLN*, *ZEB1-2*, *TWIST1*, *SNAI2*); while the TA subtype has more heterogeneous expression due to TA cells' spread along a differentiation gradient from intestinal stem cells to differentiated enterocyte and goblet-like cells.

While the CRCAssigner subtypes were named for their transcriptomic similarities to normal crypt cells, they do show enrichment for non-transcriptomic biomarkers (those not primarily defined by gene expression characteristics). For example, the inflammatory, stem-like and TA subtypes are enriched for the known categories of MSI and MSS, with MSI tumours falling into the inflammatory group, and MSS tumours into the stem-like and TA subtypes (Sadanandam *et al.*, 2013; Guinney *et al.*, 2015). Correspondingly, the inflammatory subtype has favourable disease-free survival (DFS), while stem-like has poor DFS (Sadanandam *et al.*, 2013). However, the TA subtype has good DFS despite being MSS (Sadanandam *et al.*, 2013).

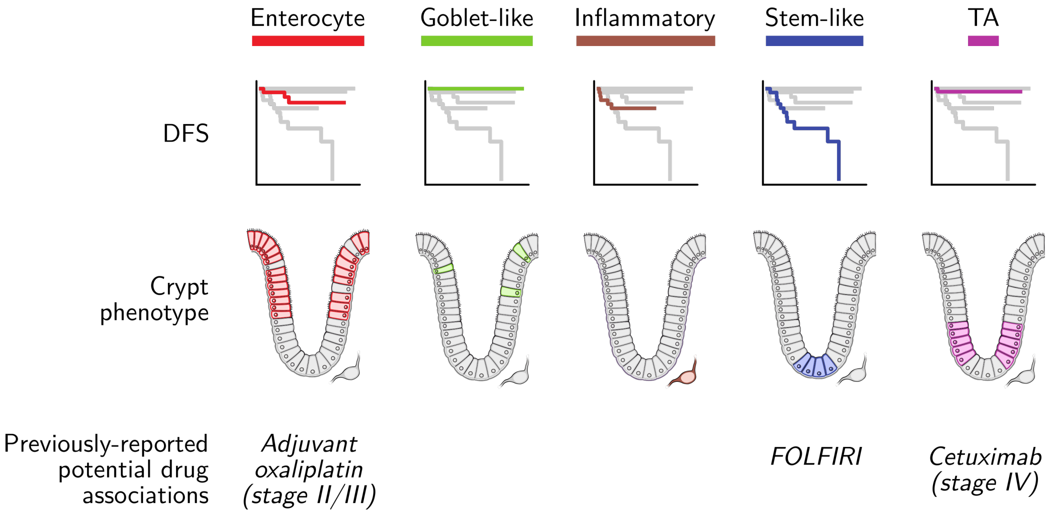


Figure 1.4: The CRCAssigner subtypes have differential prognostic power and potential drug associations, and similarities to normal colon crypt cell types. Overview of the CRCAssigner subtypes of CRC, including disease-free survival (DFS) and crypt phenotypes from the original publication (Sadanandam *et al.*, 2013) and drug sensitivities from the original (FOLFIRI and cetuximab) and follow-up (oxaliplatin (Song *et al.*, 2016)) publications. Coloured cells indicate the location of the equivalent normal cell types in the colon crypt.

CRCAssigner subtypes were shown to exhibit subtype-specific associations with cetuximab and FOLFIRI (folinic acid, 5FU and irinotecan) treatments in patients (Sadanandam *et al.*, 2013). More recently, their possible association with oxaliplatin treatment was demonstrated, wherein CRCAssigner was retrospectively identified as more able to stratify patients into those who do and do not experience increased recurrence-free survival from oxaliplatin than the CMS classifier (described in the next section), with enterocyte subtype patients representing those who have longer survival (Song *et al.*, 2016). This difference was significant in the discovery cohort, but did not reach significance in the validation cohort, although the trend remained the same.

1.1.4.2 The CMS subtypes

Five other groups have released major publications on gene expression subtypes in CRC near-concomitantly, and defined the following subtypes: (Budinska *et al.*, 2013; Marisa *et al.*, 2013; Roepman *et al.*, 2013; De Sousa E Melo *et al.*, 2013; Schlicker *et al.*, 2012)

1. Budinska *et al.*
   - i. Surface crypt-like
   - ii. Lower crypt-like
   - iii. CIMP-H-like
   - iv. Mesenchymal
   - v. Mixed
2. Roepman *et al.*
   - i. Deficient MMR epithelial
   - ii. Proliferative epithelial
   - iii. Mesenchymal
3. De Sousa E Melo *et al.*
   - i. CIN
   - ii. MSI
   - iii. Serrated
4. Marisa *et al.*
   - i. CIN immune down
   - ii. Deficient MMR
   - iii. *KRAS*-mutant
   - iv. Cancer stem cell
   - v. CIN Wnt up
   - vi. CIN normal-like
5. Schlicker *et al.*
   - i. 1.1 (Strongly mesenchymal, late stage)
   - ii. 1.2 (Mesenchymal, MSI)
   - iii. 1.3 (Mesenchymal, MSS)
   - iv. 1.4 (Epithelial)
   - v. 1.5 (Epithelial, MSS)

A consortium effort consolidated these subtypes into a consensus solution (consensus molecular subtypes, CMS) of four subtypes (Guinney *et al.*, 2015):

- i. CMS1 - MSI immune (enriched for: MSI, CIMP, hypermutation, *BRAF* mutation, immune infiltration)
- ii. CMS2 - Canonical (enriched for: somatic copy number aberrations (SCNAs), Wnt and MYC activation)
- iii. CMS3 - Metabolic (enriched for: *KRAS* mutation, metabolic disregulation)
- iv. CMS4 - Mesenchymal (enriched for: SCNAs, stromal infiltration, TGFβ activation, angiogenesis)

These subtypes were defined using a network approach, whereby the overlap of samples between subtypes from different studies (measured using Jaccard index[†]) were interpreted as edges in a weighted network (Guinney *et al.*, 2015). The nodes of this network were then clustered using the Markov Cluster Algorithm (Van Dongen, 1998). This graph-based clustering algorithm relies on the idea that when some nodes are more densely clustered than others, a random walk from one node to another is more likely to stay within a cluster than travel between clusters. Hence, by performing many random walks through the graph, dense regions of nodes can be grouped into clusters.

These consensus subtypes have been adopted for many subsequent studies, with clinically-relevant but sometimes contradictory findings (Lenz *et al.*, 2018; Mooi *et al.*, 2018). However, efforts continue on developing the six original publications' subtypes due to their context-specific use as markers where the CMS subtypes may be too broad, for example in the context of personalising oxaliplatin treatment for stage II/III patients (Song *et al.*, 2016). In addition, new subtyping schemes have been published (Isella *et al.*, 2017) in reaction to the suggestion that mesenchymal/stem-like subtypes may be reflecting stromal gene expression, rather than tumour gene expression (Isella *et al.*, 2015), discussed in the section below.

1.1.4.3 Contamination of gene expression profiles by stromal cells and the CRIS subtypes

Human tumour samples contain a mixture of cancerous cells and non-cancerous stromal cells. Whilst impossible to determine using microarrays on human tumour samples, the recent use of RNA-seq to profile patient-derived mouse xenograft (PDX) samples has allowed for the delineation of which genes are expressed mostly in cancer or stromal cells. This can be achieved because PDX samples will contain human cancer cells, but mouse stroma; hence, by aligning sequencing reads to the human and the mouse genomes, the relative expression of each gene in cancer and stromal cells can be calculated.

[†] The Jaccard index of two sets of labels A and B is defined at the intersection of the label sets divided by the union of the label sets, i.e.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

In the context of CRC gene expression signatures, recent work (Isella *et al.*, 2015) has shown that the expression of $\sim 85\%$ of stem-like subtype (and $\sim 50\%$ of inflammatory subtype) genes is higher in the PDX's stroma than in its human cancer cells. In reaction to this result, new subtypes (CRC intrinsic subtypes, or CRIS) were developed from PDX data (Isella *et al.*, 2017) , using an otherwise similar methodology to the derivation of the CRCAssigner subtypes using patient samples (Sadanandam *et al.*, 2013), the major change being the removal of genes for which more than 50% of expression could be attributed to stromal cells.

The differences between these two classifiers is to be expected, as the CRCAssigner subtypes were discovered using patient samples having tumour content $> 65\%$, meaning non-cancer cells were included in the expression profiling. However, whether it is preferable to include or exclude stromal expression when classifying tumours is not settled. While it is certainly important to understand which signals are coming from the cancer and non-cancerous cells for furthering biological understanding, stromal signalling can heavily influence tumour progression (Mueller & Fusenig, 2004), and the inclusion of stromal expression in a subtyping scheme utilised in a clinical setting could be informative.

1.1.5 Summary

Given these reported differences in subtypes' prognosis and drug-specific associations from cell lines and retrospective analyses of patient data, the question arises of how to make prospective decisions about patients' care that are informed by these insights. This question will be the subject of the next section.

1.2 Clinical translation of molecular subtypes

In CRC, the clinical evaluation of *RAS* and *BRAF* mutations is performed with polymerase chain reaction (PCR)-based methods or newer pyrosequencing, next-generation sequencing (NGS) or Sanger methods (Westwood *et al.*, 2014). MSI can be assessed by determining the loss of MMR genes by immunohistochemistry (IHC) or PCR-based testing to compare the counts of nucleotide repeats in established microsatellite markers between tumour and normal samples (Ryan *et al.*, 2017) — but there is no evaluation of transcriptomic subtypes in routine clinical management of CRC.

In breast cancer, however, the five breast cancer subtypes first defined in the year 2000 (Perou *et al.*, 2000) are arguably the most-studied cancer gene expression subtypes, and there have been efforts to translate them into routine clinical use (the Prosigna PAM50 assay (Dowsett *et al.*, 2013) below). This assay, and others not based on these subtypes, have been successful in predicting prognosis and guiding chemotherapy decision-making in patients with ER+/HER2- early breast cancer, and have been FDA-approved (Pond, Piccart-gebhart & Brand, 2019). As such, they could provide a model for developing a similar tool in CRC. These assays include:

i. Oncotype DX (Cronin *et al.*, 2007)
   21 genes measured using reverse transcription (RT)-PCR on FFPE samples giving a high/intermediate/low risk call
ii. MammaPrint (Mook *et al.*, 2007)
   70 genes measured using microarrays on fresh frozen (FF) or FFPE samples, giving a high/low risk call and subtype information
iii. Prosigna PAM50 (Dowsett *et al.*, 2013)
   50 genes measured using nCounter (NanoString Technologies) on FFPE samples giving high/intermediate/low risk and subtype information

The results of these tests are often discordant with each other (Vieira & Schmitt, 2018). A prospective clinical trial is ongoing which will compare Oncotype DX, MammaPrint and Prosigna PAM50 and other breast gene signatures in the same samples, with preliminary results showing agreement as low as 39.4% (Bartlett *et al.*, 2016). Some previous studies have shown greater prognostic power using the Prosigna PAM50 test

(Dowsett *et al.*, 2013), but which assay is the most effective at stratifying patients so that some can be safely spared chemotherapy is yet to be settled.

Two gene expression tests for colon cancer, designed by the developers of the MammaPrint and Oncotype DX assays, were made available in the early 2010s:

i. Oncotype DX Colon Cancer (Kerr *et al.*, 2009)
   12 genes meaured using RT-PCR on FFPE samples giving a high/intermediate/low risk call
ii. ColoPrint (Salazar *et al.*, 2011)
   18 genes measured using microarrays on FF samples, giving a high/low risk call

These assays aim to identify patients which with stage II colon cancer should receive adjuvant chemotherapy in addition to surgery. However, they have not been approved by the FDA (U.S. Food & Drug Administration, 2019). This is because the difference in risk between groups in colon cancer was lower using these tests than between the equivalent risk groups in breast cancer (Kelley & Venook, 2011), and while they provide information on risk of recurrence, evidence is lacking of their ability to predict responses to chemotherapy (Kelley & Venook, 2011; Sharif & O'Connell, 2012).

The breast cancer tests listed above are routinely performed in appropriate cases of that disease, as are the above-listed mutation and MSI assessments in CRC — but no equivalent assays for the determination of subtype/risk have previously been successfully developed for use beyond research in CRC. One aspect of implementation of such a gene expression assay for clinical use that must be assessed is its ability to handle tissue preserved via different methods, discussed next.

1.2.1 Tissue preservation and platform considerations

In research, fresh frozen tissue (that which has been snap frozen in liquid nitrogen and stored at ultra-low temperatures) provides the highest quality DNA, RNA and proteins for subsequent profiling (Klopfleisch, Weiss & Gruber, 2011). In clinical practice, frozen tissue is prohibitively expensive to store long-term and cannot be used for routine pathology, e.g. haemotoxylin and eosin (H&E) staining and IHC.

Instead, tissue is fixed by immersing it in formalin, passing ethanol and xylene through it to dehydrate it and make it permeable to paraffin, followed by infiltrating it with liquid paraffin which is then allowed to cool and harden (formalin-fixed, paraffin-embedded; FFPE) (Iles & Butler, 2012). However, this has severe molecular implications. DNA, RNA and protein are cross-linked to each other by addition of methylol groups, which can progress into methylene bridge formation (Masuda *et al.*, 1999). These chemical alterations cause damage that can severely impede and distort DNA, RNA and protein profiling of FFPE samples (Vermeulen *et al.*, 2011; Kong *et al.*, 2014).

A number of publications have evaluated the applicability of RNA profiling techniques such as qRT-PCR, microarrays and nCounter to FFPE tissue. In general, these techniques can be applied to FFPE tissue with moderately high (but variable) concordance with FF tissue results (see e.g. qRT-PCR (Sánchez-Navarro *et al.*, 2010; Mullins *et al.*, 2007); microarrays (Fedorowicz *et al.*, 2009; Duenwald *et al.*, 2009; Frank *et al.*, 2007)). Differences between these platforms' concordance between FF and FFPE can be attributed to several factors. For example, differing probe/primer designs for the same transcript between technologies can target different locations on the mRNA, leading the degradation associated with FFPE tissue to have different effects on expression quantification (Etienne *et al.*, 2004). Normalisation of qRT-PCR data is also done in relation to a handful of reference or housekeeping genes, presumed to be uniformly expressed across samples, whereas microarray data normalisation relies on the expression of all transcripts measured (Guénin *et al.*, 2009). nCounter — while a newer technology and as such not so extensively studied — exhibits particularly high correlation between FFPE and FF tissue gene expression profiles, in the range of 0.87–0.9 (Kolbert *et al.*, 2013; Norton *et al.*, 2013; Reis *et al.*, 2011). While these studies may have had only modest sample sizes, the requirement of the nCounter probes for only a relatively short section of RNA to be intact provides a theoretical basis for this technology being particularly suited to FFPE tissues (see Chapter 2.3.1 for more details of the platform).

The nCounter platform (Geiss *et al.*, 2008) has previously been exploited to develop the FDA-approved breast cancer PAM50 assay described above, as well as assays to predict medulloblastoma (Northcott *et al.*, 2012) and lymphoma (Scott *et al.*, 2014) subtypes. This platform measures gene expression in the form of discrete counts of a set of pre-defined, barcoded mRNAs, and requires no amplification step, eliminating a potential source of bias. While nCounter can profile fewer genes simultaneously than microarrays (which measure the relative expressions of a large panel of pre-defined transcripts) or RNA-seq (which directly reads the base sequences of all transcripts in the sample), it has less hands-on time for the user, and faster turn-around time (Ragulan *et al.*, 2019).

Another feature of subtyping assays that must be considered is the process by which the data is normalised, and how each sample is then assigned to a subtype, as examined in the following section.

1.2.2 Data normalisation and subtype calling

Gene expression subtype classification methods vary greatly in their procedures for both normalising the data and assigning a subtype to an expression profile. These depend on both the gene expression profiling protocol (e.g. platform and tissue type), and how the subtypes were defined (e.g. a characteristic "average" gene expression profile, relative expression of genes) and discovered (e.g. clustering algorithm). Consideration also needs to be given to whether the classifier should be "single sample", that is, whether the classification of a particular sample stay constant regardless of the other samples processed alongside it.

The different studies' CRC subtypes described above in Chapter 1.1.4 all derived from clustering (usually hierarchical) of microarray gene expression profiles from fresh frozen tissues. The classifiers were more varied, including linear discriminant analysis (LDA) (Budinska *et al.*, 2013), a shrunken-centroid variant of LDA (De Sousa E Melo *et al.*, 2013), and clustering of new samples based on differentially expressed genes (Schlicker *et al.*, 2012).

The CRCAssigner classifier was built on subtype centroids that were derived in the original publication (Sadanandam *et al.*, 2013), from clusters of samples that were discovered using non-negative matrix factorisation (NMF) clustering (Brunet *et al.*, 2004). These centroids are characteristic profiles of each subtype, consisting of genes selected using significance analysis of microarrays (SAM) (Tusher, Tibshirani & Chu, 2001) and prediction analysis of microarrays (PAM) methodologies (Tibshirani *et al.*, 2002).

SAM is a permutation-based method to identify differentially expressed genes between subtypes (Tusher, Tibshirani & Chu, 2001). PAM can then be used to derive centroids for the subtypes by calculating the mean and standard deviation of the expression of each gene within each class, and dividing the two. These centroids are then thresholded by a given value to set non-informative genes to zero [‡] (Tibshirani *et al.*, 2002). The subtype of a sample can then be predicted by finding the nearest centroid to that sample, in squared Euclidean distance. Different thresholds can be tested by measuring the misclassification error rate for each threshold during K-fold cross-validation.

1.2.3 Summary

In Chapter 2, I will show how classification of patient samples into the CRCAssigner subtypes can be achieved using both fresh-frozen and FFPE tissue using the clinically-approved nCounter platform. This new assay could facilitate subtype-driven prospective clinical trials to personalise CRC therapy based on patient transcriptomes.

While the intertumoural heterogeneity described in this section has clear implications for CRC biology and therapy, it is increasingly clear that the variability *within* individual tumours — intratumoural heterogeneity — has an additional confounding effect on patient outcomes (Dunne *et al.*, 2016), as discussed in the next section.

[‡] This can be represented as

$$C_{ij} = \begin{cases} \frac{\mu_{ij}}{\sigma_{ij}} - \theta & \text{if } \frac{\mu_{ij}}{\sigma_{ij}} > \theta, \\ 0 & \text{otherwise,} \end{cases}$$

where C_{ij} is the centroid value, μ_{ij} is the mean, and σ_{ij} is the standard deviation for gene i in subtype j, and θ is the threshold value.

1.3 Characterisations, and *in silico* deconvolution, of CRC intratumoural heterogeneity

Intratumoural heterogeneity is the molecular and morphological variability that can be observed within individual tumours. Intratumoural heterogeneity has an impact on patient outcomes that has only relatively recently been subject to investigation on a pan-cancer scale (Morris *et al.*, 2016; Raynaud *et al.*, 2018; Andor *et al.*, 2015). Intratumoural heterogeneity is influenced by factors such as the profusion of metabolites and nutrients, the abundance of stroma, and immune infiltration in different regions of the tumour (Yuan, 2017).

CRC intratumoural heterogeneity has, in the past two decades, been investigated at the level of *KRAS* and *TP53* mutations, loss of heterozygosity at chromosomes 5q and 18q (Losi *et al.*, 2005), and more recently single-gland mutations(Sottoriva *et al.*, 2015), copy number aberrations (Sottoriva *et al.*, 2015) and methylation (Siegmund *et al.*, 2009), and single-cell PCR (Dalerba *et al.*, 2011) and RNA-seq (scRNA-seq) (Li *et al.*, 2017a). These works have conflicting conclusions over the extent of selection exerted on the tumour during its development, showing either a drop-off in heterogeneity in more advanced cases (Losi *et al.*, 2005), or a lack of selection that means the prevalence of a clone only depends on the time it has had to expand (Sottoriva *et al.*, 2015). In epigenetic terms, it was found that while there is diversity in methylation patterns across malignant glands, this diversity was not spatially dependent, indicating a large early expansion untouched by later selection (Siegmund *et al.*, 2009). An early single-cell study that utilised PCR found that phenotypic differentiation, independent of any clonal or genetic diversification, was an important factor in the intratumoural heterogeneity of CRC (Dalerba *et al.*, 2011). The most recent work, applying single-cell RNA-seq technology, found diverse subgroups of tumour epithelial cells that were associated with the CRCAssigner subtypes (Li *et al.*, 2017a).

The current gold-standard for quantifying intratumoural heterogeneity lies in single-cell techniques such as scRNA-seq and fluorescence-activated cell sorting (FACS). These techniques allow for the direct quantification of cells having particular mutations or expressing pre-selected cell surface markers, and as such give the most high-resolution data regarding the cellular composition of a tumour. However, limitations in sample throughput (due to both technical factors and cost) mean these approaches are not yet suited to profiling the kind of large, well-annotated datasets needed to get a population-level understanding of CRC intratumoural heterogeneity. Hence, computational or *in silico* methods can be turned to, to predict various kinds of intratumoural heterogeneity from data that is not at single-cell resolution. These methods can utilise both genomic and transcriptomic data, as discussed below.

1.3.1 Mutation- and copy number-based deconvolution in CRC

In cancer, the majority of efforts in computationally quantifying intratumoural heterogeneity have focussed on genomic strategies utilising single-nucleotide variation (SNV) and copy number aberration (CNA) data. In CRC in particular, multi-region whole-exome sequencing and copy number profiling allowed for the reconstruction of phylogenetic trees showing the clonal history of primary tumours and their metastases (Kim *et al.*, 2015). This analysis revealed that mutations in *APC* are both highly common and clonally/regionally universal, while mutant *KRAS* was also universal but less common. A subsequent independent study mirrored these results in another set of tumours, and also identified arm-level amplifications of 7p, 7q, 13q, 20p and 20q as common founder events (Uchi *et al.*, 2016).

These multi-region analyses give valuable insights into genomic heterogeneity in CRC. However, they usually only consider a limited sample set due to the complexities of obtaining and profiling multiple sections of tissue from one tumour. In order to draw conclusions about outcomes such as survival in the wider CRC patient population, techniques can be used that estimate genomic intratumoural heterogeneity from bulk data.

One study that took such an approach (Joung *et al.*, 2017) utilised PyClone (Roth *et al.*, 2014), a model that uses Bayesian hierarchical clustering of somatic mutations to identify tumour clones while accounting for possible multiplication of the mutation through copy number aberrations. It found that the presence of more than two clones was significantly associated with a detrimental effect on patient survival (Joung *et al.*, 2017).

While these findings of intratumoural genomic heterogeneity in CRC have undoubtedly increased understanding of the disease and its mechanisms, the question remains as to how this understanding can be translated into patient benefit in the clinic.

Meanwhile, intertumoural transcriptomic heterogeneity in CRC has been shown to have multiple therapeutic implications, both in the basic research setting and in retrospective analysis of clinical trials (see Chapter 1.1.4). Hence, it is worth exploring whether intratumoural transcriptomic heterogeneity could be an alternative route to bring knowledge of the implications of intertumoural heterogeneity into the clinic.

1.3.2 Transcriptomic deconvolution strategies

While it has been known for a short time that some tumour samples can have gene expression characteristic of multiple different transcriptomic subtypes simultaneously (Guinney *et al.*, 2015) — hypothesised to be due to the concomitance of cells of different subtypes within the same tumour — the extent and implications of this phenomenon have not yet been widely investigated in CRC. Due to the present scarcity of single-cell data in CRC, it is also not yet possible to directly answer this question by subtyping individual cells. Hence, computational methods can be turned to in order to explore this observation.

Several pioneering tools in the field of transcriptome-based cell type deconvolution have come from the field of immunology. Normal immune cell types do not exhibit genomic variability within the same human, i.e. somatic mutations and copy number aberrations that would be needed to utilise the tools described in the previous section (Chapter 1.3.1). Instead, several methods for the deconvolution of immune cell types from transcriptomic data have been adopted.

Methods of transcriptomically quantifying cell types can be broadly categorised by their input and output (Finotello & Trajanoski, 2018). Firstly, lists of marker genes for cell types can be input for the calculation of enrichment of each cell type within each sample, although this is not usually interpretable as cell type subpopulations. Secondly, characteristic gene expression profiles of cell types can be used in a supervised fashion to evaluate the proportion of gene expression in a sample attributable to each cell type — this can be taken as an estimate of cell type subpopulations. Finally, totally unsupervised deconvolution of gene expression into characteristic cell type transcriptomic profiles *and* proportions of these cell types is possible. However, this requires further downstream analysis to identify and understand these cell types. Additionally, none of these methods can be used to directly estimate the number of cells belonging to each subpopulation, only relative proportions, due to the fact that the RNA from all the cells in the sample has been pooled and can no longer be attributed back to individual cells.

The second approach listed above is utilised in this thesis, due to the cell types of interest (CRCAssigner subtypes) having well-defined characteristic gene expression. While several methods exist based on linear (Abbas *et al.*, 2009; Li *et al.*, 2016) and constrained least squared regression (Gong *et al.*, 2011; Gong & Szustakowski, 2013; Racle *et al.*, 2017), it is likely that the most popular recent method of this category is CIBERSORT (Newman *et al.*, 2015). CIBERSORT utilises a variant of support vector regression (SVR), which is similar to simple linear regression in that it treats the gene expression profile of interest as a linear combination of the cell types' gene expression profiles. SVR estimates coefficients [§] that can then be normalised and interpreted as cell type proportions within the sample (Figure 1.5). For more details on SVR and its implementation in CIBERSORT, see Chapter 3.2.3.

[§] Linear deconvolution of cell subpopulations can be represented by

$$\underline{g}_i = \left(\sum_{j=1}^{J} \beta_{ij} \underline{c}_j \right) + \underline{\epsilon}_i$$

where $\underline{g}_i$ is the gene expression profile vector of sample i, $\underline{c}_j$ is the gene expression profile vector of cell type j, β_{ij} is the coefficient representing the subpopulation of each cell type in each sample, and $\underline{\epsilon}_i$ is the residual error.

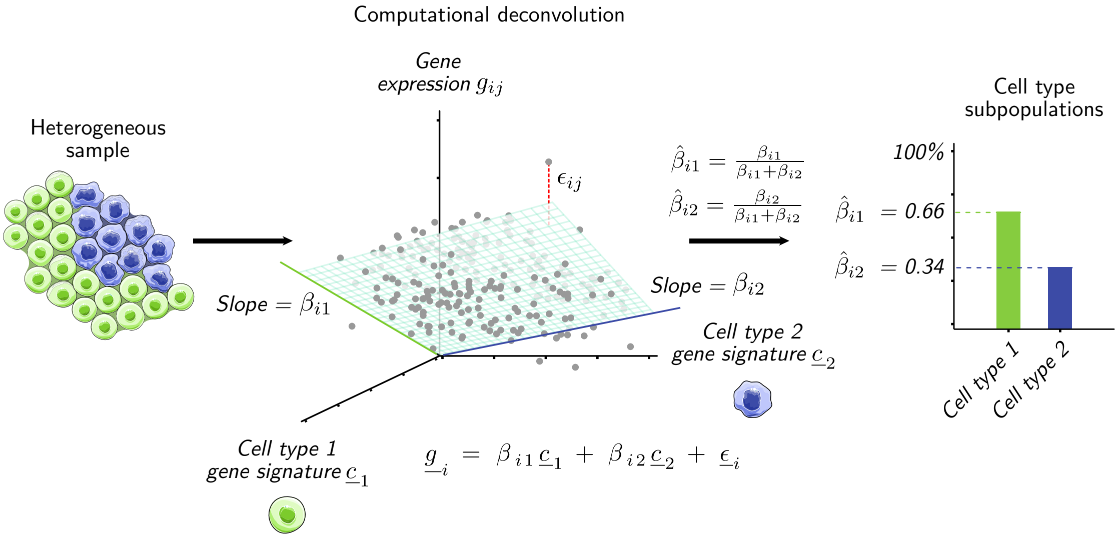

Computational deconvolution
Heterogeneous sample
Gene expression g_{ij}
ϵ_{ij}
Slope $= \beta_{i1}$
Slope $= \beta_{i2}$
$\hat{\beta}_{i1} = \frac{\beta_{i1}}{\beta_{i1}+\beta_{i2}}$
$\hat{\beta}_{i2} = \frac{\beta_{i2}}{\beta_{i1}+\beta_{i2}}$
Cell type subpopulations
100%
$\hat{\beta}_{i1}$ = 0.66
$\hat{\beta}_{i2}$ = 0.34
Cell type 1
Cell type 2
Cell type 2 gene signature $\underline{c}_2$
Cell type 1 gene signature $\underline{c}_1$
$\underline{g}_i = \beta_{i1}\underline{c}_1 + \beta_{i2}\underline{c}_2 + \underline{\epsilon}_i$

Figure 1.5: Computational deconvolution can be used to predict cell type subpopulations in a heterogeneous sample. Summary of the transcriptomic deconvolution methodology employed by CIBERSORT for the estimation of cell type subpopulations. $\underline{g}_i$ is the gene expression profile vector of sample i, $\underline{c}_k$ is the gene expression profile vector of cell type k, β_{ik} is the coefficient representing the subpopulation of each cell type in each sample, and $\underline{\epsilon}_i$ is the residual error. g_{ij} and ϵ_{ij} are the expression and error of the fit of gene j in sample i. $\hat{\beta}_{ik}$ is the normalised coefficient so that $\sum_{k=1}^{K} \hat{\beta}_{ik} = 1$.

CIBERSORT's algorithm was validated against automated cell counting and flow cytometry, and benchmarked against other deconvolution tools such as least squares regression. It was found to predict immune cell subpopulations with high accuracy, and produced lower errors than competing methods (Newman *et al.*, 2015). These exercised included testing CIBERSORT on simulated tumour/leukocyte mixtures, and on 14 patient follicular lymphoma tumours with known immune cell subpopulations measured using flow cytometry. In Chapter 3, I will adopt CIBERSORT's approach to quantify the subpopulations of cells belonging to different transcriptomic cancer subtypes within individual tumours. While this application is different to the original intent of CIBERSORT in that I aim to enumerate different tumour subtype subpopulations, as opposed to immune cell subpopulations, the core problem of quantifying the heterogeneity of cell types present using bulk gene expression profiles and cell type signatures remains identical.

1.3.3 Summary

Elucidation of transcriptomic inter- and intratumoural heterogeneity will likely garner new insights that can bring patient benefit in the relatively near future. Beyond this, the next generation of personalised treatment will require the understanding of each patient's cancer as a complex biological system, consisting of multiple levels of molecular signalling and feedback. How the integration of molecular data of different "omics" types can be achieved for the realisation of this level of comprehension is the subject of the next section.

1.4 Multiomics data integration and analysis

1.4.1 Strategies for step-wise integration of omics data

“Omics” data refers to information — usually high-dimensional — on the molecular landscape of a sample, e.g. mutations and copy number alterations (genomics), DNA methylation (epigenomics), gene/mRNA and microRNA expression (transcriptomics), and protein expression (proteomics). The most common strategy for the integration of data of one omics type to another is through simple step-wise evaluation of the association of postulated features/groups of interest with known features/groups of interest. An quantitative example of this would be the use of Fisher’s exact test to evaluate the enrichment of various mutations in the known transcriptomic CMS subtypes (Guinney *et al.*, 2015), but qualitative analysis is also common.

In the Sadanandam Lab, previous work on pancreatic neuroendocrine tumours has integrated mRNA and microRNA subtypes using a hypergeometric test, revealing that the samples in subtypes had a high overlap (Sadanandam *et al.*, 2015).

In CRC, these kinds of analyses have shed light on the relationships between many omics features/subtypes. Continuing the example of the CMS subtypes (Guinney *et al.*, 2015), CMS1 was found to be enriched for *BRAF* mutation, hypermethylation, hypermutation, CIMP and immune pathway proteins. CMS2 was enriched for CN gain of oncogenes and loss of suppressor genes, and MYC-associated micro-(mi)RNAs, while CMS3 was enriched for *KRAS* mutations and metabolic disregulation. CMS4 had enrichment of proteins from the pathways of stromal invasion, mesenchymal activation and complement pathways.

Another example in CRC specifically addressed the question of a transcriptomic signature for the prognosis of patients treated with 5FU, and how mutations and CNAs were enriched in the high- and low-risk groups defined by this signature (Tong *et al.*, 2016). Fisher’s exact tests was used to find several arm-level features (7p, 8q, 13q, 20p, 20q amplification; 8p, 17p, 18p, 18q deletion) enriched in the high-risk group.

Extending this concept to incorporate proteomics, one study (Zhang *et al.*, 2014) determined that while CNAs had strong impacts on mRNA expression, mRNAs were not reliably associated with their translated protein’s abundance. Using hierarchical consensus clustering (Zhang *et al.*, 2014) of proteomics data, they then found five subtypes of CRC. Then, using Fisher's exact tests, they found that one subtype was enriched for hypermutation, MSI and *BRAF* mutation. These five subtypes did not significantly overlap with the CRCAssigner subtypes (Zhang *et al.*, 2014).

As the size of omics datasets has grown, the methods used for post-hoc integration have been pushed further and further. The most complex multiomic analysis of CRC to date included mutations, CNAs, miRNA, mRNA, proteins and phosphoproteins in paired tumour and adjacent normal tissues (Vasaikar *et al.*, 2019). Among the results of this study, were: i. the identification of personalised neoantigens using mutations and proteomics; ii. the detection of increased glycolysis associated with CD8 infiltration in MSI cancers, determined by proteomics and transcriptomic data; iii. that phosphorylation of RB1 is a driver of CRC proliferation that could be targeted by CDK2 inhibition, as inferred from phosphoproteomics data. These three key results give avenues of exploration for novel therapies (i. personalised cancer vaccines; ii. combined glycolysis and checkpoint inhibition; iii. CDK2 inhibition), and illustrate the power of integrative omics analysis.

There is, however, a limitation to this approach to data analysis. The discrete, pairwise tests of association between features or groups only provide information on those specific variables, and may make it harder to discover wider patterns of alterations that give a global view of CRC. This is the reasoning behind the development of the tools described below, whose focus is on the simultaneous analysis of multiple omics datasets.

1.4.2 Parallel integration of multiple omics data types

One of the conceptually simplest approaches that can be taken to simultaneously integrate multiple omics data types is to firstly cluster the samples using each data type separately, followed by clustering the samples using their class assignments in each data type ("cluster of cluster assignments").[¶] This was the methodology adopted by The Cancer Genome Atlas (TCGA) consortium to find subtypes using the vast data they collected from 3,527 patients on six omics data types (mutations, CN, methylation, m/miRNA and protein expression) (Hoadley *et al.*, 2014) and 12 cancer types (including CRC). As may have been expected, most of the samples were tightly clustered within their respective cancer type. In particular, all 255 CRC samples clustered together without exception.

We have previously applied a similar approach in uveal melanoma (Eason, Nyamundanda & Sadanandam, 2018) — although the motivation in that case was to integrate clusters from different clustering algorithms, this approach can also be applied to integrate clusters from different omics types. However, the drawback of these types of approaches to integration is that it groups samples based on their discrete binning according to each separate omics type, and does not take into account the correlations between features of different omics types. These correlations are abundant in multiomics biological data due to, e.g., direct physical interactions between molecules, or the resulting patterns of regulation across pathways and networks. Losing this information is likely to lead to an unstable, less reproducible grouping of patients.

One solution to this loss of information is to integrate data using latent variable models. The underlying conceptual assumption of these models is that there exist a set of unmeasurable or “latent” variables which explain the correlations in the data (hence, conditional on these latent variables, the features in the data are independent of each other) (Akalin, 2019). These latent variables can be interpreted as patterns of co-expressed features that are likely regulated through the same (or closely-linked) pathways. The latent variables can further be utilised for sample clustering, as the hundreds of thousands of features that represent one sample at the molecular level can be reduced to just a few latent variables, whose weighting represents the strength of the signalling in key biological networks in that sample.

Perhaps the most widely-adopted latent variable model for multiomics integration is iCluster (Shen, Olshen & Ladanyi, 2009). iCluster has been utilised to identify multiomics subtypes of cancers including breast (Shen, Olshen & Ladanyi, 2009), lung adenocarcinoma (Shen, Olshen & Ladanyi, 2009) and glioblastoma (Shen *et al.*, 2012). An extension of the original iCluster software, iClusterPlus (which can handle binary

[¶] The matrix that is finally clustered has the form

$$M = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1T} \\ s_{21} & s_{22} & \cdots & s_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ s_{N1} & s_{N2} & \cdots & s_{NT} \end{bmatrix},$$

where s_{nt} is the classification of sample $n \in [1..N]$ in the omics data type $t \in [1..T]$. The rows of M are clustered to define “clusters of clusters".

data such as mutations alongside continuous data), has previously been applied to CRC (Mo *et al.*, 2013). Using exonic mutations, copy number, promoter methylation and mRNA expression, three subtypes were identified from 189 samples:

i. Cluster 1
    - CIN positive
    - High fraction of genome altered (FGA)
    - *TP53* mutations
    - CIMP-L/0
    - MSI-L/MSS
    - Amplified chromosome 8q

ii. Cluster 2
    a. Cluster 2a
        - CIN negative
        - Low FGA
        - *BRAF* mutations
        - Hypermutated
        - CIMP-H
        - MSI-H

    b. Cluster 2b
        - CIN low
        - Moderate FGA

iii. Cluster 3
    - CIN positive
    - High FGA
    - *TP53* mutations
    - CIMP-L/0
    - MSI-L/MSS
    - Normal chromosome 8q

These groups largely overlap with the known associations of MSI, CIMP and mutations outlined in Chapter 1.1.4. Because they did not show significant prognostic power (Appendix A), it is not certain how these subtypes could be used to stratify patients in a clinical setting.

It is possible that the clinical applicability of integrated molecular analysis could be augmented by the inclusion of clinical variables within the model as covariates, rather than evaluating the model's output against clinical variables as a post-analysis. This increases the statistical power to find clinical variables associated with groups of patients. This concept will be introduced next.

1.4.3 Inclusion of clinical covariates in omics models

The addition of clinical covariates to multiomics latent variable models is little-explored in the literature. In cancer in general, some progress has been made in including clinical covariates in latent variable modelling of a single omics data type, but this does not appear to be the case in CRC. One example is phenMap (Nyamundanda, Eason & Sadanandam, 2017), developed by Dr Gift Nyamundanda from the Sadanandam Lab, which models latent variables as a function of clinical covariates. This novel extension to standard latent variable models allows for the interpretation of the latent variables in the light of important potential confounders such as age, gender, or cancer stage, facilitating a more direct clinical interpretation of the results and any putative biomarkers revealed. This is in contrast to the alternative of treating clinical covariates in the same way as molecular data, which does not facilitate understanding patterns of molecular features the in specific context of different clinical characteristics.

phenMap has not previously been applied to CRC. Notably, it can be extended to model multiple omics data types simultaneously, alongside clinical covariates (Chapter 4.2.1). The resulting latent variables can be referred to as metavariables, in reference to their representing the integration of multiple omics data types, as well as clinical variables. In Chapter 4, I will use this extended version of phenMap to model TCGA CRC data.

1.5 Summary of chapters, specific aims and hypotheses

1.5.1 Summary of chapters

In this thesis, I present several characterisations of CRC's multi-faceted heterogeneity. In Chapter 2, I demonstrate how the existing transcriptomic CRCAssigner intertumoural subtypes can be profiled in a clinically-practicable manner using fewer genes than was possible previously. Chapter 3 expands on our understanding of these subtypes by quantifying their co-existence within individual tumours, and the prognostic power and potential drug associations that knowledge of this intratumoural heterogeneity can give. In Chapter 4, I move beyond transcriptomics to delineate the mechanisms driving CRC on a pan-molecular scale, holistically integrating genomics, epigenomics, transcriptomics and proteomics with clinicopathological data. Finally, Chapter 5 summarises and discusses the implications that the results of these analyses could have for the fields of CRC research and medicine.

1.5.2 Specific aims and hypotheses

Chapter 2: Development and analytical validation of a clinically practicable assay for detecting gene expression CRC subtypes to translate intertumoural heterogeneity

Hypothesis: To assess the CRCAssigner subtypes' value prospectively, a low-cost, fast and accurate assay is required, ideally that can be applied to both fresh frozen and FFPE tissue samples. The nCounter platform can be used to profile the gene expression of tissue preserved using both these techniques, using a modified lower-cost protocol which is equivalent to the standard protocol. These gene expression profiles can then be used for subtyping.

Aims: To develop and analytically validate an nCounter-based gene expression assay to classify patients into the CRCAssigner subtypes, specifically:

1. To test the correlation of a lower-cost modified assay protocol with the standard protocol, in both fresh frozen and FFPE tissue samples
2. To test the correlation of technical replicates of the modified assay protocol, in both fresh frozen and FFPE tissue samples

3. To select an optimal gene set to use for subtype classification, from the 48 subtype-specific genes pre-selected to be included in the assay
4. To compare the results of microarray/RNA-seq-based and nCounter-based classification
5. To compare the results of nCounter-based classification in fresh frozen and FFPE tissue samples

Chapter 3: Comprehensive quantification of intratumoural subtype heterogeneity in CRC using *in vitro*-validated machine learning models in large, clinically-annotated datasets

Hypothesis: There exists substantial intratumoural transcriptomic heterogeneity, which can be studied through the quantification of intratumoural subpopulations of the CRCAssigner subtypes. This intratumoural subtype heterogeneity could have clinicallys prognostic implications and predictive power.

Aims: To quantify the level of intratumoural transcriptomic subtype heterogeneity in CRC and assess its relationship with survival outcomes and responses/sensitivity to different therapies, specifically:

1. To test the SVR method of gene expression deconvolution using co-cultured cell line and single-cell gene expression data
2. To test whether subtype subpopulation information adds significant prognostic information over bulk subtype
3. To compare the subtype subpopulations of patients, PDXs and cell lines with their drug responses/sensitivity to FOLFIRI and anti-EGFR therapy
4. To compare the subtype subpopulations of MSI/MSS patients, and the effects of subtype subpopulations on their prognosis

Chapter 4: Integrated multiomics factor analysis of CRC molecular profiles, interactions between clinicopathological categories, and prognostic implications

Hypothesis: Previous methods developed to integrate multiomics data do not model the potentially confounding effect of clinical covariates. Including these variables explicitly could increase the interpretability of multi-omics integration by adding important information about context.

Aims: To integrate multi-omics molecular data in parallel with clinical variables in order to understand correlated molecular features across data types and their specificity to different clinicopathological contexts, specifically:

1. To select appropriate features/clinical variables to include in the model that will be the most informative in downstream analyses
2. To explore the molecular features and clinical variables most highly weighted in each metavariable, and their relationship to each other
3. To test for the enrichment of known biological pathways in the highly weighted features in each metavariable
4. To test whether the metavariables are prognostic of patient survival

Chapter 2

Development and analytical validation of a clinically practicable assay for detecting gene expression CRC subtypes to translate intertumoural heterogeneity

2.1 Introduction

As described in Chapter 1.1.4, previously, CRCs were classified into five CRCAssigner subtypes with different prognoses and potential treatment responses (Sadanandam *et al.*, 2013). These subtypes and those described by five other groups were merged into a consensus subtyping scheme (CMS), that bears strong resemblance to CRCAssigner subtypes (Guinney *et al.*, 2015; Ragulan *et al.*, 2019). In this chapter, the analytical development and validation of a custom NanoString nCounter platform-based biomarker assay (*NanoCRCA*) to stratify CRC into subtypes is demonstrated.

To reduce costs, the standard nCounter protocol was switched to a custom modified protocol. The assay included a reduced 38-gene panel from the original 786-gene signature that was selected using an in-house computational pipeline of methods (fully described in Chapter 2.2.6) comprising a consensus of gene selection and class prediction methods. NanoCRCA was applied to 295 samples from 237 CRC patients. From the fresh frozen samples (n = 237), a subset had matched RNA-seq/microarray profiles (n = 47) or FFPE samples (n = 58). Further, the assay results were compared with the CMS classifier, different platforms (microarrays/RNA-seq) and gene-set classifiers (38 and 786 genes).

Full details of the datasets and methods utilised in this chapter are given in Chapter 2.2.

2.2 Methods and data sources

2.2.1 Patient cohorts and datasets

Five CRC cohorts of primary tumour samples collected prior to treatment were studied, as summarised in Table 2.1 and detailed below; three derived from fresh frozen, one from FFPE samples and one from matched fresh frozen and FFPE. The first included RNA samples from 17 stage IV patients (Montpellier cohort) from a published study (Del Rio *et al.*, 2007). A second cohort of RNA samples (OriGene; n=17) was purchased from OriGene (Rockville, MD, USA). A third cohort included 145 fresh frozen samples (Singapore) from patients participating in an on-going observational study approved by the institutional review board (Singhealth Centralised IRB — 2013/110/B). The fourth cohort consisted of 12 FFPE CRC samples from a retrospective tissue collection from The Royal Marsden Hospital, UK (RETRO-C cohort: IRB and ethical approval NRES Committee East of England-Cambridge Central: 10/H0308/28). The fifth and final cohort consisted of 58 stage II-III CRC patients (INCLIVA-Valencia cohort) with matched prospectively collected fresh frozen and FFPE tissue (ethical approval Comité Etico de Investigacion Clinica del Clínico Universitario de Valencia: F-CE-GEva-15). Cellularity in this cohort was scored by a trained pathologist on H&E stained slides. All the patients provided informed consent.

Table 2.1: Overview of patient cohorts for assay development. Sample numbers, types, platforms, and clinical characteristics. All samples are from surgical/biopsy specimens collected prior to treatment.

Cohort	Number of patients	Sample preservation	Platforms	Clinical characteristics
Montpellier	17	Fresh frozen	NanoCRCA, Microarray	Stage IV primary CRCs
OriGene	17	Fresh frozen	NanoCRCA, Microarray	Mixed stage CRCs
Singapore	145	Fresh frozen	NanoCRCA, RNA-seq (13 matched)	Mixed stage primary CRCs
RETRO-C	12	FFPE	NanoCRCA	Stage IV primary CRCs
INCLIVA-Valencia	58	Matched fresh frozen and FFPE	NanoCRCA	Stage II and III primary CRCs

Sample material from the Montpellier cohort was kindly provided by Dr Maguy Del Rio and Dr Pierre Martineau of Université Montpellier, France; from the Singapore cohort by Dr Koo Si-Lin, Dr Tan Wah Siew, and Dr Iain Beehuat Tan of the National Cancer Centre Singapore and Singapore General Hospital, Singapore; from the RETRO-C cohort by Dr Francesco Sclafani, Ms Ruwaida Begum, Dr Larissa Sena Teixeira Mendes, and Prof. David Cunningham of the Royal Marsden NHS Foundation Trust, UK; and from the INCLIVA-Valencia cohort by Dr Noelia Tarazona and Prof. Andrés Cervantes of the University of Valencia, Spain.

2.2.2 nCounter assay protocols and data normalisation

The nCounter assay hybridisation and data collection was kindly performed by Ms Chanthirika Ragulan and Dr Elisa Fontana of the Institute of Cancer Research, UK.

nCounter Max Analysis System (NanoString Technologies, Seattle, WA, USA) was used to perform the assay using either standard or modified (*Elements* chemistry) protocol as per the manufacturer's instructions. For the standard protocol, custom CodeSets (pre-built capture probes and barcoded reporter probes having sequences complementary to the target genes) for the selected genes were designed and built by NanoString Technologies. For the modified protocol, nCounter Elements TagSets (only capture and reporter tags; NanoString Technologies) and custom-designed target-specific oligonucleotide probe pairs were procured separately (Integrated DNA Technologies, Inc., Leuven, Belgium).

For both standard and modified protocols, 100 ng of total mRNA (20 ng/µL) from fresh-frozen or FFPE tissues was used. Hybridisation reactions were prepared according to manufacturers' instructions for either 18 hours at 65 °C using Standard CodeSets reagents for the standard protocol or for 20 hours at 67 °C using Elements TagSets reagents for the modified protocol. Hybridised samples were pipetted using the nCounter Prep Station and immobilised on to the sample cartridge for data quantification and collection using nCounter Digital Analyzer (NanoString Technologies). The nCounter Prep Station and Digital Analyzer together constitute the nCounter Max Analysis System.

Data quality from nCounter assays was checked and data normalization was performed using nSolver v3.0 analysis software (NanoString Technologies). Firstly, counts were corrected to background noise using geometric means of 8 negative control probes, followed by the correction using geometric means of 6 internal positive control spike-ins in each lane/sample. These negative and positive probes were built-in to both standard and modified protocols. Only those housekeeping genes with raw molecular counts greater than 50 and those selected by geNorm algorithm (part of the nSolver analysis software) were retained for further analysis. Variations due to RNA input volume were corrected by normalising to the expression of geNorm selected housekeeping genes. The normalised final count data were $\log_2$ transformed for further analysis.

2.2.3 Microarray/RNA-seq protocols and data normalisation

OriGene microarray data was kindly generated by Ms Chanthirika Ragulan of the Institute of Cancer Research, UK.

For the OriGene cohort, 100 ng of total RNA was used for first strand cDNA synthesis and labelled according to the manufacturer's protocol. Labelled single stranded cDNA was hybridised in GeneChip Human Transcriptome Array (HTA) 2.0 (Affymetrix, High Wycombe, UK) then arrays were washed (Gene Chip Fluidics station 450) and scanned (Gene Chip Scanner).

RNA-seq data for the Singapore cohort was kindly provided by Dr Iain Beehuat Tan of the Singapore General Hospital, Singapore, with additional advice provided by Mr Yatish Patil of the Institute of Cancer Research, UK.

RNA-seq libraries were prepared using TruSeq Stranded mRNA Library Prep Kit (Illumina, Singapore). Libraries were quality controlled using KAPA qPCR (Roche, Singapore) and Agilent Bioanalyzer, before pooling and sequencing on the Illumina HiSeq (Illumina) to a median of 22 million paired reads per sample. Fastq files were checked for read counts for paired end reads and read quality using *fastqc* (v0.11.4) (Andrews, 2016). All 17 samples had mean Phred score greater than 34. Mapping quality was checked using *RSEM* (v1.2.22) (Li & Dewey, 2011) - *samtools-flagstat* (1.3.1) (Li *et al.*, 2009) using reference transcriptome (GRCh37). RNA quality (ribosomal, coding, intronic and intergenic) was checked using the *CollectRnaSeqMetrics* function of *picard* (v2.1.0) (Broad Institute, 2015) on mapping to reference genome (GRCh37). Transcripts per million (TPM) values were calculated using *RSEM*, genes with <20% missing values were retained and $\log_2(TPM + 1)$ transformed.

2.2.4 Clustering and heatmaps of gene expression

All sample clustering was performed using Euclidean distance and complete linkage as implemented in the *hclust* and *dist* functions of the R package *stats* (v3.3.2) (R Core Team, 2017). Arc plots were generated using the *circlize* package (v0.4.0) (Gu *et al.*, 2014). Heatmaps were plotted from gene-wise median-centred expression data thresholded to $[-3, 3]$ using the *heatmap.plus* package (v1.3) (Day, 2012)

2.2.5 Subtype assignment

CRCA subtypes were assigned by performing Pearson correlation of gene-wise median-centred expression profiles for each sample with corresponding centroids for the subtypes. The subtype with the highest correlation was then assigned to that sample. Samples were marked as having undetermined subtype if the sample's correlation with the subtype centroid had a value (Pearson's r) ≤ 0.15, or if the correlation was high for multiple subtype centroids (Pearson's r difference between first and second highest subtypes ≤ 0.06), in line with the published CMS classifier (Guinney *et al.*, 2015).

CMS subtypes were determined from microarray or RNA-seq data using the *CMSclassifier* R package (v1.0.0)(Reynies & Guinney, 2015) and the *classifyCMS* function, using the single sample prediction (SSP) classifier (Guinney *et al.*, 2015).

2.2.6 Gene selection pipeline

This gene selection pipeline was developed by Dr Gift Nyamundanda of the Institute of Cancer Research, UK.

Having selected samples with minimal intratumoural subtype heterogeneity using the support vector regression (SVR) method fully described in Chapter 3.2.3, these samples are passed to a gene selection pipeline detailed here (dubbed "intPredict" (Ragulan *et al.*, 2019)). Firstly, the samples (and their known subtypes) are randomly partitioned 50 times into training and test sets (Monte Carlo cross-validation). On each training set, prediction strength (Golub *et al.*, 1999), PAM (Tibshirani *et al.*, 2002) and between-within group sum of squares ratio (Dudoit, Fridlyand & Speed, 2002) are used to select the genes that are most strongly discriminant of the classes, according to each method. This process is repeated across the range of possible gene set sizes (in this case $[2, 48]$).

Each of the gene sets obtained from the training sample sets, as well as their respective test sample sets, are then passed to four subtype prediction methods: random forest (Breiman, 2001), diagonal linear discriminant analysis (Dudoit, Fridlyand & Speed, 2002), and linear and radial SVM (Cortes & Vapnik, 1995). The number of genes p with the lowest median MCR – defined as the proportion of samples in the training set that are assigned an incorrect subtype based on these subtype prediction methods – is then selected. The p most frequently identified genes from the gene selection methods are then taken as the final gene set.

2.3 Results

2.3.1 Evaluation of a modified lower-cost protocol for nCounter assays in fresh frozen tissue

In the standard chemistry nCounter protocol (Geiss *et al.*, 2008), two probes having complementary sequences to different sections of the target mRNA are hybridised to said target (Figure 2.1a). One probe has the purpose of capturing the mRNA for adherence to a cartridge surface that immobilises it, using a "capture tag". The second probe barcodes the target for subsequent identification. This is achieved using a DNA backbone to which a pre-defined sequence of fluorophore-labelled RNA segments have been annealed. Using the nCounter Digital Analyzer, the cartridge can be optically scanned, barcodes identified, and mRNAs quantified.

More flexibility can be achieved using the nCounter *Elements* protocol. Here, the two probes described above are assembled by the user using barcodes and capture tags manufactured by NanoString, and target-specific oligonucleotides procured by the user (Figure 2.1b). This approach allows for more cost-effective assay runs as large batches of oligos can be sourced at relatively low cost.

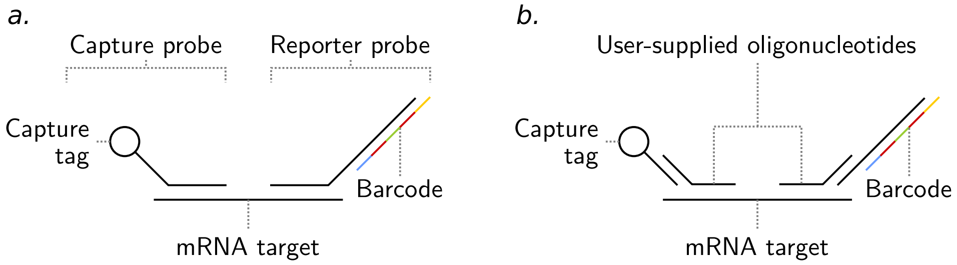


Figure 2.1: Two protocols (chemistries) for quantification of mRNA using the NanoString nCounter platform. *(a)* Standard chemistry probes and an mRNA target. *(b)* Modified (*Elements*) chemistry probes and an mRNA target. In the modified chemistry, the user procures oligonucleotides that attach the capture tag and the barcode to the mRNA target.

For both the standard and modified (*Elements*-based) protocols, the same initial set of 50 genes were selected for profiling (prior to the commencement of this PhD project, by Dr Anguraj Sadanandam of the Institute of Cancer Research), 48 of which were from the original 786-gene CRCAssigner signature, henceforth refered to as “CRCA-786”. The 50 genes were initially selected based on the original report (Sadanandam *et al.*, 2013) and criteria fully defined in Appendix B. Briefly, these constituted the qRT-PCR/IHC marker genes proposed in said report; genes that scored highly in the CRCA-786 centroids; genes which may distinguish cetuximab-resistant from cetuximab-sensitive samples (Sadanandam *et al.*, 2013); and genes from signalling pathways characteristic of each subtype (Sadanandam *et al.*, 2013).

2.3.1.1 Description of data collected using nCounter

Data collected using the nCounter platform comes in the form of integer counts per barcode, which uniquely identify the genes of interest, synthetic DNA positive controls, and negative control probes that should not hybridise. The counts of endogenous genes can range from as low as 25, to over 50,000 (Geiss *et al.*, 2008). The normalisation procedure (fully described in Chapter 2.2.2) includes correction for positive/negative controls and housekeeping genes. This normalisation shifts the data into non-integer continuous values, which are then $\log_2$ transformed, and a final matrix of genes of interest by samples was used for downstream analysis.

2.3.1.2 Concordance of modified with standard protocol profiles

Firstly, we sought to identify whether the modified protocol could accurately reproduce the gene expression profiles measured using the standard protocol (using paired aliquots of RNA extracted from the same piece of tissue, to avoid spatial sampling effects). I performed hierarchical clustering on the normalised (gene-wise median centred (Sadanandam *et al.*, 2013)) gene expression measured using the standard and modified protocols in 22 CRC samples (from the Montpellier and OriGene cohorts; Chapter 2.2.1), which showed clustering of profiles based on the source sample, rather than the protocol used (Figure 2.2).

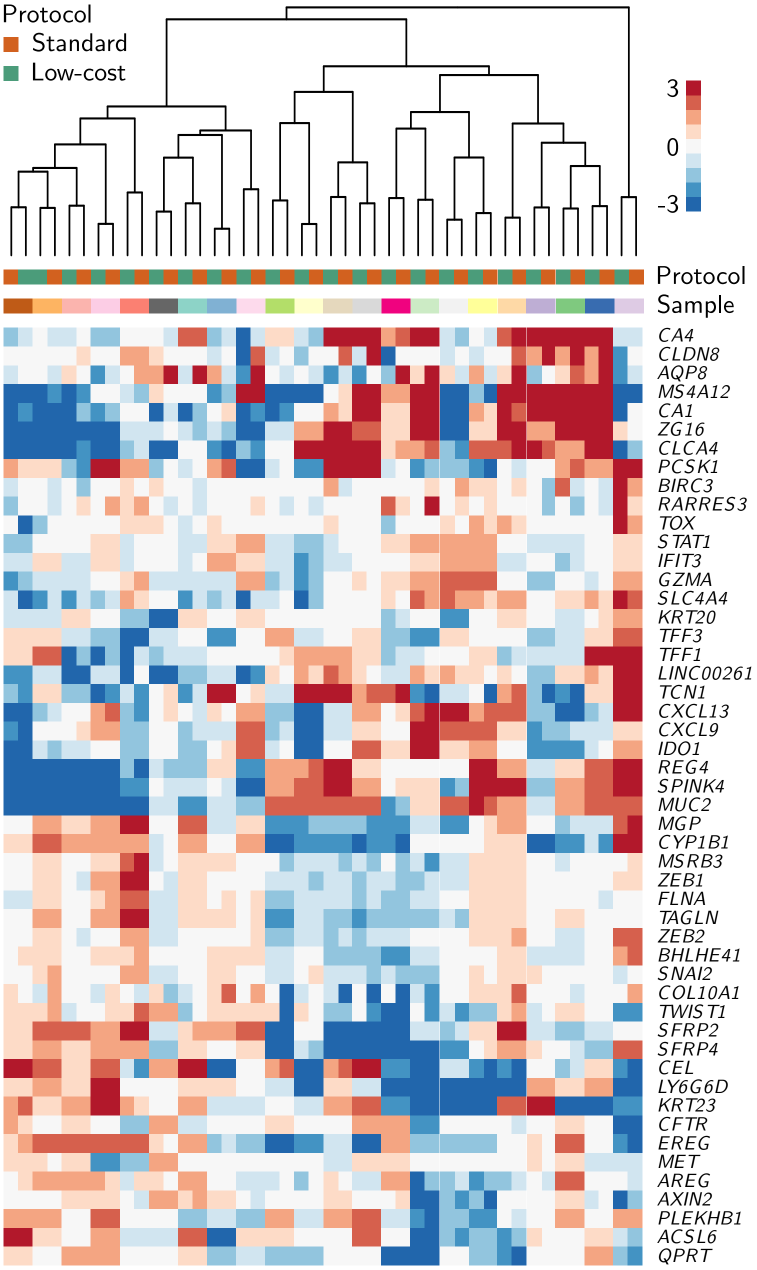


Figure 2.2: Repeat measurements of fresh frozen samples using the standard and modified protocols cluster by sample, rather than by protocol. Hierarchical clustering and heatmap of gene expression of samples profiled using standard and modified assay protocols ($n = 22$). nCounter data collected by Ms. Chanthirika Ragulan.

I then merged data from each protocol after normalisation and subjected them to prinicpal component analysis (PCA), which demonstrated high similarity between samples measured on the two protocols (Figure 2.3a). Calculating the correlation coefficient between the two datasets confirmed they were highly concordant (Pearson's $r = 0.90$; $p < 0.001$; Figure 2.3b).

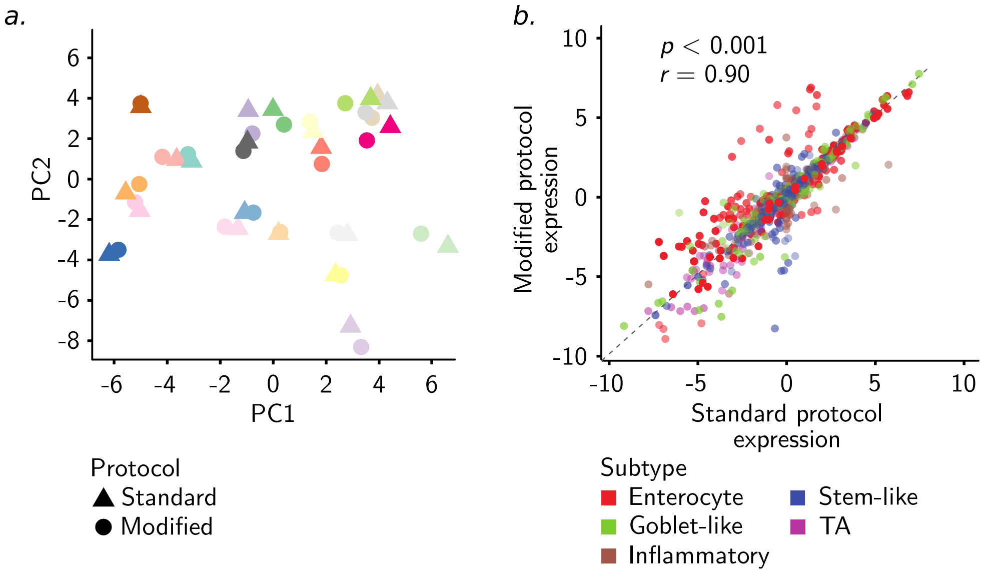


Figure 2.3: Gene expression profiles of fresh frozen samples measured using the standard and modified protocols are highly concordant. *(a)* PCA analysis and *(b)* scatter plot of normalised gene expression profiles from the standard and modified assay protocols ($n = 22$). Colours in *a* and *b* represent the sample and the subtype association of each gene, respectively. nCounter data collected by Ms. Chanthirika Ragulan.

2.3.1.3 Technical reproducibility of modified protocol

To test the reproducibility of the modified assay protocol, gene expression profiles were measured from 5 samples twice, on dates a maximum of 40 weeks apart. Figure 2.4 shows the clustering of replicates, and Figure 2.5a the grouping of replicate samples in the principal space, with no visible batch effect. There was high correlation between the replicates (Pearson's $r = 0.98$; $p < 0.001$; Figure 2.5b). This establishes the high reproducibility of this assay over non-negligible periods of time. It was not assessed whether RNA dropped in quality between timepoints, because RNA from each replicate was thawed immediately prior to running each assay, meaning the quality should be consistent for each replicate.

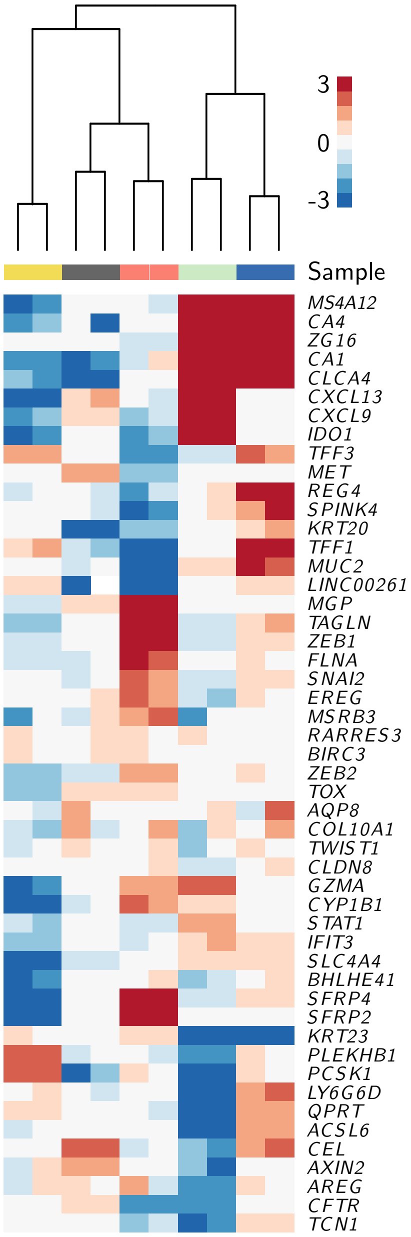


Figure 2.4: Replicates of fresh frozen samples using the modified protocol cluster together. Hierarchical clustering and heatmap of gene expression of replicates of samples profiled using the modified assay protocol ($n = 5$). nCounter data collected by Ms. Chanthirika Ragulan.

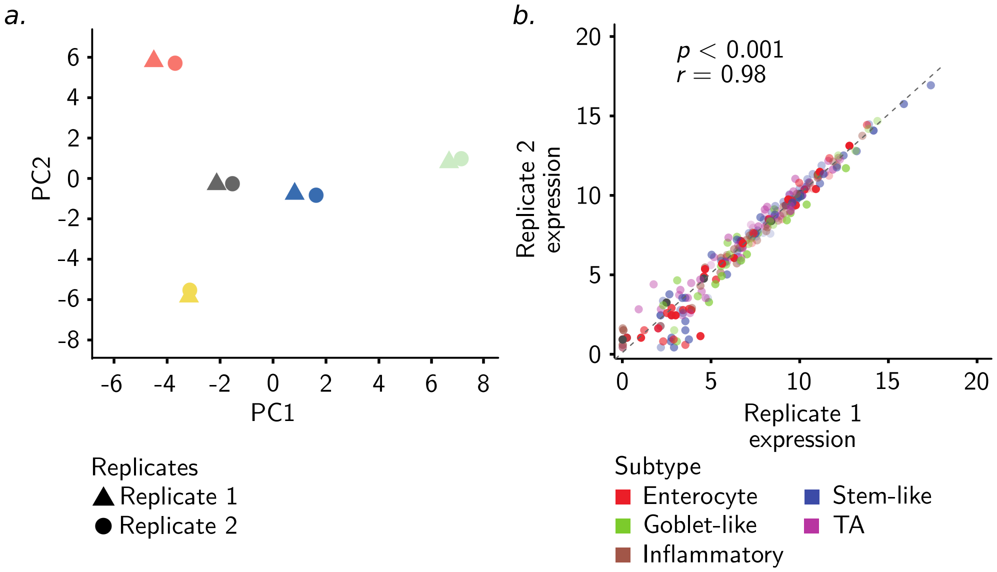


Figure 2.5: Gene expression profiles measured using replicate fresh frozen samples and the modified protocol are highly concordant *(a)* PCA analysis and *(b)* scatter plot of normalised gene expression profiles from replicates of the modified assay protocol ($n = 5$). Colours in *a* and *b* represent the sample and the subtype association of each gene, respectively. nCounter data collected by Ms. Chanthirika Ragulan.

2.3.2 Evaluation of a modified lower-cost protocol for nCounter assays in FFPE tissue

Although FFPE samples may contain a low abundance or highly degraded RNA, they also represent the most frequently available type of samples for diagnosis and biomarker assessment (Chapter 1.2.1). Therefore, an efficient biomarker assay for FFPE samples is crucial for routine clinical application.

Details of the RETRO-C cohort of samples used for this analysis are given in Chapter 2.2.1.

2.3.2.1 Concordance of modified with standard protocol profiles and technical reproducibility

I repeated the analysis as described in Chapter 2.3.1 for fresh frozen samples for FFPE tissues. This again achieved successful clustering of gene expression profiles by sample, rather than by protocol (Figures 2.6 and 2.7a). Correlation of gene expression between the two profiles near-matched that of fresh frozen tissue (Pearson's $r = 0.88$; $p < 0.001$; Figure 2.7b).

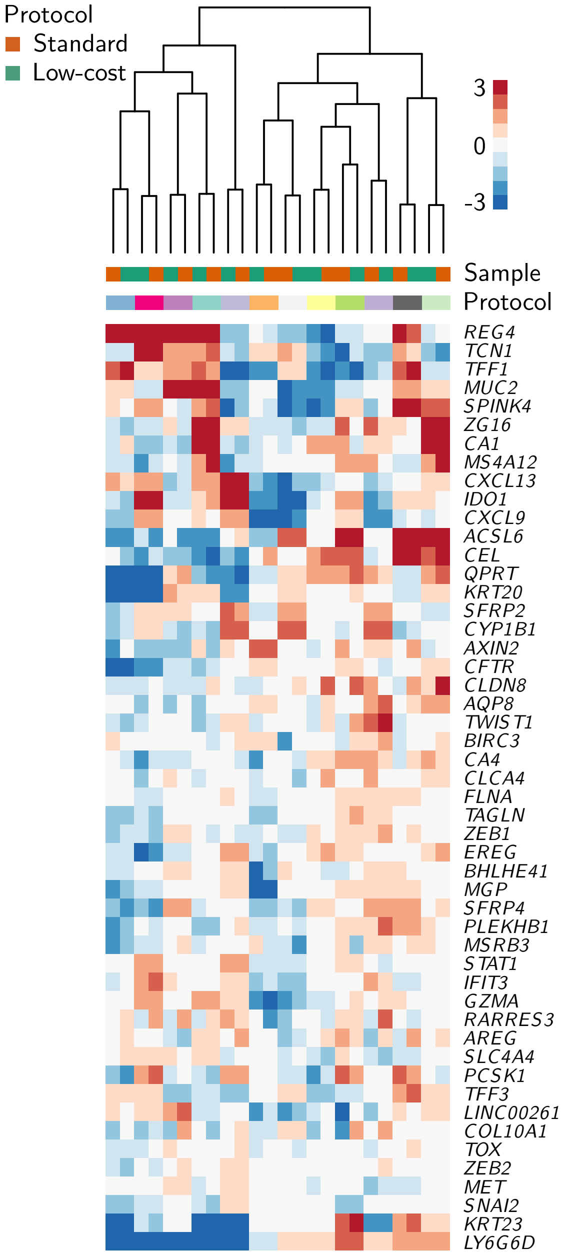


Figure 2.6: FFPE gene expression profiles show the same levels of expression regardless of assay protocol used. Hierarchical clustering and heatmap of gene expression of FFPE samples profiled using standard and modified assay protocols ($n = 12$). nCounter data collected by Ms. Chanthirika Ragulan and Dr Elisa Fontana.

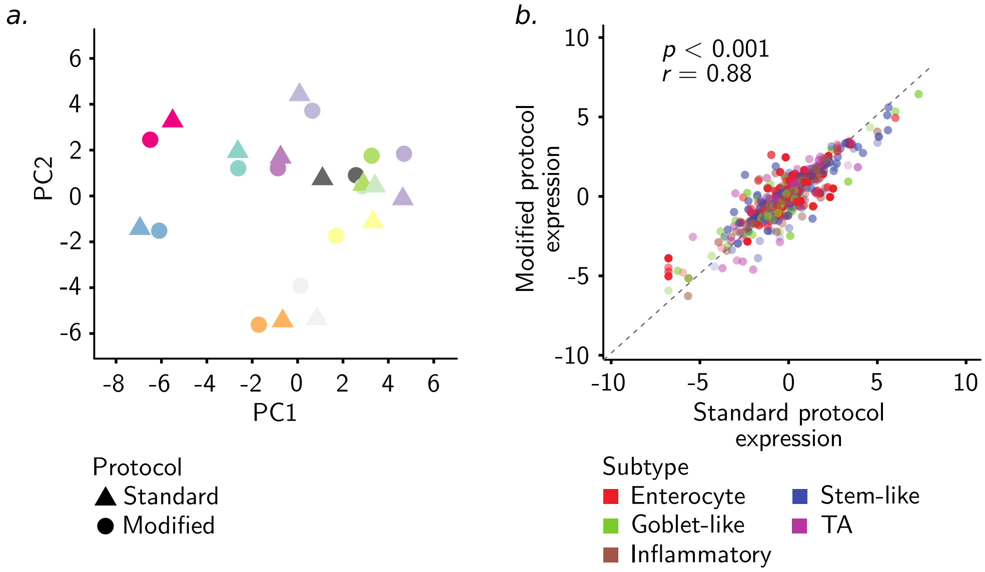


Figure 2.7: Gene expression profiles measured on FFPE samples measured using the standard and modified protocols are highly concordant *(a)* PCA analysis and *(b)* scatter plot of normalised gene expression profiles from the standard and modified assay protocols in FFPE tissue ($n = 22$). nCounter data collected by Ms. Chanthirika Ragulan and Dr Elisa Fontana. Colours in *a* and *b* represent the sample and the subtype association of each gene, respectively.

Moreover, 5 pairs of technical replicates also showed highly reproducible results (Figures 2.8 and 2.9a) with a correlation of 0.96, similar to that for fresh frozen samples (Figure 2.8b).

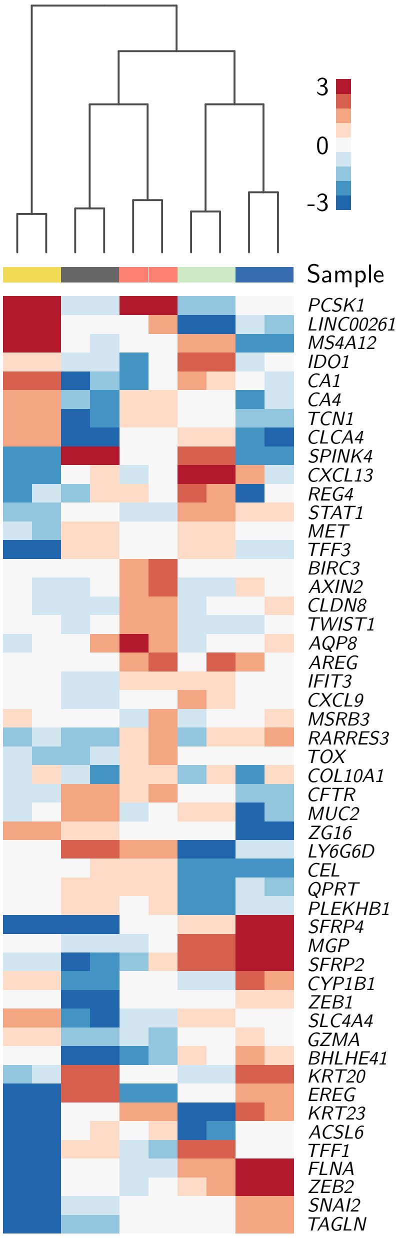


Figure 2.8: Replicates of FFPE samples using the modified protocol cluster together. Hierarchical clustering and heatmap of gene expression of replicates of samples profiled using the modified assay protocol in FFPE tissue ($n = 5$). nCounter data collected by Ms. Chanthirika Ragulan and Dr Elisa Fontana.

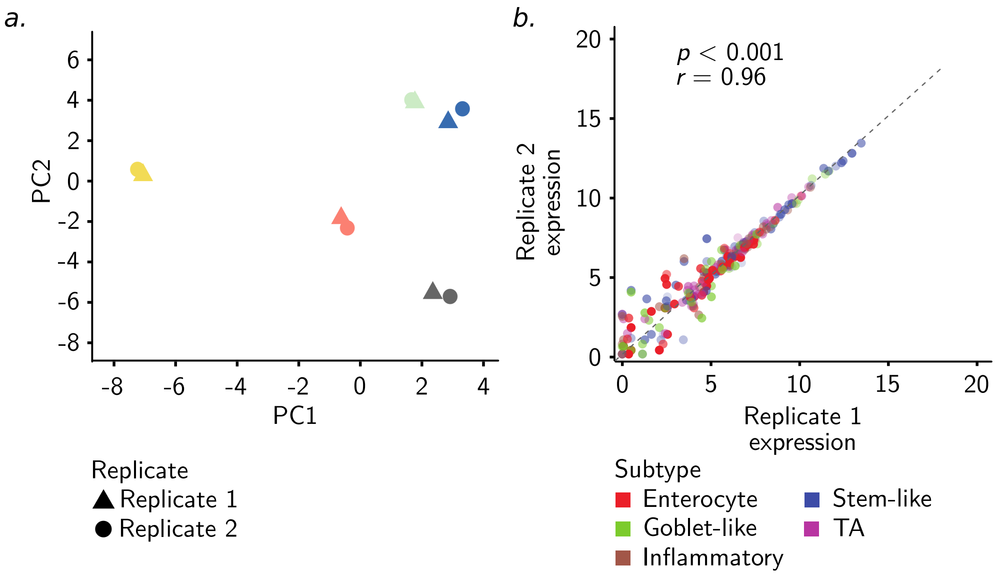


Figure 2.9: Gene expression profiles measured on replicate FFPE samples are highly concordant. *(a)* PCA analysis and *(b)* scatter plot of normalised gene expression profiles from replicates of the modified assay protocol in FFPE tissue ($n = 5$). Colours in *a* and *b* represent the sample and the subtype association of each gene, respectively. nCounter data collected by Ms. Chanthirika Ragulan and Dr Elisa Fontana.

2.3.3 Gene selection and subtype centroid design

Successful clinical biomarker assays should be able to classify samples into subtypes with high concordance, and this requires a robust set of genes. Hence, the accuracy of the 48-gene preliminary gene set was tested using two in-house bioinformatics tools, described briefly below, and in detail in Chapters 2.2.6 and 3.2.3.

2.3.3.1 Selection of maximally subtype-homogeneous samples

Since mixed subtype samples comprising more than one subtype are present in CRC (Guinney *et al.*, 2015), only the samples from the original CRCAssigner dataset (Sadanandam *et al.*, 2013) that showed at least 70% of their gene expression attributable to a single subtype were used for gene selection. This was achieved using an SVR method based on CIBERSORT (Newman *et al.*, 2015) (Figure 2.10a; method described in full detail in Chapter 3.2.3).

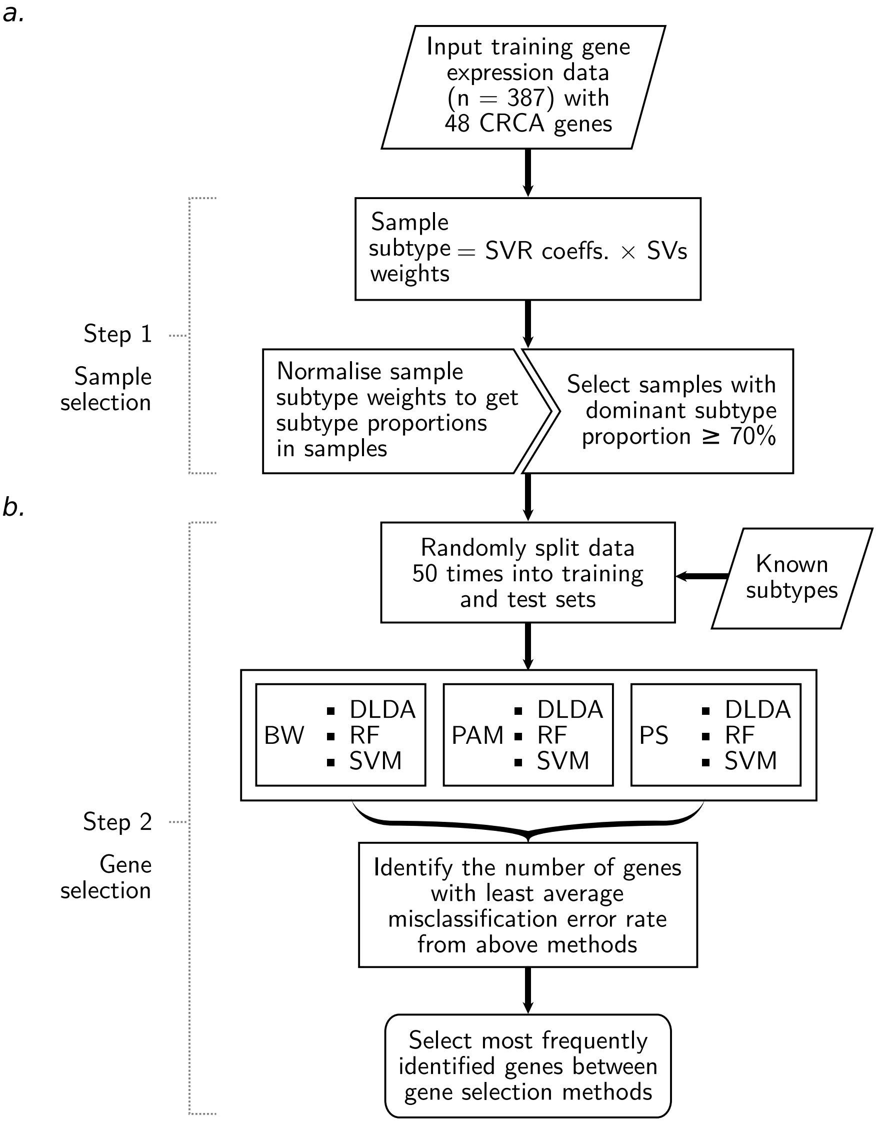

a.
Input training gene expression data (n = 387) with 48 CRCA genes
Sample subtype weights = SVR coeffs. × SVs
Step 1
Sample selection
Normalise sample subtype weights to get subtype proportions in samples
Select samples with dominant subtype proportion ≥ 70%
b.
Randomly split data 50 times into training and test sets
Known subtypes
BW
PAM
PS
DLDA
RF
SVM
Step 2
Gene selection
Identify the number of genes with least average misclassification error rate from above methods
Select most frequently identified genes between gene selection methods

Figure 2.10: Training sample selection and gene set selection pipeline. Overview of the process and pipelines used to select a robust gene set for the NanoCRCA assay using in-laboratory developed computational tools, *(a)* sample selection and *(b)* gene selection. SVR: Support vector regression; SV: support vector; BW: Between-within sum of squares; DLDA: Diagonal linear discriminant analysis; RF: Random forest; SVM: Support vector machine; PAM: Prediction analysis of microarrays; PS: Prediction strength.

2.3.3.2 Minimisation of gene set and centroid derivation

Furthermore, I used the second tool, which comprises a pipeline of supervised class prediction methods (Chapter 2.2.6), to identify the 38 genes (out of the 48 subtype genes measured on the panel, which were selected prior to the commencement of this PhD project) that classify samples into their known subtypes (Sadanandam *et al.*, 2013) with the lowest MCR (Figure 2.10b). In order to subsequently classify samples into CRC subtypes using the selected 38 genes, I derived new 38-gene centroids (*CRCA-38*) using PAM (Tibshirani *et al.*, 2002).

2.3.4 Subtyping of fresh frozen tissue samples using multiple gene sets and platforms

To determine if the 38-gene centroids and modified assay protocol could be used to accurately classify new samples into CRCAssigner subtypes, the *NanoCRCA* assay was applied to fresh frozen CRC samples (n = 179; combined from the Montpellier, Singapore and OriGene cohorts). I subtyped samples by calculating the correlation of gene expression profiles with the CRCA-38 centroids.

All 5 subtypes were identified by the assay (with 89% (159/179) of the samples being classifiable) and demonstrated distinct patterns of gene expression (Figure 2.11). There was some expression of goblet-like and TA genes in samples classified as belonging to the enterocyte subtype. Because the CRCAssigner subtypes are distinguished by their expression of genes in normal colon cells, this could be a result of the proximity of enterocyte, goblet-like and TA cells in the colon crypt (Figure 1.2), and their overlapping phenotypic traits. There is also some expression of stem-like genes in inflammatory subtype samples. Because stem-like genes can be expressed in stroma (Isella *et al.*, 2015), this could be due to higher stromal infiltration in the inflammatory subtype.

A small proportion of samples (11%) were found to be of undetermined subtype (in that they cannot be classified into any one of the 5 subtypes), either because they have a mixed subtype or are of poor quality, attributes that were determined using correlation coefficient cut-offs as discussed in the CMS publication (Guinney *et al.*, 2015) (Chapter 2.2.5).

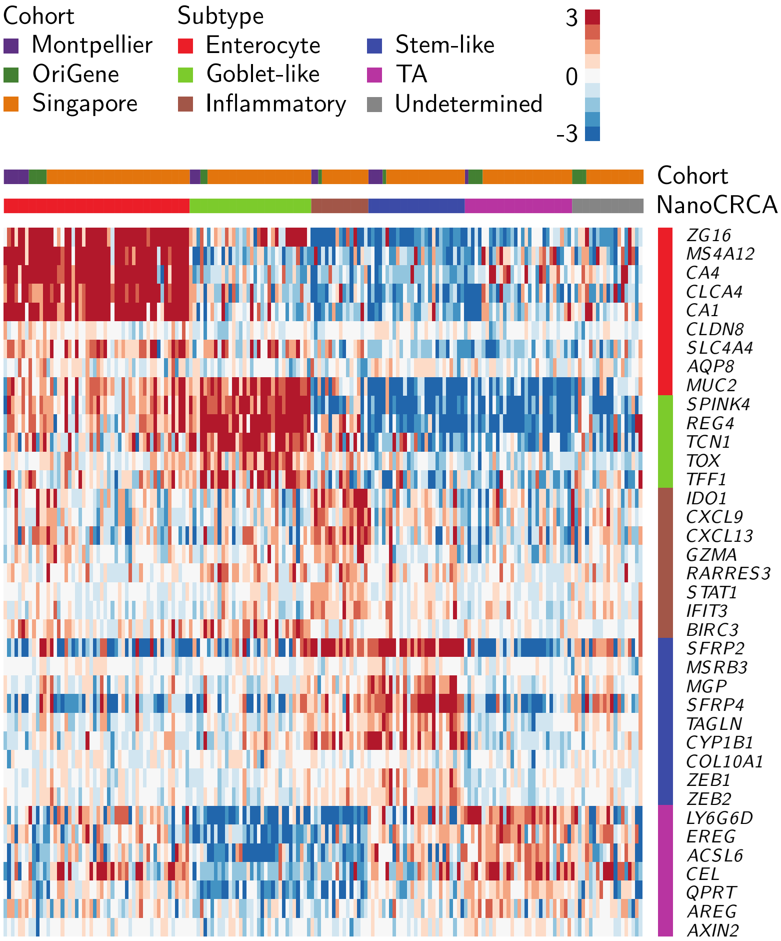


Figure 2.11: nCounter gene expression profiles of the 38-gene panel show samples expressing the characteristic genes of each subtype Heatmap showing the expression of the 38 genes in the fresh frozen samples, as measured by NanoCRCA (n = 179). nCounter data collected by Ms. Chanthirika Ragulan and Dr Elisa Fontana.

I then assessed whether subtyping using the NanoCRCA assay mirrored the results of subtyping using CMS subtypes and platforms such as microarrays and RNA-seq. Matched microarray or RNA-seq data for the fresh frozen Montpellier, Singapore and OriGene cohorts were generated or obtained from public repositories (Chapter 2.2.1; n = 47). I determined samples' subtypes by correlation of the RNA-seq/microarray gene expression profiles with both the new CRCA-38 and original 786-gene centroids (Sadanandam *et al.*, 2013) (*CRCA-786*). I predicted CMS subtypes using the CMS classifier (Guinney *et al.*, 2015).

Figure 2.12 shows the expression of the CRCA-38 classifier genes in these samples as measured by the NanoCRCA assay, alongside their subtypes as assigned by: NanoCRCA; the RNA-seq/microarray platform plus the CRCA-38 classifier; the RNA-seq/microarray platform plus the CRCA-786 classifier; and the RNA-seq/microarray platform plus the CMS classifier.

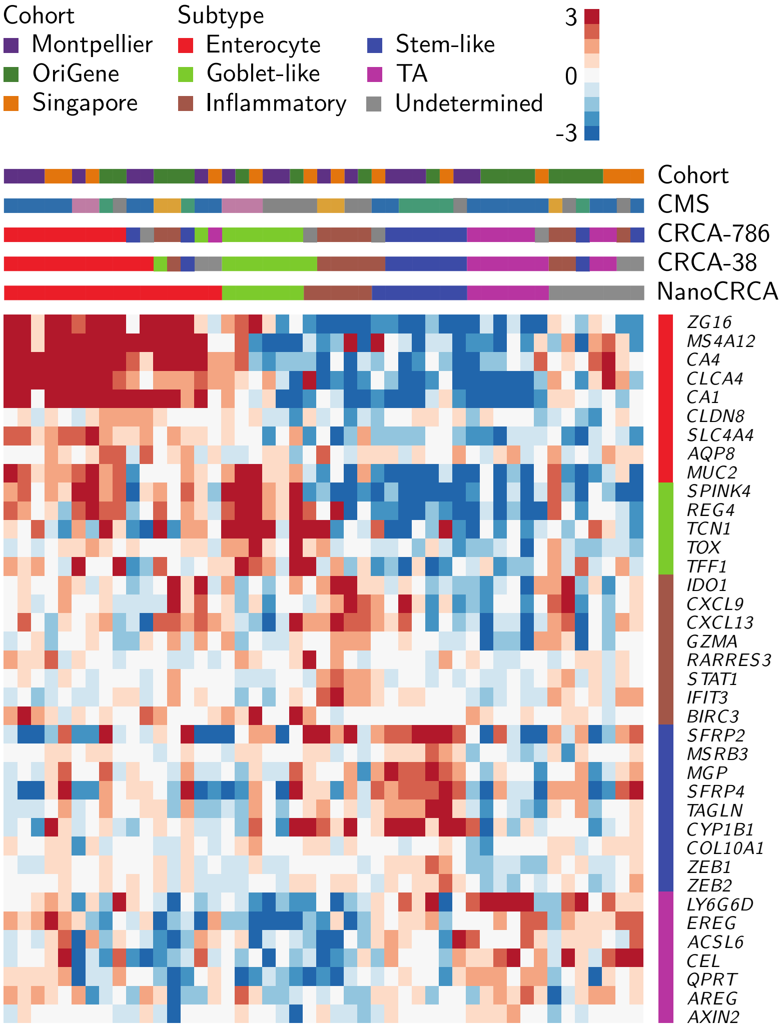


Figure 2.12: Expression as measured on the nCounter platform is concordant with the subtypes predicted by RNA-seq or microarrays. Heatmap showing the expression of the 38 genes in the fresh frozen samples having matched RNA-seq/microarray data, as measured by NanoCRCA ($n = 47$). nCounter data collected by Ms. Chanthirika Ragulan and Dr Elisa Fontana.

To confirm that platform and gene set differences did not bias the distribution of subtypes assigned to the samples, I pooled the subtypes from the three cohorts with matched RNA-seq/microarray data ($n = 47$) and tested for enrichment of each subtype in each classifier/platform combination. Figure 2.13 shows there is no significant difference ($p > 0.05$, proportion test) in the distribution of each subtype across the three CRCA assays.

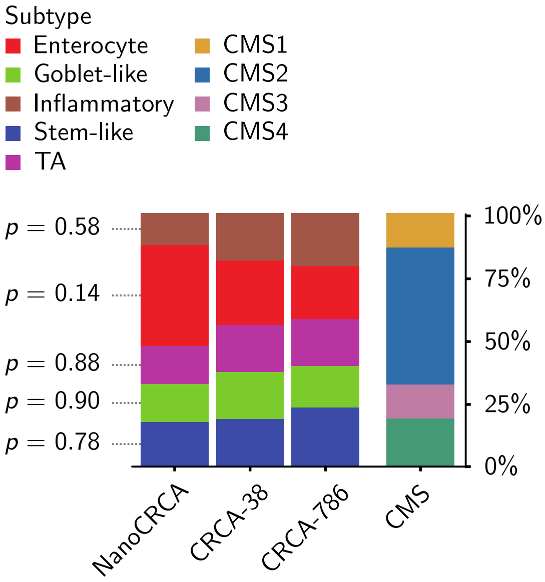


Figure 2.13: There is no significant difference in the distribution of the subtypes found using each platform. Bar chart showing the proportion of subtypes as classified by each classifier/platform combination ($n = 47$).

Additionally, when I performed pairwise comparisons between all 4 classifier/platform combinations (including CMS, Appendix C), all assays were significantly associated with all other assays ($p < 0.001$, Fisher's exact test), whether undetermined samples were considered or not.

2.3.4.1 Concordance of subtyping using original and reduced gene sets

Concordance was highest between the CRCA-38 and CRCA-786 classifiers (Appendix C), demonstrating high agreement between these different gene sets on the RNA-seq/microarray platforms (95%; 37/39; $p < 2.2 \times 10^{-16}$). This indicates that the reduction of the gene set has not caused misclassification between subtypes. There were 4 samples that were classified as undetermined by CRCA-38, but were classifiable by CRCA-786. Although these samples were classifiable by the 786-gene signature, they did display expression of signature genes from multiple subtypes (Figure 2.14). This intratumoural heterogeneity may have caused this change in classification between gene sets, and is explored extensively in Chapter 3.

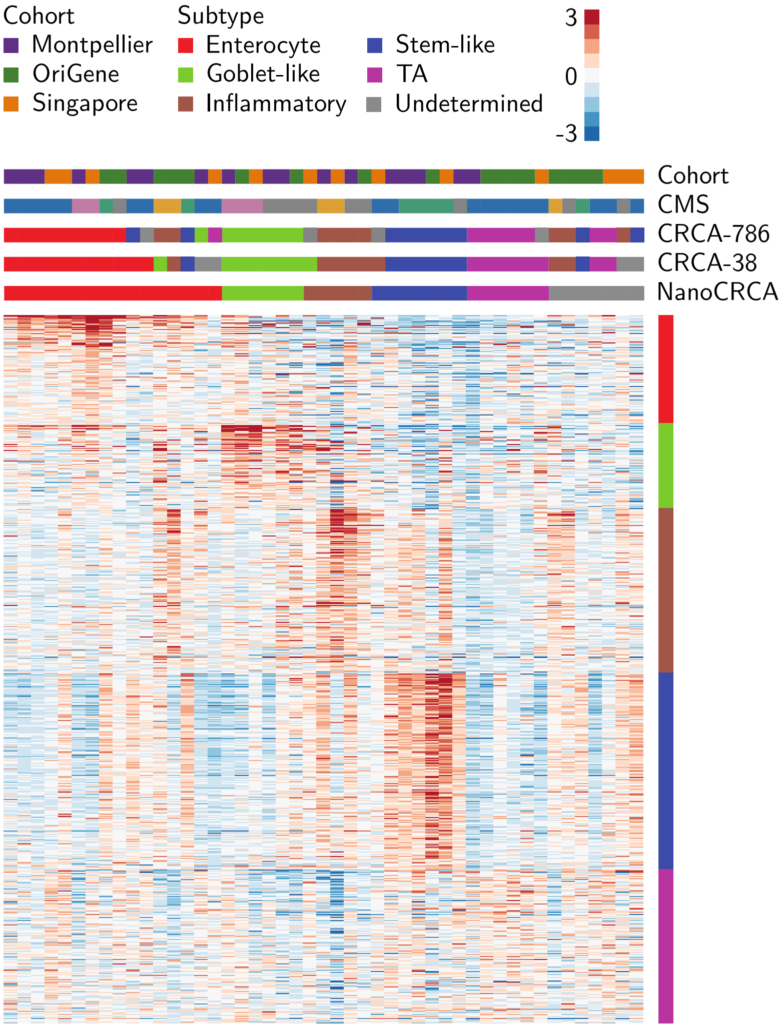


Figure 2.14: Subtypes predicted using the NanoCRCA assay align with gene expression profiles measured using RNA-seq or microarrays. Heatmap showing the expression of the 786 genes in the fresh frozen samples as measured by RNA-seq/microarrays ($n = 47$). Gene expression data collected from various sources; see Chapter 2.2.

2.3.4.2 Concordance of subtyping using the reduced gene set using nCounter and RNA-seq/microarrays

NanoCRCA was concordant with the CRCA-38 RNA-seq/microarray classifications at 87% (33/38; $p = 3.3 \times 10^{-16}$; Appendix C), despite the fundamental differences between these platforms. This indicates that the 38-gene classifier used to classify samples in both these cases is generally robust to these technical differences.

However, the largest difference between the NanoCRCA and CRCA-38 classifications came from the increased number of samples classified as enterocyte by NanoCRCA. The majority of samples that were classified into a different subtype by NanoCRCA than CRCA-38, were classified by NanoCRCA as enterocyte (Figure 2.12 and Appendix C).

Figure 2.15 shows the mean measured expression of the 38 genes in samples that were classified as enterocyte by NanoCRCA, but not by CRCA-38. This reveals an increase in the levels of the genes *CA4* and *ZG16* measured by NanoCRCA. These genes have the first and third highest weights, respectively, in the 38-gene centroids used to classify data from both CRCA-38 and NanoCRCA. Hence, high measured expression of these genes is likely to result in the classification of a sample as being enterocyte. How often this technical difference between platforms occurs will need to be assessed in a larger cohort.

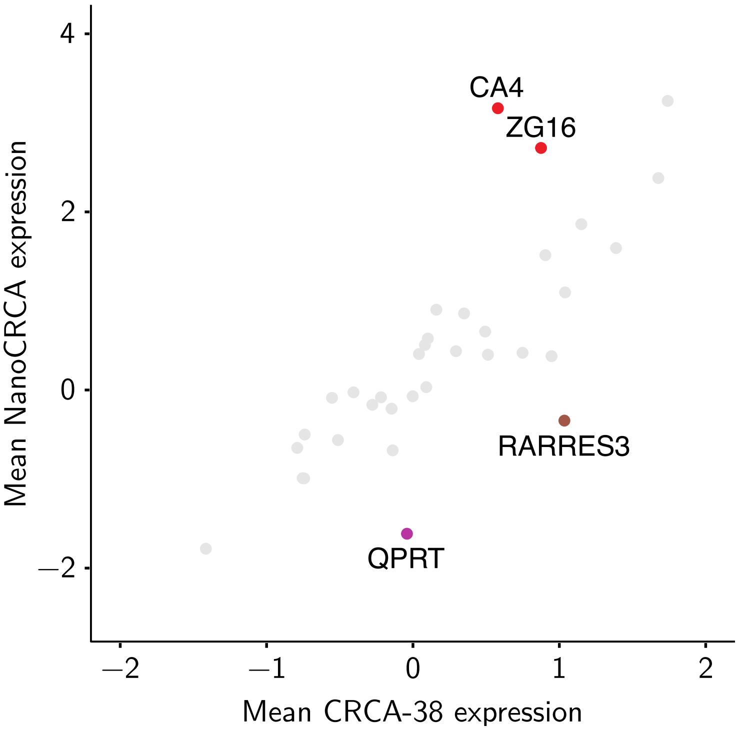


Figure 2.15: Increased measured expression of enterocyte genes *CA4* and *ZG16* changes the classification of samples to enterocyte in the NanoCRCA assay. Scatter plot showing the mean expression of the 38 genes in samples classified as enterocyte by NanoCRCA, but as other subtypes by CRCA-38 ($n = 5$). Highlighted genes have expression greater than 1 standard deviation away from a linear regression line. Gene expression data collected from various sources; see Chapter 2.2.

2.3.5 Subtyping of tumour-matched fresh frozen and FFPE tissue samples

Finally, the equivalence of subtyping using FFPE and fresh frozen tissue must be established, for the reasons of clinical applicability set out in Chapter 1.2.1. This analysis was performed using the INCLIVA-Valencia cohort detailed in Chapter 2.2.1.

2.3.5.1 Differential cellularity between matched samples and effects on subtyping

Tumour cellularity (as scored by a pathologist) between the fresh frozen and FFPE samples often differed (Figure 2.16) - in the most extreme case, by 60%. This could have a detrimental effect on subtyping, as when normal colon tissue is classified into the CRCA subtypes it falls near-exclusively into the enterocyte subtype (Appendix D), indicating that there is a high level of expression of enterocyte subtype genes in normal colon epithelium. Indeed, when I performed subtyping of all samples regardless of cellularity, only 45% (26/58) of samples were classified into the same subtype using their fresh frozen and FFPE tissues.

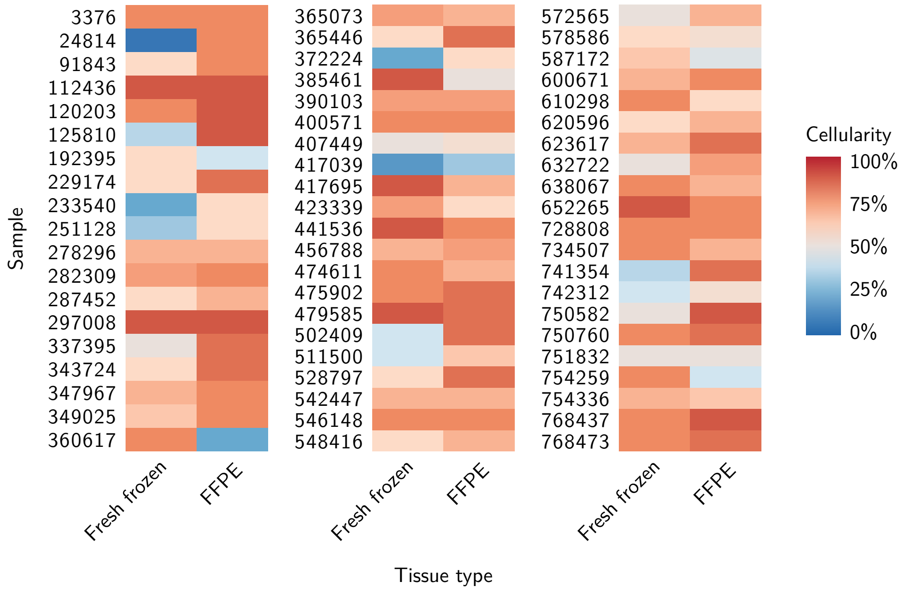


Figure 2.16: The same samples' tissue preserved by freezing or FFPE can have different levels of tumour cellularity. Heatmap showing the tumour cellularity of samples in their matched fresh frozen and FFPE tissues ($n = 58$).

While there was a significant difference in cellularity between the subtypes called using fresh frozen tissues (Figure 2.17a), there was no such difference between the FFPE tissue-derived subtypes (Figure 2.17b). As the FFPE tissues had been macrodissected prior to RNA extraction, this indicated that cellularity was creating a bias in subtype assignment because of the high expression of enterocyte genes in normal colon epithelium, rather than the alternative conclusion that different cancer subtypes have inherently differing levels of cellularity due to interactions with the tumour microenvironment.

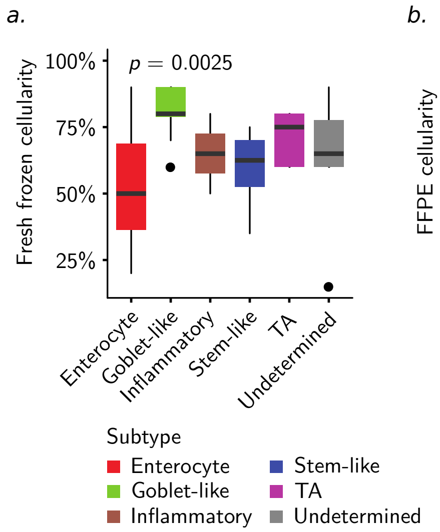


Figure 2.17: Tumours classified as enterocyte by their fresh frozen tissue samples have lower cellularity than other subtypes. Box plot showing the cellularity of samples of each subtype in *(a)* fresh frozen and *(b)* FFPE samples ($n = 58$).

2.3.5.2 Subtype concordance in selected matched samples

For subsequent analysis of fresh frozen and FFPE subtyping concordance, I considered only the 24 samples having $\geq 70\%$ cellularity in both tissue types in order to minimise the effect described in the previous section.

All of the samples – excluding those classified as enterocyte by their fresh frozen tissue – had the same subtype in the fresh frozen and FFPE tissues (Figure 2.18). However, fresh frozen enterocyte samples were classified as TA or stem-like by their FFPE tissues. Due to the limited sample size, the reason for this misclassification cannot be determined from this dataset (only two samples were misclassified, precluding stastical analyses from finding significant associations with confounding variables), and further investigation is required to conclude if this is a systematic effect by gathering a larger cohort of matched fresh frozen and FFPE tissues. Speculatively, this could partially be explicable by the similarity of the enterocyte and TA subtypes, that combined together form the CMS2 subtype in the CMS classification (Guinney *et al.*, 2015). It is possible that the enterocyte-related genes from the 38-gene classifier are subject to more variation between fresh frozen and FFPE tissues than genes characteristic of other subtypes.

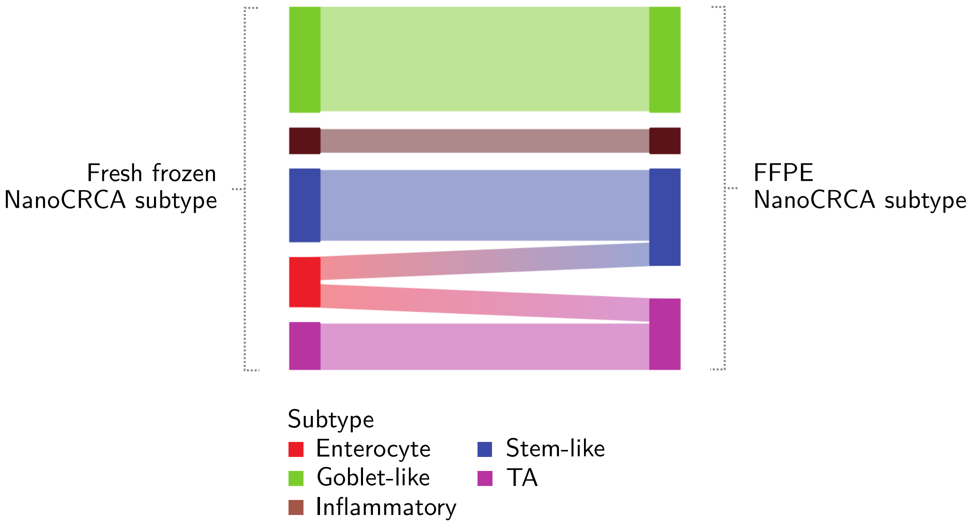


Figure 2.18: Subtypes are concordant between fresh frozen and FFPE tissues for high-cellularity samples Alluvial diagram showing the subtypes of matched fresh frozen and FFPE samples having high cellularity ($n = 14$, excluding undetermined subtype samples).

2.4 Chapter discussion and conclusions

As an analytical validation, these analyses alone cannot confirm whether there is prognostic or predictive value in the subtypes. However, efforts towards testing this power are ongoing at the Sadanandam Lab with the collection of samples from multiple international clinical trials.

The technical reproducibility of the assay was in line with that of previously reported nCounter cancer subtyping assays (Veldman-Jones *et al.*, 2015). However, the major limitation of this chapter was the analysis of the concordance between fresh-frozen and FFPE subtyping results, due to the limited number of samples having the appropriate level of cellularity. Collecting large matched fresh-frozen and FFPE tissue samples is a challenge, so confirming whether subtyping using these two tissue types consistently yields the same results could take some time. Previous work has indicated correlations between fresh frozen and FFPE RNA profiles profiled using nCounter can reach 90% (Kolbert *et al.*, 2013; Norton *et al.*, 2013; Reis *et al.*, 2011), hence subtype concordance can also be expected to be high.

In this work, it was not explored whether this assay is reproducible across different laboratories. All NanoCRCA assays were performed in the Sadanandam Lab at the Institute of Cancer Research, and hence, the level of inter-laboratory variance cannot be assessed. Previous studies have shown that high inter-laboratory reproducibility can be achieved using the nCounter platform (Nielsen *et al.*, 2014). In addition, the samples analysed in this chapter were from a variety of sources: the majority came from prospective collections for clinical trials, but one cohort was sourced from a commercial supplier of tumour RNA. For the purposes of the analytical development presented here, the retrospective nature of these analyses and the use of commercially supplied RNA is acceptable, but the future development of this assay must rely on the prospective collection of samples using a uniform collection protocol.

Subtype prediction by the NanoCRCA assay was highly concordant with more multiplexed platforms, and predicted subtypes show the expected association with the CMS subtypes, highlighting the similarities between these two classification systems. In addition, the same 38-gene signature can be applied to subtype both whole-transcriptome data (microarrays or RNA-seq) and nCounter data, allowing for more equivalent interpretation of results from both of these platform types.

If this assay can be proved to be prognostic or predictive in a way that is clinically meaningful, the question will remain as to how patients whose tumour subtype could not be predicted should be managed. These gene expression profiles could be unclassifiable for a number of reasons, including low RNA quality, normal tissue contamination, or intratumoural heterogeneity. The former can be overcome with more robust sample collection and preparation procedures; the latter will be further explored in Chapter 3.

In summary, this chapter showed the development and analytical validation of the NanoCRCA biomarker assay based on a refined 38-gene classifier, and the classification of CRC samples into molecular subtypes. Since multiple CRC clinical trials will require low-cost, reproducible and rapid clinically implementable assays to prospectively validate CRC subtypes for subtype-specific studies, the NanoCRCA assay may potentially facilitate this process in the clinic using FFPE samples.

Chapter 3

Comprehensive quantification of intratumoural subtype heterogeneity in CRC using *in vitro*-validated machine learning models in large, clinically-annotated datasets

3.1 Introduction

The intratumoural genomic heterogeneity of CRC is beginning to be illuminated, as detailed in Chapter 1.3. Early multi-regional analyses showed spatial diversity of *KRAS* and *TP53* mutations and the loss of heterozygosity at the *APC* and *DCC* loci (Losi *et al.*, 2005), as well as discordant mutation statuses between the central tumour and the invasive front of *BRAF* and *PIK3CA* (Baldus *et al.*, 2010). Indications that dominant clonal mutations arise early in cancer development and are unlikely to be overtaken by later aberrations, regardless of their fitness, have been found by more recent statistical models (Sottoriva *et al.*, 2015). However, it is as yet unclear how this genomic understanding potentially translates into clinical practice.

At the intertumoural scale, transcriptomic CRC subtypes exist which are associated with patients' responses to both cytotoxic and targeted therapeutic regimes (Schlicker *et al.*, 2012; De Sousa E Melo *et al.*, 2013; Sadanandam *et al.*, 2013; Roepman *et al.*, 2013). Previous work defining the CMS subtypes indicated that a significant proportion of tumours express the phenotype of multiple subtypes simultaneously (Guinney *et al.*, 2015). The hypothesis of this chapter is that this intratumoural transcriptomic heterogeneity is important in further understanding both the biology of

CRC and patients' responses to therapies, in addition to known key genomic features.

Quantification of intratumoural subtype heterogeneity is ideally performed via single-cell RNA-seq and subsequently classifying individual cells. However, this technique is not yet feasible at large (clinical) scales due to restrictive costs, and suffers from a lack of translatability to the clinic due to its inapplicability to FFPE-preserved tissues. Low-cost assessment of intratumoural subpopulations could instead be achieved in a more readily translational fashion by computational means.

This chapter introduces a machine learning approach to quantify subtype subpopulations within individual tumours. This method is based on a statistical framework as introduced by validated deconvolution tools to quantify immune cell type subpopulations (Newman *et al.*, 2015). I assessed the utility of intratumoural subtype subpopulations in increasing biological understanding of CRC, alongside their power as prognostic markers and their associations with drug responses, in multiple independent datasets.

3.2 Methods and data sources

3.2.1 Patient/pre-clinical cohorts and datasets

Seven publicly-available CRC cohorts were studied, as summarised in Table 3.1 and detailed below:

TCGA: Level 3 normalised ($\log_2(\text{TPM}+1)$) tumour RNA-seq data were downloaded from the UCSC Xena Browser (Goldman *et al.*, 2018) for 380 CRC patients.

Single-cell RNA-seq: FPKM values from single-cell RNA-seq performed on 375 tumour cells from 11 CRC patients (Li *et al.*, 2017a) were downloaded from GEO and converted to TPM. Deconvolution validation was performed on a "pseudo-bulk" dataset generated by pooling reads from all cells of a given sample (downloaded from the EGA accession EGAS00001001945), then processing these pooled files using a standard RNA-seq gene quantification pipeline (RSEM v1.2.29, Bowtie 2 v2.2.6, Samtools v1.3.1, Picard Tools 2.1.0, aligned to human transcriptome version GRCh37). One sample (CRC11) was excluded for containing no epithelial cells.

GSE14333: Affymetrix CEL files for 290 CRC specimens (Jorissen *et al.*, 2009) were downloaded from GEO and RMA normalised.

FOLFIRI: RMA normalised data for 21 mCRC patients treated with FOLFIRI chemotherapy (Del Rio *et al.*, 2007) were downloaded from GEO.

Khambata-Ford: Affymetrix CEL files for 80 patients with mCRC treated with cetuximab monotherapy (Khambata-Ford *et al.*, 2007) were downloaded from GEO and RMA normalised.

Cell line panel: Normalised Illumina Beadchip data for 155 CRC cell lines (Medico *et al.*, 2015) were downloaded from GEO. "AUC index" from Supplementary Table 1 of the original publication were converted to "Cetuximab effect" from Figure 2 of the same manuscript by the formula: Cetuximab effect $= 13,000 -$ AUC index.

Novartis PDX panel FPKM normalised RNA-seq data for 68 CRC PDX models (Gao *et al.*, 2015) were downloaded from Supplementary Table 1 of the original publication and converted to TPM values.

Table 3.1: Overview of patient cohorts for deconvolution analysis: sample numbers, platforms, and clinical characteristics.

Cohort	Number of patients	Platforms	Public data availability	Clinical characteristics
TCGA	380	RNA-seq	GDC Portal; UCSC Xena Browser	Mixed stage primary CRCs
Single-cell RNA-seq	375 cells from 11 patients	Single-cell RNA-seq	GSE81861	Stage II–IV primary CRCs
GSE14333	290	Microarray	GSE14333	Mixed stage primary CRCs
FOLFIRI	21	Microarray	GSE62080	Stage IV CRC metastases
Khambata-Ford	80	Microarray	GSE5851	Stage IV CRC metastases
Cell line panel	155	Microarray	GSE59857	NA
Novartis PDX panel	68	RNA-seq	Gao *et al.*(Gao *et al.*, 2015)	NA

3.2.1.1 Coculture flow cytometry and gene expression

Co-culture experiments, FACS and gene expression data were kindly performed/collected by Ms Chanthirika Ragulan of the Institute of Cancer Research, UK.

The two mycoplasma-negative and STR-validated cell lines HCT116 and LS174T were fluorescently labelled with mCherry (excitation 561nm, emission 614nm) and Venus (excitation 488nm, emission 513nm) respectively.

Cells were first sorted using BeckmanCoulter MoFlo Astrios and Summit software (v6) to remove unlabelled/dead cells, before being seeded at the reported starting ratios (100:0, 75:25, 50:50, 25:75 and 0:100). After culturing for at least their doubling time, the cells were sorted again to quantify their respective subpopulations.

3.2.2 Data preprocessing and normalisation

For all datasets, where necessary, mapping was performed between gene- and platform-specific gene/probe symbols and IDs using *biomaRt* v2.32 and human genome version GCRh37. Multiple platform features mapping to single gene identifiers were reduced by selection of the platform feature with the highest standard deviation across samples. In RNA-seq datasets, features having >20% zero values were removed (with the exception of the single-cell dataset, where high zero-inflation meant a higher threshold of 80% zeros was utilised to ensure enough genes for subtyping), and data was not quantile normalised (as recommended by the authors of CIBERSORT (Newman *et al.*, 2015)).

For deconvolution into the CRCA subtypes, data were median centred (in $\log_2$ space) prior to deconvolution and entered into the model in non-log-linear space using HGNC symbols as feature identifiers.

3.2.3 Deconvolution model

The deconvolution model adopted from CIBERSORT (Newman *et al.*, 2015) consists of support vector regression (SVR — specifically, ν-SVR). The principal of SVR is similar to simple linear regression in that it aims to calculate a coefficient that best predicts a dependent variable/outcome from an independent variable/predictor. In this specific context, the dependent variable is the gene expression of a given sample, and the independent variable is the gene expression of a subtype. The coefficient can be interpreted as the proportion of the sample's overall gene expression attributable to that subtype.

Simple linear regression aims to minimise the sum of the squared errors between the data and the regression line. SVR instead fits a regression line where each data point is a maximum distance of ϵ away (Figure 3.1), for which the magnitude of the coefficients is minimal (to prevent overfitting). As this criterion is not always attainable, some points can be allowed to fall a distance further than ϵ away from the regression line, but these will be penalised (through a term in the loss function that is minimised by the fitting procedure). ν-SVR introduces a new parameter ν to allow the user to pre-specify what proportion of data points are allowed to fall in- and outside the ϵ boundary.

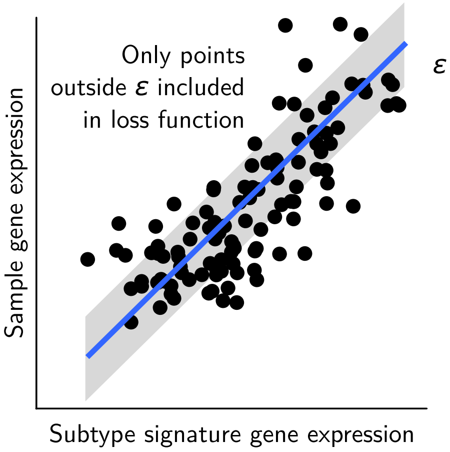


Figure 3.1: ν-SVR is more robust to noise than simple linear regression. Scatter plot illustrating the ϵ boundaries in SVR (data illustrative only).

As discussed by the authors of CIBERSORT (Newman *et al.*, 2015), SVR is more likely to be appropriate than simple linear regression in the biological setting due to its robustness to noise, overfitting (Cherkassky & Ma, 2004), and correlations between subtype signatures (multicollinearity) (Wang, Zhu & Zou, 2006).

In the CIBERSORT implementation adopted here (Newman *et al.*, 2015), ν values of 0.25, 0.5 and 0.75 are all tested, and a final value selected by the minimum RMSE. In addition, CIBERSORT sets any negative coefficients to zero, before normalising all the coefficients to sum to 1. It is these steps that allow the coefficients to be interpreted as subtype subpopulations.

The subtype signatures used for deconvolution in all datasets except the co-cultured cell lines described below are the 786-gene centroids that define the CRCAssigner subtypes (Sadanandam *et al.*, 2013), containing genes that are highly expressed within each subtype.

3.2.3.1 Deconvolution of co-cultured cell lines

For deconvolution of co-cultured cell lines into their constituent subpopulations, marker genes representative of each line had to be identified. From the cell line dataset GSE59857, ratios of gene expression for each gene in the custom CRC NanoString panel from Chapter 2 (Table B.1) were calculated between LS174T and HCT116 and scaled to the range [-1, 1]. Genes with scaled ratio > 0 or < -0.8 were selected as markers for the two cell lines respectively (due to the higher overall expression of these genes in LS174T than HCT116). The expression of these marker genes from the cell line panel GSE59857 were then used as signatures for deconvolution of the co-cultures.

3.2.4 Bulk subtype assignment

Bulk subtypes were assigned to samples using Pearson's correlation of gene-wise median-centred expression values to the CRCA centroids, as previously described (Sadanandam *et al.*, 2013).

3.2.5 Statistical methods

All statistical analysis was performed in R 3.4. Simpson's diversity index was calculated using the *vegan* v2.5 package. MCR for deconvolution was calculated as the number of samples for which the largest subpopulation was not the same as the bulk subtype. Differences in subtype subpopulations between patient groups were calculated with the Mann-Whitney U (2 groups) or Kruskal-Wallace rank sum (>2 groups) tests using the *stats* v3.4 package. Resulting *p*-values were corrected for multiple testing in the case where multiple subtypes were being analysed simultaneously using the Benjamin-Hochberg procedure (FDR) implemented in the *stats* package. Kendall's tau correlation coefficient was calculated to quantify correlations between subtype subpopulations and other variables as it is non-parametric and subtype subpopulations are bound between 0% and 100% (i.e. not normally distributed, an assumption of Pearson's correlation coefficient). Trends in anti-PD1 immunotherapy scatter plots were visualised using a generalised linear model with binomial outcome and cauchit link function (*stats* package).

3.2.5.1 Survival analysis

Kaplan-Meir curves were plotted using the *survival* v2.41 and *survminer* v0.42 packages, with *p*-values calculated using the *survConcordance* function. Patients were dichotomised separately for each subpopulation-based survival analysis using the *surv_cutpointf* function with default parameters to calculate an optimal cutoff for the relevant subtype to use as a threshold between high and low subtype subpopulations. This function returns the cutoff providing the maximal log-rank statistic. Cox proportional hazards regression was performed using default parameters in the *survival* package. Where Cox proportional hazards survival models were compared between bulk subtype classification and dichotomised subpopulations, the *anova* function from R package *stats* v.3.4 was input with the relevant nested models to generate p-values from the likelihood ratio test.

3.3 Results

3.3.1 *In vitro* validation of deconvolution models

3.3.1.1 Quantification of CRC cell lines in co-culture via FACS and deconvolution

I firstly aimed to test whether SVR-based computational deconvolution could be used to estimate cell type subpopulations where gene expression data was collected from cells kept under controlled laboratory conditions, and whose relative subpopulations were known by robust laboratory methods. Two fluorescently labelled CRC cell lines, HCT116 and LS174T, were seeded at different starting ratios as shown in Figure 3.2 (HCT116:LS174T starting ratios 100:0, 75:25, 50:50, 25:75 and 0:100) and co-cultured (experimental conditions detailed in Chapter 3.2.3.1; experiments performed by Ms Chanthirika Ragulan at the Institute of Cancer Research). After 4 days, to allow time for intermixing of the cells, FACS was performed on each co-culture to empirically determine the subpopulation of each cell line. This provided a “ground truth” estimate for the proportion of cells belonging to each cell line, using standard laboratory techniques. RNA was extracted immediately afterwards for gene expression profiling, providing the equivalent to a “bulk” gene expression profile of a tumour. I then applied computational deconvolution to these profiles to estimate the cell line subpopulations *in silico* (details of methodology can be found in Chapter 3.2).

When I applied computational deconvolution, it was able to estimate these proportions with a root mean square (RMS) error of 14.1% (Figure 3.2b). While there is no generally-applicable threshold for an acceptable RMS error, in this context being able to computationally estimate cell type populations to within, on average, 14.1% of the empirical estimate justifies applying this method where no ground truth estimates are available.

Deconvolution detected both the dominant and the minority subpopulations in two of the three mixed conditions (50:50 and 25:75). In the third mixed condition (75:25) it did not detect the LS174T subpopulation, a false negative result. At the time of FACS sorting and RNA extraction, this subpopulation comprised 16% of the cells. Conversely, a negligibly small false positive subpopulation (<1%) of HCT116 cells in the 100% LS174T condition was reported. This result gives an initial indication that computational deconvolution can be utilised to estimate transcriptomic phenotype subpopulations.

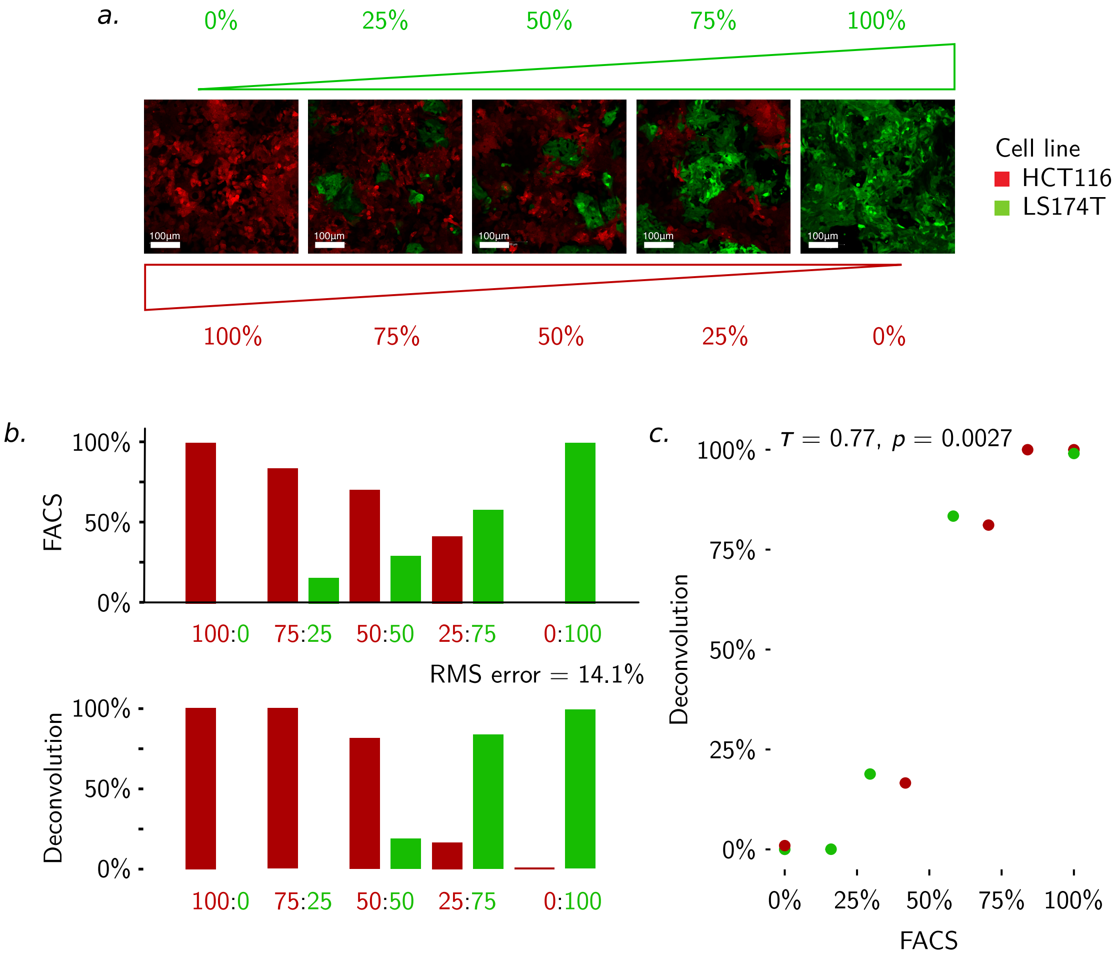


Figure 3.2: Computational deconvolution can estimate the subpopulations of two different cell lines co-cultured at different cell count ratios. *(a)* Fluorescence microscopy images of co-cultures of two CRC cell lines labelled with GFP (LS174T) and mCherry (HCT116). Percentages indicate the starting seeding ratio of the two cell lines. *(b)* Bar plot and *(c)* scatter plot showing subpopulations of co-cultured cell lines as quantified using FACS and computational deconvolution , demonstrating that computational deconvolution can reconstruct the subpopulation ratios of the two cell lines. Microscopy images, FACS and gene expression data were collected by Ms. Chanthirika Ragulan of the Institute of Cancer Research, UK.

3.3.1.2 Quantification of CRC subtypes in tumours via single-cell RNA-seq and deconvolution

In order to validate the computational deconvolution approach against benchmark subpopulation data in the context of CRC subtypes, I applied it to 363 cells from 10 CRC samples previously profiled by scRNA-seq (Li *et al.*, 2017a). The vast majority of cells were epithelial (77%), but immune (18%), fibroblast (4%) and endothelial

(1%) cells were also present., Pseudo-bulk RNA-seq data (generated by pooling reads from single cells, see Chapter 3.2) was used to deconvolve the subtype subpopulations in each sample of five CRCA subtypes; enterocyte, goblet-like, inflammatory, TA and stem-like. The cells from the scRNA-seq data for each sample were then classified into the same five subtypes, to estimate the ground-truth subpopulation of each subtype. The correlation of the single cells with the CRCAssigner centroids (mean 0.18; SD 0.1) was lower than is seen in bulk datasets(Sadanandam *et al.*, 2013), likely due to the higher number of drop-outs in single-cell RNA-seq data (Kiselev, Andrews & Hemberg, 2019).

Figure 3.3 shows the proportion of cells classified into each subtype for each sample by scRNA-seq, and the estimation of the same by computational deconvolution. There were no false positive subtype subpopulations (i.e. subpopulations that were predicted to exist by computational deconvolution but for which there was no evidence in the scRNA-seq data). Where deconvolution failed to report an existing subtype subpopulation, the stem-like subtype was the most likely to be missed. The majority of these false negatives fell within the range of 1-17%. However, in the sample CRC10, a subpopulation of the enterocyte subtype comprising 31% of the sample was not detected. Nevertheless, the overall RMS error was 19% between subpopulation estimates by scRNA-seq and bulk deconvolution, and indicates fair agreement between these two results.

Analysis of the correlation between scRNA-seq and computational deconvolution-based results showed that the inflammatory, TA and stem-like subtypes were all significantly correlated between the two methods ($\tau > 0.8$; $p < 0.01$). The goblet-like and enterocyte subtypes did not correlate significantly ($\tau < 0.4$; $p > 0.05$). This could be due to the difficulty in distinguishing enterocyte tumour subtype and normal colon tissue (due to the high expression of enterocyte subtype genes in normal colon epithelium, as discussed in Chapter 2.3.5), which could have a knock-on effect on the estimation of the goblet-like subtype due to the similarity between these two subtypes – the most differentiated of the CRCAssigner subtypes.

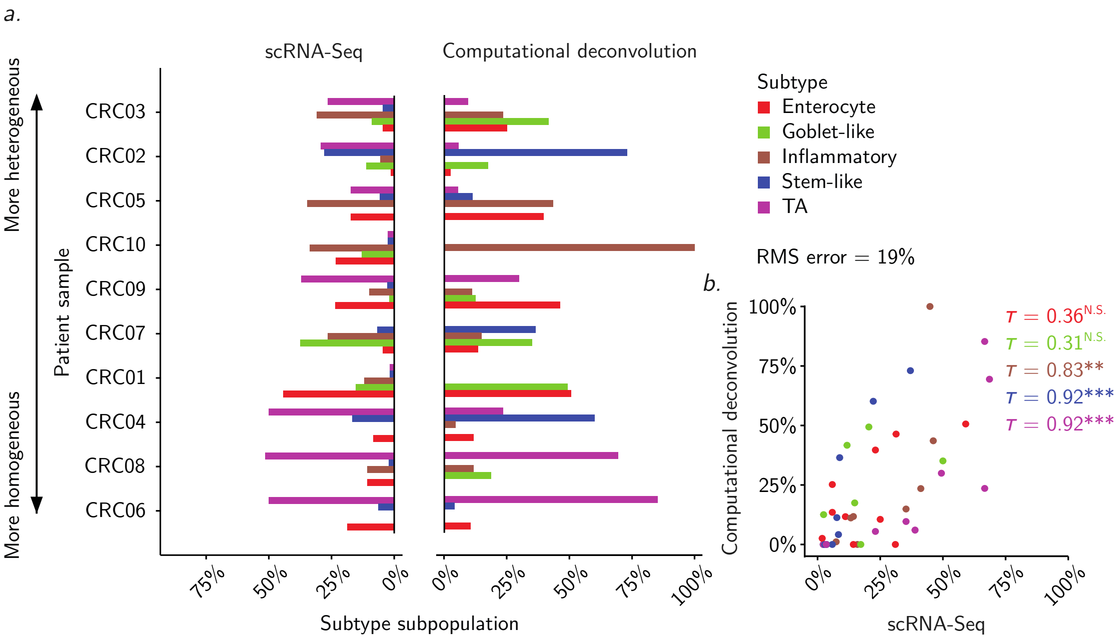


Figure 3.3: Computational deconvolution can estimate the cellular subpopulations present in patient tumours. *(a)* Bar plot and *(b)* scatter plot showing intratumoural subtype subpopulations as quantified using scRNA-seq ($n = 363$ cells) and computational deconvolution ($n = 10$ tumours). Tumours in *(a)* are ordered by their level of heterogeneity, calculated using Simpson's diversity index(Schleuter *et al.*, 2010). RMS: root mean square. N.S.: not significant; ** $p < 0.01$; *** $p < 0.001$.

3.3.2 Prognostic power and potential drug response discrimination using intratumoural subtype populations versus bulk subtypes

To explore intratumoural subtype heterogeneity in a larger cohort of samples, I carried out deconvolution of 380 CRC samples from TCGA (see Chapter 3.2.1). Figure 3.4 shows the results of the deconvolution, giving the proportion of each subtype present in each sample, as well as the "bulk subtype" assigned by the CRCAssigner classifier (see Chapter 3.2). I computed 94% of these samples to be composed of a mixture where the largest subtype component was the same as the bulk CRCAssigner subtype. This aligned with the expectation that the subtype with the highest subpopulation would dominate the gene expression profile of a sample, and hence cause its classification into that subtype when treated as a bulk sample.

I found that there was a near-uniform distribution of subtypes in these samples: the TA subtype was slightly dominant, being detected in 66% of samples, while stem-like, inflammatory, enterocyte and goblet-like could be found in 65%, 65%, 63% and

61% of samples respectively. TA had the highest mean subpopulation at 25%, while inflammatory had the lowest at 16% (Figure 3.4). By comparison, the bulk subtype assignments categorised 25% of samples as stem-like, 25% as TA, 20% as enterocyte, 15% as inflammatory and 14% as goblet-like.

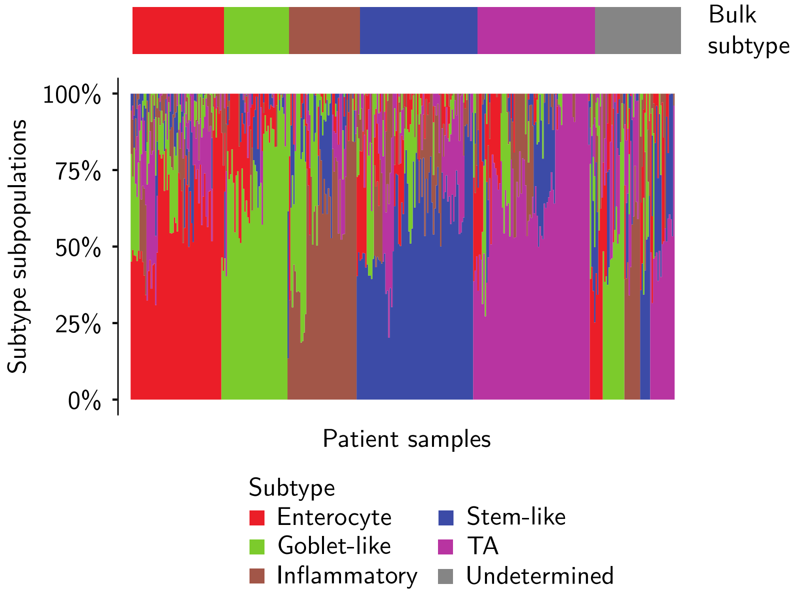


Figure 3.4: Patient tumours are transcriptomically heterogeneous, with the subtype as determined by bulk subtyping representing the largest transcriptomic subpopulation in the tumour. Bar plot showing the subtype subpopulations present in each sample of the TCGA dataset (as calculated by computational deconvolution), alongside the bulk subtype of each sample (n = 378). Each vertical stack of bars represents the different subpopulations present in a single sample.

3.3.2.1 Stem-like subtype

3.3.2.1.1 High stem-like subpopulations are associated with advanced stage and poor prognosis in patients. When I investigated how the subpopulations varied with stage, I found that the stem-like subpopulation increased significantly over the tumour stages ($p < 0.005$; $\epsilon^2 = 0.0595$; Figure 3.5a). This aligned with the increasing proportion of patients whose bulk tumour was classified as being stem-like at later stages (Figure 3.5b).

I stratified stage I-III patients into high- and low-stem-like groups based on their stem-like subpopulations, using a subpopulation cutoff that maximises the prognostic difference between groups (Chapter 3.2.5.1). There was significantly poorer RFS in the high-stem-like group (p = 0.037, HR = 1.86 (1.03-3.36); Figure 3.5b) — this effect was not seen when stratifying patients based on their subpopulations of other subtypes. However, stratifying the same patients into stem-like versus other subtypes did not give a significant difference in survival (p = 0.89, HR = 1.05 (0.52-2.13); Figure 3.5d). Including high-stem-like status in a Cox proportional hazards model of survival significantly improved its ability to predict OS time relative to a model that only included bulk stem-like status, as measured by a likelihood ratio test (which tests whether high-stem-like status has a significant effect on survival, given bulk stem-like status; $p = 0.015$). In contrast, bulk stem-like status did not add prognostic information to a Cox model with high-stem-like status ($p = 0.19$). However, neither high-stem-like or bulk stem-like status added significant prognostic information to stage ($p = 0.13$ and $p = 0.80$, respectively).

I validated this finding of reduced RFS in early-stage, highly stem-like patients in an independent cohort of 290 CRCs (GSE14333, see Chapter 3.2.1), again showing significantly poorer prognosis in patients with a large stem-like component in their tumours ($p < 0.0001$, HR = 6.98 (2.81-17.35); Figure 3.5e).

Previous analyses have shown poor prognosis in the stem-like subtype in treatment naive patient samples (Sadanandam *et al.*, 2013), but this result was not stratified by stage, and as such it is possible that stem-like patients' poor prognosis was due to their advanced stage. The results presented here indicate that while bulk stem-like subtype status is not prognostic in early-stage CRCs, the stem-like subpopulation within each sample is prognostic in these patients.

Given that the poor prognoses of patients with high stem-like subpopulations seemed to be the result of their advanced tumour stage, I wanted to investigate whether stem-like subpopulations could have value as predictors of therapeutic responses in metastatic cancers. One example is given in the next section.

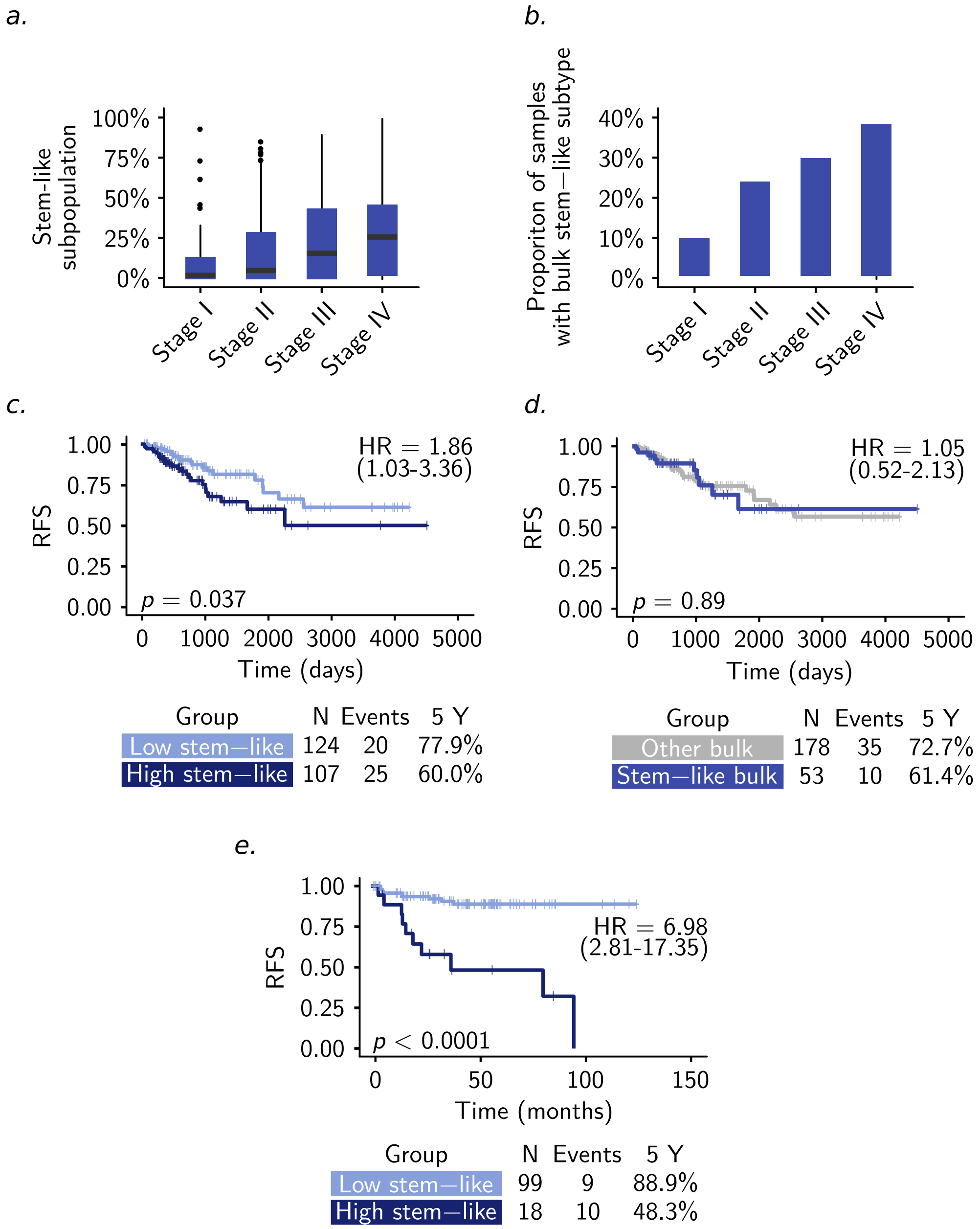


Figure 3.5: Dichotomising early-stage patients by stem-like subpopulations provides greater discrimination in outcomes than dichotomising by bulk subtype, but does not add significant value over stage information. *(a)* Box plot showing the subpopulations of the stem-like subtype in patients at different stages. *(b)* Bar plot showing the proportion of patients whose bulk tumour falls into the stem-like subtype at different stages. *(c,d)* Kaplan-Meier survival curves for patients with *(c)* high- or low-stem-like tumours (as determined by computational deconvolution; cutoff 9.2% stem-like) and *(d)* bulk stem-like or other bulk subtype tumours in the TCGA cohort. *(e)* Kaplan-Meier survival curves for patients with high- or low-stem-like tumours in the GSE14333 cohort. As samples with undetermined bulk subtype could not be included in KM curves for bulk stratification, they were also excluded from the computational deconvolution stratification curves.

3.3.2.1.2 Potential associations with FOLFIRI therapy outcomes in highly stem-like patients. Folinic acid, fluorouracil and irinotecan (FOLFIRI) is a chemotherapy regimen that can be used in the treatment of metastatic CRC (Chapter 1.1.3). Increased sensitivity of the stem-like subtype to FOLFIRI has previously been demonstrated (Sadanandam *et al.*, 2015). Using the same dataset as that publication (Del Rio *et al.*, 2007) (which includes stage IV patients only; see Chapter 3.2.1), deconvolution reveals a trend towards increased tumour shrinkage in patients with a higher proportion of the stem-like subtype (Figure 3.6a). FOLFIRI responders (as defined by the WHO criteria (Miller *et al.*, 1981)) had significantly larger stem-like subpopulations than non-responders ($p = 0.03$; Figure 3.6b).

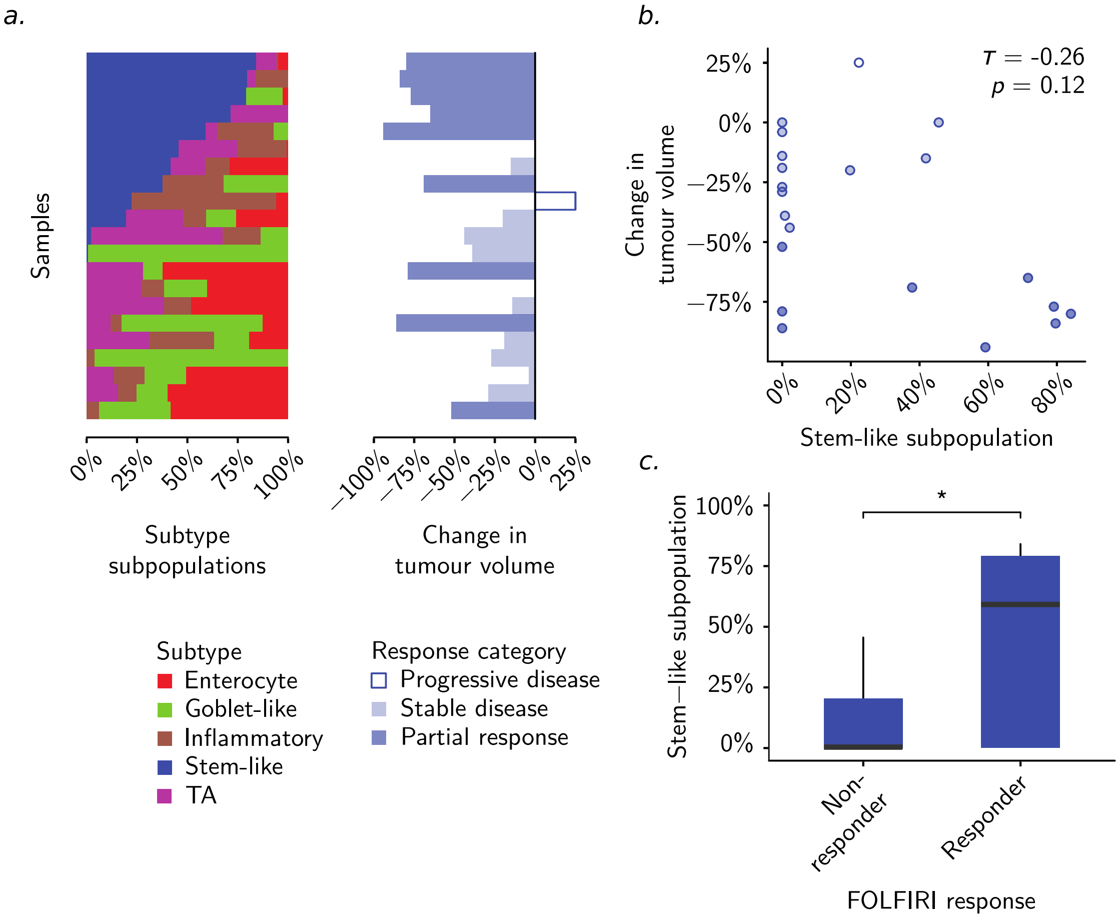


Figure 3.6: Patients with higher responses to FOLFIRI had higher subpopulations of the stem-like subtype. *(a)* Bar plots showing the subpopulation of each subtype in each sample (left panel) against the response (change in tumour volume) of each patient to FOLFIRI chemotherapy (right panel) from GSE62080 ($n = 21$). Fills of each bar indicate the overall response category. *(b)* Scatter plot showing the stem-like subpopulation versus the change in tumour volume of each patient. Fills of each point indicate the overall response category. *(c)* Boxplot showing the stem-like subpopulation in patient responders and non-responders to FOLFIRI therapy.

I then dichotomised patients using their stem-like subpopulations into high-stem-like and low-stem-like groups (using a cutoff of 59% stem-like subpopulation, determined by maximising the sum of the negative and positive predictive values). While stratifying patients by stem-like subpopulation did not identify FOLFIRI responders with more sensitivity than bulk subtype status, it did do so with more specificity (Table 3.2). Given the increased side effects and cost per year of life of FOLFIRI versus the similarly-effective oxaliplatin-based regimen FOLFOX (Neugut *et al.*, 2019), predicting FOLFIRI responders using stem-like subpopulations could save patients from unnecessary toxicity and decrease the costs of care when compared to stratifying patients using bulk subtypes.

Table 3.2: Stratifying by stem-like subpopulations increases the specificity of FOLFIRI response prediction. Confusion matrix showing the sensitivity and specificity of bulk- and subpopulation-based stratifications of patients in predicting response to FOLFIRI.

	Non-responders	Responders	Sensitivity/specificity
Other bulk	8	4	0.56/0.80
Stem-like bulk	2	5	
Low-stem-like	12	4	0.56/1.00
High-stem-like	0	5	

Taking into account the poor prognosis of patients with advanced stage/highly stem-like tumours (Chapter 3.3.2.1.1), a personalised therapy that could improve their outcomes would be highly valuable. The association of tumour responses to FOLFIRI and stem-like tumour subpopulations should be further explored to understand if this is a reproducible effect.

3.3.2.2 TA subtype

3.3.2.2.1 TA subpopulations and responses to cetuximab in patients. It has previously been established that a subset of patients who fall into the TA subtype respond to the anti-EGFR monoclonal antibody cetuximab (cetuximab-sensitive (CS)-TA; cf. cetuximab-resistant (CR)-TA) (Sadanandam *et al.*, 2015). When I quantified TA intratumoural subtype subpopulations in 80 CRC metastases treated with cetuximab monotherapy (see Chapter 3.2.1), I found significantly higher TA subpopulations in the liver metastases of *KRAS*-WT patients who responded to the therapy, versus those who did not respond ($p = 0.04$; $n = 32$; Figure 3.7).

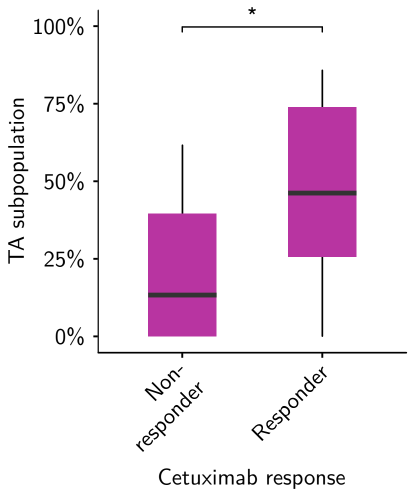
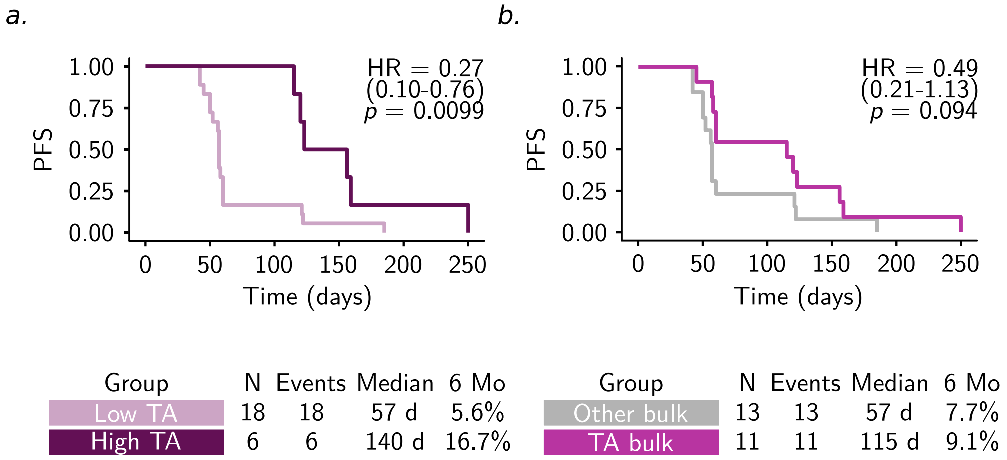


Figure 3.7: Responders to cetuximab have higher subpopulations of the TA subtype. Box plots showing the intratumoural subpopulations of the TA subtype in *KRAS*-WT cetuximab responder and non-responder liver metastases in the Khambata-Ford dataset ($n = 32$). * $p < 0.5$.

When I grouped these patients into high- and low-TA groups (using a cutoff of 62% TA subpopulation, which maximises the prognostic difference between groups, see Chapter 3.2.5.1), the p-value and hazard ratio of log-rank survival analysis decreased when compared to grouping of TA and non-TA patients (p = 0.0099, HR = 0.27 (0.10-0.76) vs. p = 0.094, HR = 0.49 (0.21-1.13); Figure 3.8). A Cox proportional hazards model that added high-TA status to bulk TA status was significantly better able to predict PFS over one that included bulk TA status alone (p = 0.029), whereas bulk TA status did not add significant prognostic information to high-TA status (p = 0.605).



Group	N	Events	Median	6 Mo
Low TA	18	18	57 d	5.6%
High TA	6	6	140 d	16.7%

Group	N	Events	Median	6 Mo
Other bulk	13	13	57 d	7.7%
TA bulk	11	11	115 d	9.1%

Figure 3.8: Dichotomising patients by TA subpopulations provides greater discrimination in predicting length of response to cetuximab than dichotomising by bulk subtype. Kaplan-Meier survival curves for patients with *(a)* high- and low-TA tumours (cutoff 62% TA) and *(b)* bulk TA or other bulk subtype tumours.

I then hypothesised that the CS-TA group originally identified in the original CRCA publication (Sadanandam *et al.*, 2013) could be comprised of patients who had a particularly high subpopulation of TA in their metastases, whereas the CR-TA group had a high subpopulation of other subtypes despite also being classified as a bulk TA tumour. Indeed, there was a significant difference in TA subpopulation between *KRAS*-WT responders versus non-responder liver metastases with bulk TA subtype ($n = 11$; $p = 0.04$; Figure 3.9). Unfortunately, the very small sample size used to test this hypothesis (7 responders and 4 non-responders) means that it requires extensive further exploration.

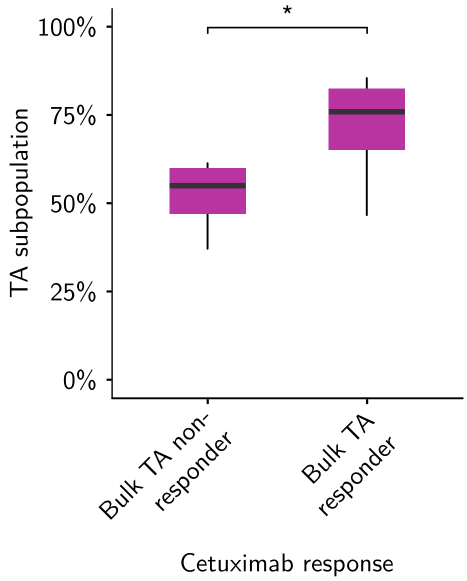


Figure 3.9: In tumours with bulk TA subtype, the TA subpopulation is higher in responders to cetuximab. Box plot showing the TA subpopulation in responders and non-responders to cetuximab having bulk TA subtype ($n = 11$). * $p < 0.5$.

3.3.2.2.2 TA subpopulations and responses to cetuximab in pre-clinical models. When I quantified subtype subpopulations in 152 CRC cell lines (Medico *et al.*, 2015) (see Chapter 3.2.1), the results showed non-trivial intra-cell-line heterogeneity (Figure 3.10), such as has previously been reported in single-cell transcriptomic analysis of lung ademocarcinoma (Suzuki *et al.*, 2015) and multiple myeloma (Mitra *et al.*, 2016). Two possible explanations for this heterogeneity are: i) that is a true and sustained reflection of the heterogeneity of the cells used to seed the culture; or ii) that it is the result of non-uniform phenotypic drift between the cells, perhaps owing to slight differences in environmental conditions between different physical locations in the culture. Unfortunately, it is not possible to test these hypotheses in this dataset due to a lack of longitudinal data.

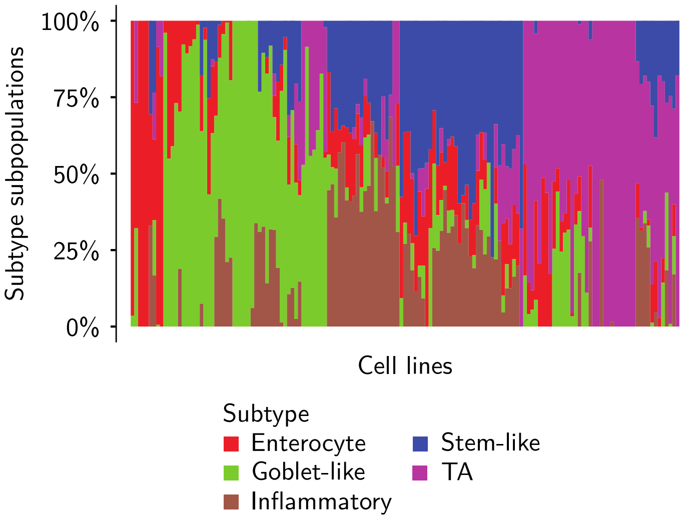


Figure 3.10: CRC cell lines exhibit subtype-based transcriptomic heterogeneity. Bar plot showing the proportion of subtype subpopulations in cell lines, as determined by computational deconvolution (n = 152). Each vertical stack of bars represents the different subpopulations present in a single cell line.

Similarly to patients' samples, there was a high subpopulation of TA in *KRAS*-WT cetuximab-sensitive cell lines ($p < 0.001$; Figure 3.11). Resistant lines were enriched for subpopulations of the inflammatory and stem-like subtypes ($p < 0.01$; Figure 3.11). There was no significant difference in the goblet-like or enterocyte subpopulations between resistant or sensitive cell lines ($p > 0.05$).

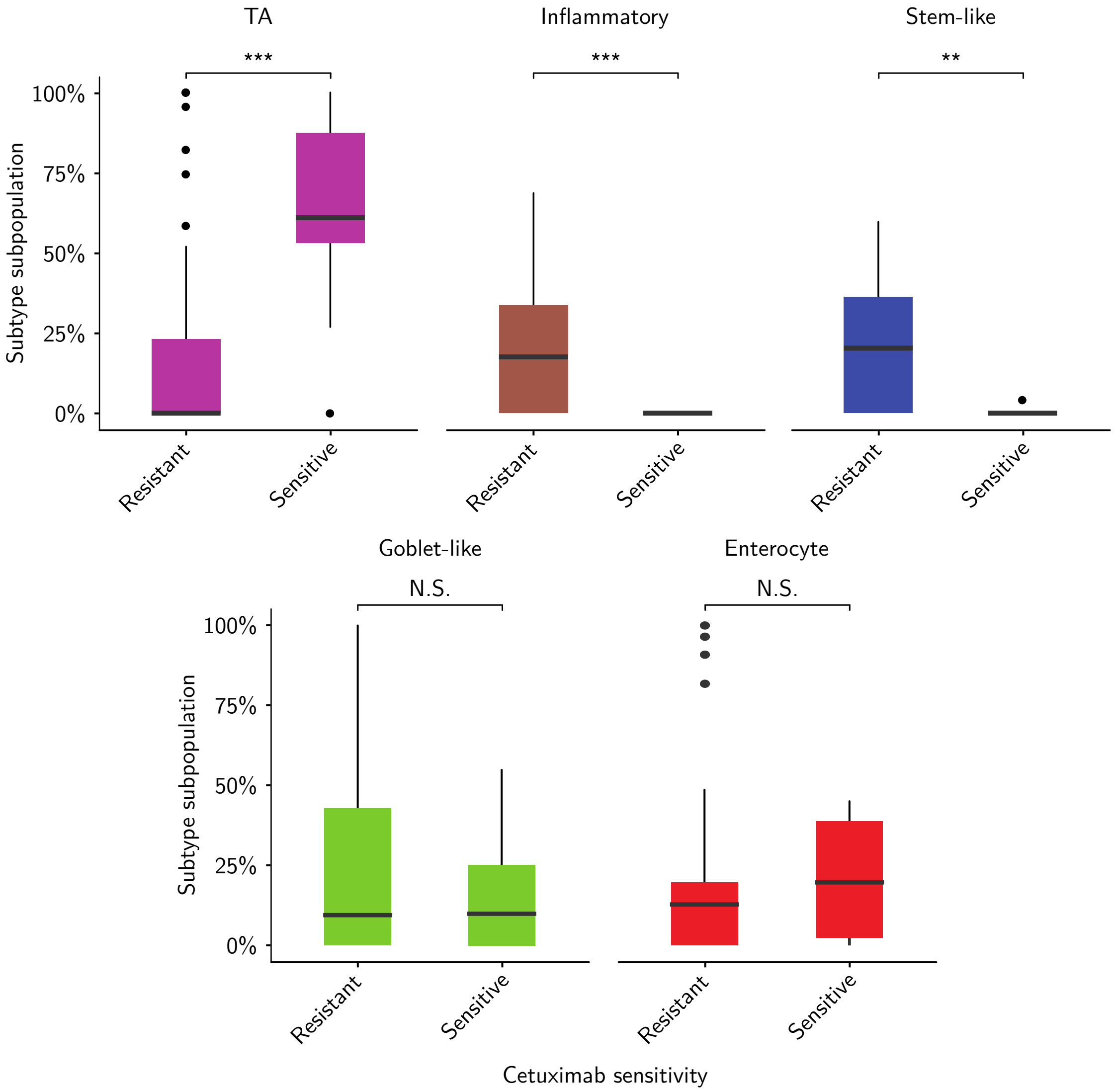


Figure 3.11: Cell line subtype subpopulations recapitulate patterns of patients' sensitivity to cetuximab Boxplots showing the intra-cell line subpopulations of the CRCAssigner subtypes in *RAS*-WT cetuximab-resistant and -sensitive cell lines ($n = 76$). N.S.: not significant; ** $p < 0.01$; *** $p < 0.001$.

Additionally, for further validation, I performed the same subtype quantification on 41 CRC patient gene expression profiles (Gao *et al.*, 2015) that had mutation information, and cetuximab responses from matched PDX models (see Chapter 3.2.1). This analysis revealed a significant and strong negative correlation between patient TA subpopulation and PDX change in tumour volume in *RAS*/*BRAF* WT models (Figure 3.12a). Examining the other subtypes present in these tumours revealed trends towards higher enterocyte and goblet-like subpopulations in WT patients whose PDX models did not respond to cetuximab (Figure 3.12b).

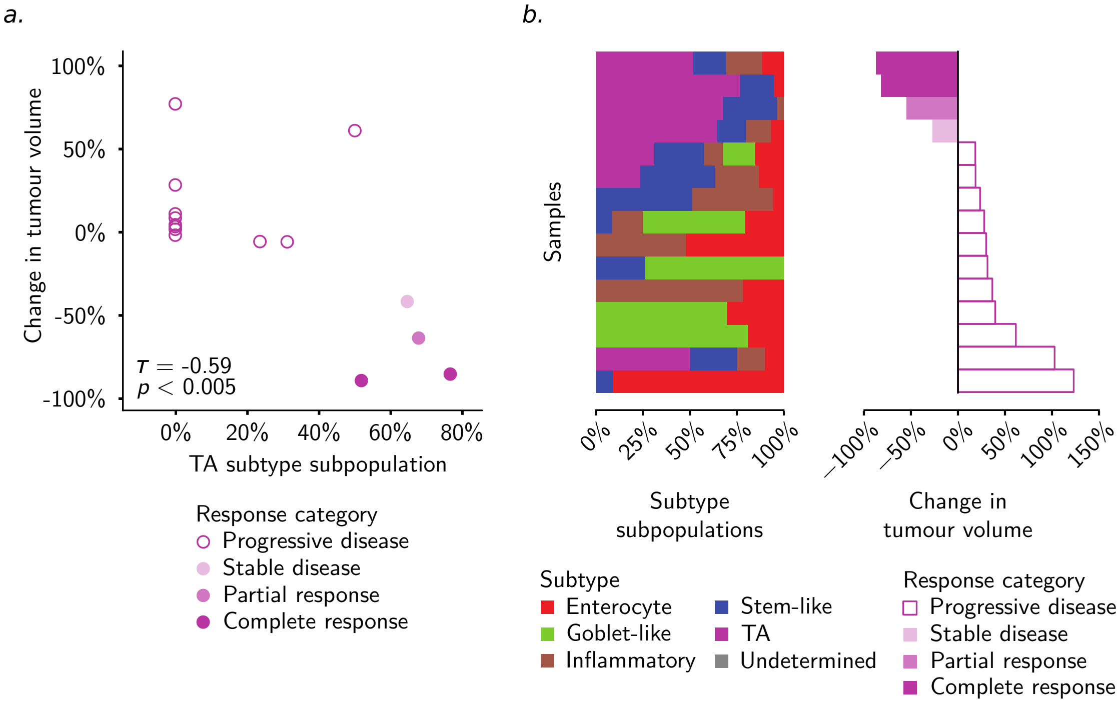


Figure 3.12: PDX models show higher responses to cetuximab when their TA subpopulation is high. *(a)* Scatter plot showing the TA subpopulation versus extent of cetuximab response in PDX models (n = 15). Fills of each point indicate the overall response category. *(b)* Bar plots showing the subpopulation of each subtype in each patient sample (left panel) against the response (change in tumour volume) of each matched PDX to cetuximab treatment (right panel) from the Novartis PDX data (n = 15). Fills of each bar indicate the overall response category.

This analysis of TA subpopulations across patients, cell lines and PDX models confirms that there is a relationship between the TA subtype and cetuximab response, as first proposed in the work defining the CRCA subtypes (Sadanandam *et al.*, 2013). Importantly, it demonstrates that this relationship is a function of intratumoural subtype heterogeneity. In the future, sensitive tumours may be identified prior to treatment by quantification of the TA subpopulation, likely with greater predictive power than could be achieved using patients' bulk subtype.

3.4 Intratumoural subtype heterogeneity within microsatellite (in-)stable tumours

Microsatellite instable tumours have recently been of interest as potentially targetable with personalised therapies, as they have been shown to be more susceptible to immunotherapies, in particular PD1 blockade (Overman *et al.*, 2018, 2017; Kim *et al.*, 2017; Le *et al.*, 2015, 2017).

3.4.0.0.1 MSI-H tumours are enriched for inflammatory and goblet-like subpopulations. Returning to the TCGA patient dataset (Chapters 3.3.2.1.1 and 3.2.1), I investigated the nature of intratumoural subtype heterogeneity within the MSI-H and MSI-L/MSS groups.

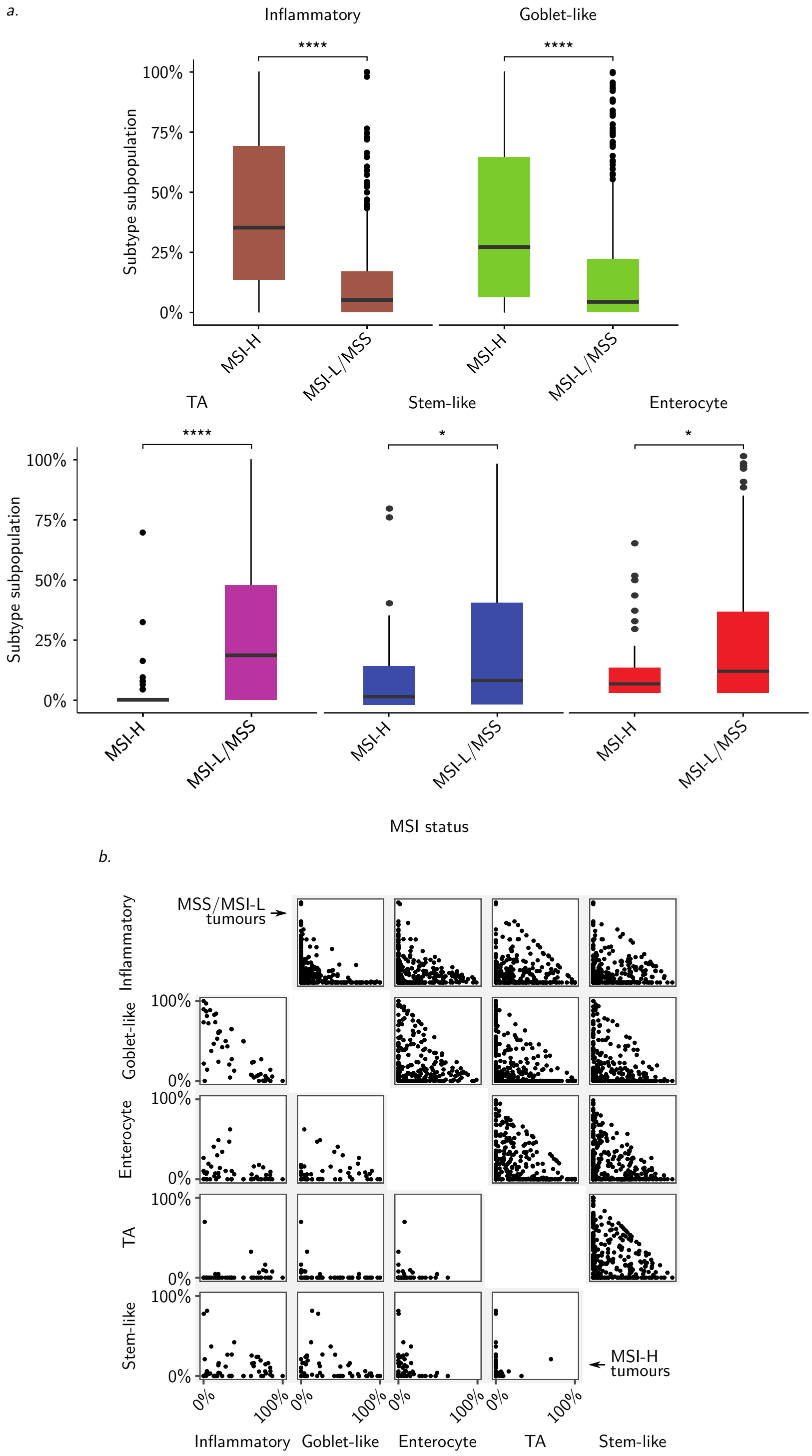
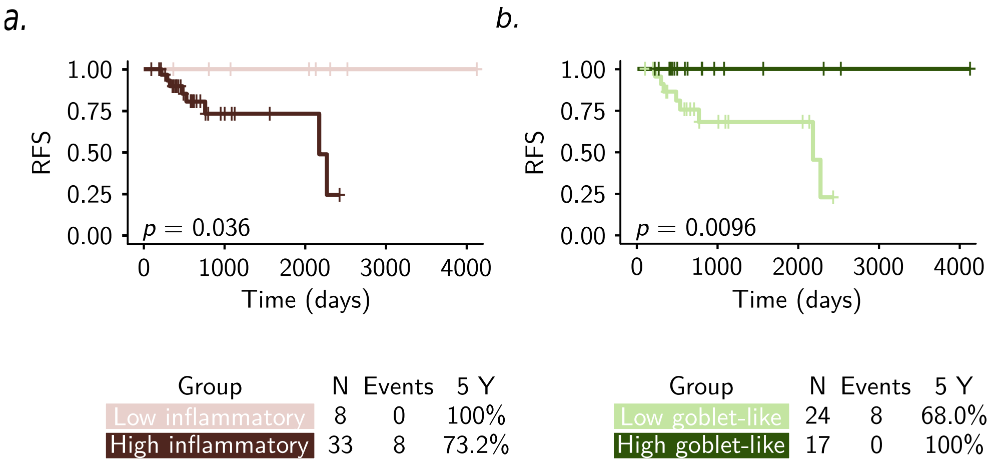


Figure 3.13: MSI-H and MSI-L/MSS patients have different distributions of subtype subpopulations, with MSI-H tumours comprised of inflammatory and goblet-like subpopulations. *(a)* Boxplots and *(b)* pairwise scatter plots showing the subtype subpopulations in MSI-H and MSI-L/MSS patient tumours in the TCGA data ($n = 368$). * $p < 0.05$; **** $p < 0.0001$

I determined that the subpopulations of all subtypes varied significantly between these two groups ($p < 0.05$, Figure 3.13a). The goblet-like and inflammatory subtypes were enriched in MSI-H tumours, whereas stem-like, enterocyte and TA subpopulations were higher in MSI-L/MSS tumours. Together, goblet-like and inflammatory subtypes made up the majority of MSI-H tumours' subpopulations, to the exclusion of other subtypes (Figure 3.13b).

3.4.0.0.2 Heterogeneous survival outcomes within MSI-H cancers by inflammatory and goblet-like subpopulations. Having shown that high levels of the stem-like subtype within tumours are indicative of poor prognosis and advanced stage (Chapter 3.3.2.1.1), I explored whether other subtype subpopulations could also be prognostic within the MSI-H group of patients. The results showed that the MSI-H group could be stratified into good- and poor-prognosis groups on the basis of intratumoural subpopulations of both the inflammatory and goblet-like subtypes ($p = 0.036$ and $p = 0.0096$ respectively; hazard ratios are incalculable for these models due to the lack of events in the low-inflammatory and high-goblet-like groups; Figure 3.14) – both of which are enriched in the MSI-H group. Interestingly, this analysis suggests that high levels of the inflammatory and low levels of the goblet-like subtype indicate poor prognosis in the MSI-H subset of patients, and vice-versa for good prognosis.

Furthermore, high-inflammatory and high-goblet-like status predicted RFS significantly better than bulk inflammatory/goblet-like alone ($p < 0.05$); the bulk subtypes did not add prognostic information to high-inflammatory/goblet-like status ($p > 0.5$) in a Cox proportional hazards model.



Group	N	Events	5 Y
Low inflammatory	8	0	100%
High inflammatory	33	8	73.2%

Group	N	Events	5 Y
Low goblet-like	24	8	68.0%
High goblet-like	17	0	100%

Figure 3.14: High inflammatory subpopulation portends a poor prognosis in MSI-H cancers, while high goblet-like subpopulation shows the opposite trend. Kaplan-Meier survival curves for MSI-H patients with high- and low- *(a)* inflammatory (cutoff 4.4% inflammatory) *(b)* goblet-like (cutoff 46% goblet-like) tumour subpopulations ($n = 41$). Hazard ratios are incalculable for these models due to the lack of events in the low-inflammatory and high-goblet-like groups.

The Kaplan-Meir curves in Figure 3.14 appear to indicate that the same subset of patients have high-inflammatory and low-goblet-like tumours (and a poor prognosis), however, the patient dichotomisations did not delineate patients in exactly the same ways. Table 3.3 shows that not all the high-inflammatory tumours were also low-goblet-like, and not all the low-inflammatory tumours were also high-goblet-like.

Table 3.3: The poor-prognosis high-inflammatory and low-goblet-like groups do not identify identical sets of patients. Cross-table showing the number of patients falling into the low/high-inflammatory and low/high-goblet-like groups.

	Low-inflammatory	High-inflammatory	*Total*
Low-goblet-like	3	21	*24*
High-goblet-like	5	12	*17*
Total	*8*	*33*	*41*

Importantly, the poor-prognosis high-inflammatory and low-goblet-like groups were not simply the tumours with the highest stage, as is shown in Table 3.4, where stage II tumours make up a larger proportion of the poor-prognosis groups than the favourable-prognosis groups.

Table 3.4: The poor-prognosis groups are not disproportionately enriched for later-stage tumours. Cross-table showing the number of patients from each subpopulation-based group falling into each tumour stage. * One patient had unknown stage.

	Stage I	Stage II	Stage III	Stage IV	*Total**
Low-inflammatory	2 (25%)	3 (38%)	2 (25%)	1 (13%)	*8*
High-inflammatory	7 (22%)	19 (59%)	4 (13%)	2 (6%)	*32*
Low-goblet-like	3 (13%)	13 (56%)	4 (17%)	3 (13%)	*23*
High-goblet-like	6 (35%)	9 (53%)	2 (12%)	0 (0%)	*17*
*Total**	*9*	*22*	*6*	*3*	*40*

The question of how to manage these subsets of patients having highly distinct prognoses within the MSI-H group must be considered in light of recent efforts to introduce immunotherapeutic options to these patients (Overman *et al.*, 2018, 2017; Kim *et al.*, 2017; Le *et al.*, 2015, 2017).

3.4.0.0.3 Correlation with predictor for anti-PD1 immunotherapy. Next, I investigated the implications that inflammatory subtype levels could have for patients' potential responses to immunotherapy-based treatment regimes. Patients' scores were calculated for a published signature of response to anti-PD1 immunotherapy (Ayers *et al.*, 2017). There was a significant positive correlation between the inflammatory subtype subpopulations and the INF-γ-related signature score ($\tau = 0.54$; $p < 0.005$). No significant correlation was found when calculated using the goblet-like subpopulation (Figure 3.15).

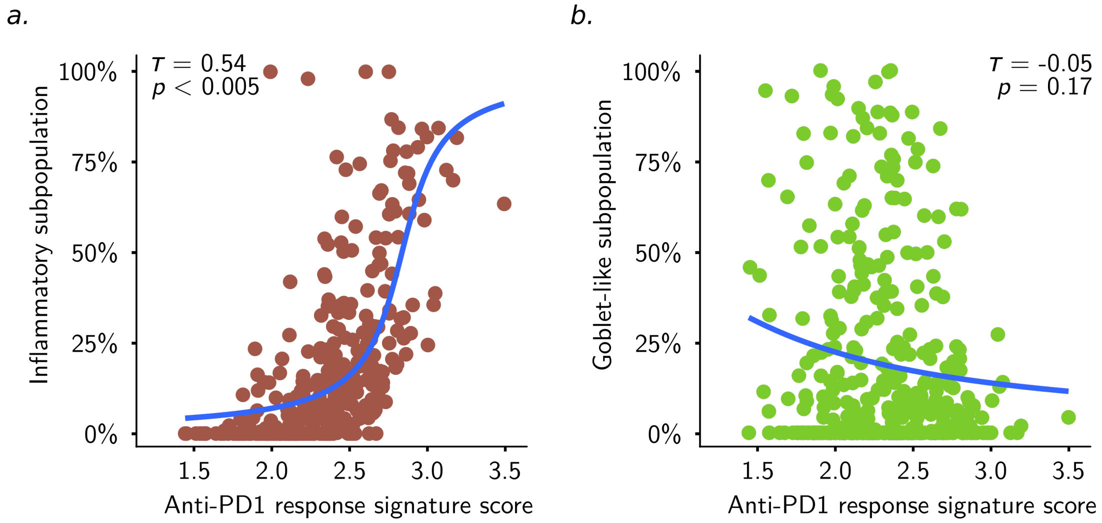


Figure 3.15: Only the inflammatory subtype correlated with anti-PD1 response signature scores when MSI status was not considered. Scatter plots showing anti-PD1 signature scores versus subtype subpopulations for the *(a)* inflammatory and *(b)* goblet-like tumour subpopulations ($n = 370$). Correlations values were calculated using Kendall's rank method, and the trend line was fitted using a binomial GLM.

However, when I repeated this analysis using only the MSI-H patients, the result for the inflammatory subtype stayed the same, whilst the goblet-like subpopulation had a significant negative correlation with the response signature (Figure 3.16). This suggests that even within this more homogeneous group of patients, intratumoural subtype heterogeneity could possibly cause large discrepancies in patient outcome.

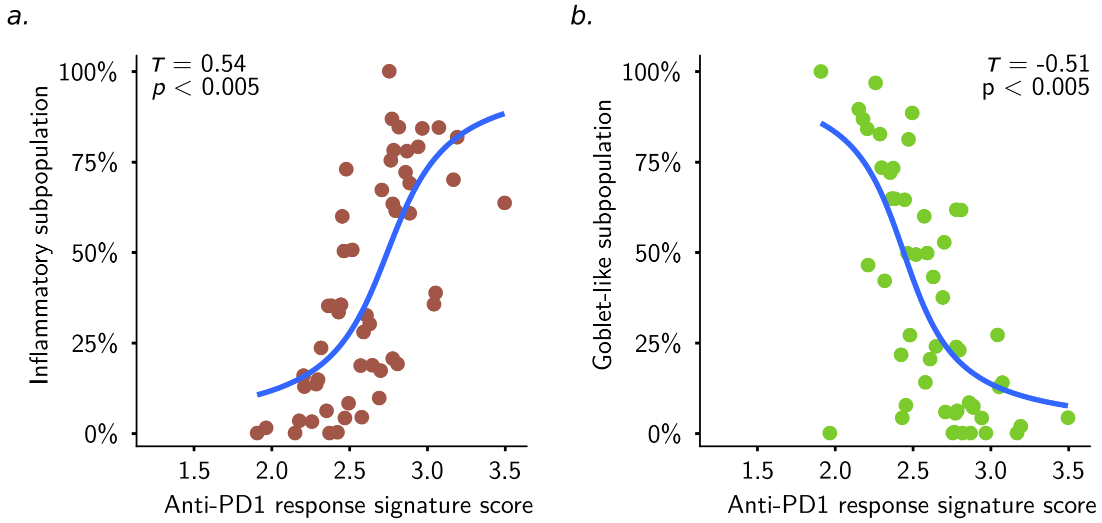


Figure 3.16: Inflammatory and goblet-like subpopulation levels positively and negatively correlate with anti-PD1 response signature scores in MSI-H tumours. Scatter plots showing anti-PD1 signature scores versus subtype subpopulations for MSI-H patients only for the *(a)* inflammatory and *(b)* goblet-like tumour subpopulations ($n = 51$). Correlations values were calculated using Kendall's rank method, and the trend line was fitted using a binomial GLM.

These results show that inflammatory subpopulations correlate with anti-PD1 scores regardless of MSI status, but that goblet-like subpopulations are only correlated with anti-PD1 scores in MSI-H patients. This could be due to the fact that MSI-H tumours are primarily composed of inflammatory and goblet-like subpopulations, as opposed to other subtype subpopulations (Figure 3.13), while MSS tumours can contain heterogeneous subtype subpopulations alongside the inflammatory component. Hence, the negative correlation with goblet-like subpopulations in MSI-H patients could be an artefact secondary to the "true" relationship between anti-PD1 score and inflammatory subpopulation.

It is not unexpected that high levels of inflammatory signalling would be correlated with a score for anti-PD1 response. A high level of tumour inflammation is predictive of responses to anti-PD1 therapies (extensively reviewed in previous literature (Cogdill, Andrews & Wargo, 2017; Chen & Mellman, 2017)). The importance of this result is therefore that it provides evidence that computational deconvolution of subtype subpopulations provides information on the tumours' constituent cell phenotypes which is accurately reflected by other metrics, and which can therefore be used for hypothesis generation in discovering new biomarkers for precision medicine.

3.5 Chapter discussion and conclusions

Some care should be taken when interpreting the results of analysis on public datasets such as those used in this chapter. The survival data from TCGA that was used to quantify the prognoses associated with high levels of the stem-like, inflammatory and goblet-like subpopulations has a higher level of censoring than would be ideal. In the dataset that included patients' responses to FOLFIRI (Del Rio *et al.*, 2007), there were fewer patients whose tumours progressed in the course of their treatment than might be expected. This could bias attempts to associate tumours' shrinkage with their stem-like subpopulations – i.e. it is unknown what the stem-like subpopulation was in tumours that did progress. Finally, when analysing patients' responses to cetuximab in light of their TA subpopulations, the mutation status of RAS family genes other than *KRAS* was unknown, hence some of the patients who appeared not to respond due to low TA subpopulations, could in fact have mutations in other RAS genes. However, the relationship of TA subpopulations with cetuximab response was demonstrated in more recent datasets from both cell lines and PDX models where full *RAS* status was known, strengthening the evidence for this association.

A further consideration should be whether additional cell types could exist in some tumours that cannot be quantified using the methods in this chapter. Because the CRCAssigner subtypes were derived from gene expression profiles of whole tumours, there could be additional CRC cell types which were obfuscated by the fact that bulk expression data is essentially an average measurement of all the cells in the tumour. Whether further cell types exist in CRC will be elucidated as single-cell sequencing becomes more readily accessible.

Additionally, because recent evidence has shown that there is high expression of stem-like genes in the stromal, but not the tumour, compartment of PDX models (Isella *et al.*, 2015), it is possible that the prognostic effect of high stem-like subpopulations is actually due to higher stromal infiltration in these tumours. These samples were not macrodissected before expression profiling so it is likely that some stromal tissue is present in some samples.

Finally, consideration must be made for how patients should be optimally partitioned based on continuous values of subtype subpopulations. In the analyses in this chapter, patients were dichotomised using optimal cutoffs that gave the most discrimination between groups. For any future work moving subtype subpopulations towards clinical applications, standard cutoffs would need to be determined in large, well-selected cohorts, and rigorously validated in unseen data.

However, if appropriate threshold values are determined in future work, the results presented here could have implications on the application of CRCAssigner-based subtyping in the clinic. Previous work has shown FOLFIRI therapy leads to increased side effects and cost per year of life when compared to FOLFOX, with no difference in effectiveness (Neugut *et al.*, 2019). Hence, using stem-like subpopulations to predict irinotecan responders could help spare patients from unnecessary toxicity and reduce care costs versus stratifying patients using bulk subtypes. With regards to anti-EGFR therapy, there appears to be a correlation between TA subpopulations and heightened response, with associated lengthened PFS. This could imply that patients' intratumoural subpopulation of the TA subtype could be used to predict their likelihood of response to anti-EGFR therapy in the future.

In conclusion, I have demonstrated that individual CRC tumours can be comprised of multiple co-existing transcriptomic subtypes, and that the subpopulations of these subtypes can be systematically quantified in large datasets using an *in silico* approach. These subpopulations may also be associated with various patient outcomes. This evidence lays a foundation for the idea that intratumoural subpopulations could become valuable clinical biomarkers for personalisation of CRC treatment in the future.

Chapter 4

Integrated multiomics factor analysis of CRC molecular profiles, interactions between clinicopathological categories, and prognostic implications

4.1 Introduction

Over time, high-throughput platforms have become more accessible and reliable, and it has become more routine to profile multiple different types of "omics" data (e.g. genomics, transcriptomics, epigenomics and proteomics) simultaneously in the same samples. As described in Chapter 1.4, often the way this data is analysed does not fully leverage correlations between features of different omics data types (e.g. cluster-of-cluster analysis). Alternatively, the integration of data types with clinical covariates happens in a post-hoc fashion, using contingency tables after clusters or features of interest have been identified separately from each data type.

Latent variable models – such as iCluster (Shen *et al.*, 2012) for integrating different omics data types, or phenMap (Nyamundanda, Eason & Sadanandam, 2017) for integrating omics data with clinical covariates (developed in the Sadanandam Lab, by Dr Gift Nyamundanda) – can fully exploit correlations between omics features and/or clinical covariates to find underlying patterns in the data. These patterns might represent biological signalling pathways that are co-regulated, and hence may help explain some of the complexities of CRC biology, as previously examined in Chapter 1.4.

Combining the two approaches of multiomics integration and omics/clinical integration could provide a novel tool with the power to explain biological signalling across omics data types using clinical data. In this chapter, I utilise an extended version of phenMap which can both integrate multiple omics data types, and model the patterns in these data types as a function of clinical covariates. This tool, named in this thesis as *integrated sparse Bayesian factor analysis with covariates* (isBFAC, also developed by Dr Gift Nyamundanda), has not previously been applied to CRC, and so could provide new insights into the mechanisms driving this disease.

Figure 4.1 gives an overview of isBFAC, its inputs, and its outputs, which are referred to in this thesis as “metavariables”, due to their being representations of latent patterns that incorportate information from many different data types. Full details of the model are given in Chapter 4.2.1.

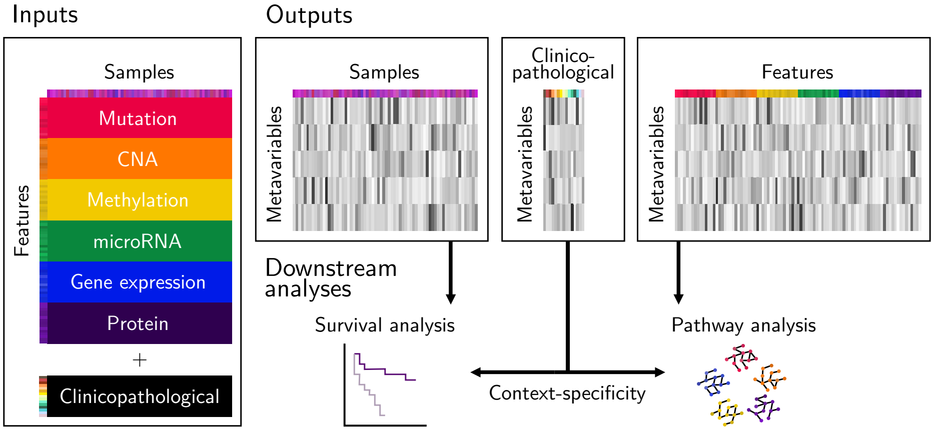


Figure 4.1: isBFAC models multiomics data alongside clinicopathological covariates. Schematic diagram illustrating the inputs and outputs of the isBFAC model, as well as downstream analyses that can be performed using the outputs.

4.2 Model, methods and data sources

4.2.1 Model structure

The isBFAC model was designed, derived and implemented by Dr Gift Nyamundanda, Postdoctoral Fellow at the Institute of Cancer Research, London. I made alterations to the implementation for speed and usability.

For a series of profiles from omics platforms $m = 1 \dots M$, a tumour n's full multiomics profile can be written as $\underline{y}_n = (\underline{y}_{n,1} \dots \underline{y}_{n,M})^{\mathrm{T}}$. Each tumour's omics profile $\underline{y}_{n,m}$ consists of a number of molecular features specific to that omics type, P_m. Then, isBFAC can be written as

$$\underline{y}_{n,m} = \mathbf{W}_m \underline{u}_n + \underline{\xi}_{n,m} \tag{4.1}$$

where $\mathbf{W}_m$ is a $P_m \times Q$ matrix relating the features in omics type m to Q metavariables, and $\underline{\xi}_{n,m}$ is the error term accounting for any remaining variance in the data. $\underline{u}_n$ is a vector of the Q metavariables' weightings in patient n.

Clinicopathological information has also been collected for the patient n, for L different variables, $\underline{x}_n = (x_{n,1} \dots x_{n,L+1})^{\mathrm{T}}$. The metavariables in $\underline{u}_n$ can subsequently be written as a function of the clinicopathological data collected for the patient,

$$\underline{u}_n = \mathbf{B}\underline{x}_n + \underline{\epsilon}_n \tag{4.2}$$

Here, $\mathbf{B}$ denotes a $Q \times (L + 1)$ matrix relating the Q metavariables to the L clinicopathological variables in $\underline{x}_n$, while $\underline{\epsilon}_n$ is the error term accounting for any remaining variance in the metavariables not accounted for by the clinicopathological variables.

$\underline{u}_n$, $\underline{\xi}_{n,m}$ and $\underline{\epsilon}_n$ all follow multivariate normal (MVN) distributions, such that

$$\underline{u}_n \sim \mathrm{MVN}_q(\mathbf{B}\underline{x}_n, \mathbf{\Phi}) \tag{4.3}$$

$$\underline{\xi}_{n,m} \sim \mathrm{MVN}_{P_m}(\underline{0}, \mathbf{\Sigma}_m) \tag{4.4}$$

$$\underline{\epsilon}_n \sim \mathrm{MVN}_q(\underline{0}, \mathbf{\Phi}) \tag{4.5}$$

where the variance matrices $\mathbf{\Phi}$ and $\mathbf{\Sigma}_m$ are diagonal with elements that follow gamma prior distributions,

$$\mathbf{\Phi} = \mathrm{diag}(\varphi_1^2, ..., \varphi_Q^2) \qquad \varphi_q \sim \Gamma(a, b) \tag{4.6}$$

$$\mathbf{\Sigma}_m = \mathrm{diag}(\sigma_{1,m}^2, ..., \sigma_{P_m,m}^2) \qquad \sigma_{n,m} \sim \Gamma(c, d) \tag{4.7}$$

and where $q = 1 \ldots Q$. For binary omics data (i.e. mutations, omics type m_{mut}), the distribution of $\underline{\xi}_{n,m_{mut}}$ is truncated to lie in the interval $(0, \infty)$ for for genes in $\underline{y}_{n,mut}$ that are mutated, and to the interval $(-\infty, 0)$ for genes that are not mutated (in both cases, the standard deviation is 1). The elements of the matrices $\mathbf{W}_m$ and $\mathbf{B}$ are made sparse to force the effects of noisy metavariables, features and clinicopathological variables towards zero. This is achieved in different ways for the two matrices. For $\mathbf{W}_m$, there is a "spike and slab" prior that selects informative metavariables. For informative metavariables, an automatic relevance determination (ARD) prior (Engelhardt & Stephens, 2010) on each element $w_{n,q,m}$ selects informative features,

$$p(w_{n,q,m}|\theta_q, \lambda_{n,q,m}) \sim N(w_{n,q,m}|0, \lambda_{n,q,m}^{-1})\theta_q + (1-\theta_q)\delta_0 \tag{4.8}$$

$$p(\lambda_{n,q,m}|e, f) \sim \Gamma(\lambda_{n,q,m}|e, f) \tag{4.9}$$

$\lambda_{j,k,m}$ is a hyper-parameter for which high values shrink $w_{j,k,m}$ to zero, resulting in few features being deemed relevant to the metavariables. θ_q is a latent binary indicator that is 1 for informative metavariables and 0 for non-informative metavariables.

To induce sparcity in $\mathbf{B}$, a Gaussian g-prior shrinks the elements of $\mathbf{B}$ for which the clinicopathological variable is non-informative (ridge regression (Hoerl & Kennard, 1970)),

$$p(\underline{\beta}_{q,*}|\varphi_q^2, g) \sim \mathrm{MVN}_{L+1}(\underline{\beta}_{q,*}|\underline{0}, g\varphi_q^2) \tag{4.10}$$

Here, $\underline{\beta}_{q,*}$ is the row of $\mathbf{B}$ for the metavariable q, and g is the parameter which controls the level of shrinkage in $\mathbf{B}$.

The values of all the hyperparameters specified $(a, b, c, d, e, f, g, \lambda, \theta)$ were set to be non-informative.

4.2.2 Implementation of the model

4.2.2.1 Preprocessing and initialisation

Within the software implementation of isBFAC, several normalisation steps are performed prior to fitting the model. Firstly, scaling is applied to the omics data (except mutations, which are treated differently inside the model; see Chapter 4.2.1), by dividing each feature by its standard deviation. Continuous clinicopathological data is linearly scaled to lie in the range $(0, 1)$, in order to remove scaling differences between the variables. Categorical variables are coded as seperate dummy variables having values of 0 or 1 for each category before being input into the software. No relationship was assumed between categories of the same variable, and each dummy category variable was included as a seperate variable.

Next, the parameters relevant to Equation 4.1 are initialised. $\mathbf{W}$ is initialised as the first Q eigenvectors of the cross-correlation of the features, and $\mathbf{\Sigma}$ as the mean of the remaining eigenvalues, both of which can then be used to initialise $\mathbf{U}$ using a maximum likelihood estimator.

Subsequently, the parameters relevent to Equation 4.2 are initialised. $\mathbf{B}$ and $\mathbf{\Phi}$ are initialised by fitting a ridge regression on the metavariables $\mathbf{U}$ using the clinicopathological variables $\mathbf{X}$.

4.2.2.2 Fitting

The model is fit using Gibbs sampling, which is a Markov chain Monte Carlo (MCMC) technique. The Gibbs sampler proceeds by sampling from the full conditional distribution of each variable; first $\mathbf{W}$, then $\mathbf{U}$, $\mathbf{\Sigma}$, $\mathbf{\Phi}$, $\mathbf{B}$, $\underline{\theta}$ and $\mathbf{\Lambda}$. This process is repeated for 50,000 iterations.

Gibbs sampling (and all MCMC methods) commonly requires a burn-in period during which the samples move away from the initialisation values and towards values in regions of the conditional distribution with higher probability. As such, the first 20,000 samples are removed from the chain. These samples are also not independent - they are conditioned on the previous iteration. To remove this autocorrelation and make the samples independent, only one in every 40 samples is retained (thinning).

Finally, due to the identifiability issues associated with most factor analysis models, each sample of $\mathbf{W}$ is unpredictably rotated with respect to the other samples. This is corrected by performing Procrustes rotation to align each sample of $\mathbf{W}$ to lie in the same direction as the first sample after burn-in. Each parameter can then be estimated by taking the average of the remaining samples (after removing burn-in and autocorrelated samples) for that parameter.

Figure 4.2 shows a random subset of the elements of the matrices $\mathbf{U}$, $\mathbf{W}$, $\mathbf{B}$ and $\mathbf{\Sigma}$ through the course of the MCMC sampling. Minimal drift of the estimate over the thinned iterations indicates that the model has successfully converged.

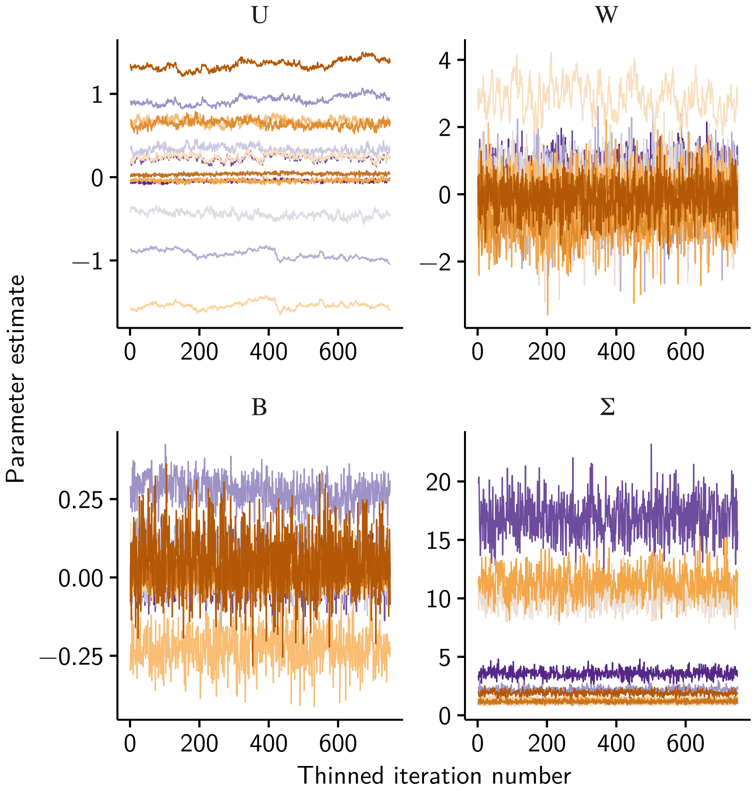


Figure 4.2: The isBFAC model successfully converged over 50,000 iterations. Trace plots showing the thinned samples from model parameters over the course of the isBFAC MCMC chain.

4.2.3 Data curation

I downloaded molecular data from TCGA through the University of California Santa Cruz (UCSC) Xena Browser for the cohorts *TCGA Colon Cancer (COAD)* and *TCGA Rectal Cancer (READ)* on the 1st October 2018, with the exception of protein data which was downloaded from The Cancer Proteome Atlas Portal on the 18th January 2019.

4.2.3.1 Data types, transformations and identifier mapping

4.2.3.1.1 Copy number data. The file *Gistic2_CopyNumber_Gistic2 _all_data_by_genes* from the UCSC Xena Browser contains GISTIC2 estimates of gene-level copy number aberrations for 451 colon and 165 rectal samples (616 total samples), for 24,776 gene identifiers as measured on the Affymetrix SNP 6.0 array. All gene identifiers had no missing data.

Data-specific transformations: None required.

Feature ID mapping: None required.

4.2.3.1.2 Gene expression data. The file *HiSeqV2* from the UCSC Xena Browser contains $\log_2(x+1)$ RNA-seq by expectation maximization (RSEM) gene-level expression estimates for 329 colon and 105 rectal samples (434 total samples), for 20,530 gene identifiers as measured on the Illumina HiSeq 2000 RNA Sequencing platform. All gene identifiers had no missing data.

Data-specific transformations: None required.

Feature ID mapping: None required.

4.2.3.1.3 Methylation data. The file *HumanMethylation450* from the UCSC Xena Browser contains beta estimates of DNA methylation for 337 colon and 106 rectal samples (443 total samples), for 485,577 CpG site identifiers as measured on the Illumina Infinium HumanMethylation450 platform. 376,065 CpG site identifiers had no missing data.

Data-specific transformations: Logit transformation to convert beta-values (ratios) to M-values.

Feature ID mapping: Use the manufacturer-provided manifest file *HumanMethylation450_ 15017482_v1-2.csv* to map CpG site IDs to HGNC symbols. Use CpG site IDs if no HGNC mapping.

4.2.3.1.4 microRNA data.

The file *miRNA_HiSeq_gene* from the UCSC Xena Browser contains $\log_2(x+1)$ reads per million (RPM) estimates of miRNA mature strand expression for 261 colon and 92 rectal samples (353 total samples), for 1,952 miRNA identifiers as measured on the Illumina HiSeq 2000 RNA Sequencing platform. 324 miRNA identifiers had no missing data.

Data-specific transformations: None required.

Feature ID mapping: Use *miRBaseConverter* (Yu *et al.*, 2018) to convert accessions to names (version 22).

4.2.3.1.5 Mutation data.

The file *mutation_bcm_gene* from the UCSC Xena Browser contains gene-level somatic non-silent mutation calls for 217 colon and 81 rectal samples (298 total samples), for 43,916 gene identifiers as measured on the Illumina HiSeq 2000 platform. All gene identifiers had no missing data.

Data-specific transformations: None required for pre-processing.

Feature ID mapping: None required.

4.2.3.1.6 Protein data.

The files *TCGA-COAD-L4* and *TCGA-READ-L4* from The Cancer Proteome Atlas Portal contains level 4 (replicate-based normalised) reverse phase protein array (RPPA) protein expression estimates (Li *et al.*, 2013, 2017b) for 327 colon and 129 rectal samples (456 total samples), for 223 protein identifiers. All protein identifiers had no missing data.

Data-specific transformations: None required.

Feature ID mapping: Use GeneCards/UniProtKB databases to manually map protein names to HGNC symbols.

4.2.3.1.7 Clinicopathological data. The file *clinicalMatrix* from the UCSC Xena Browser contains clinicopathological variables for 462 colon and 169 rectal samples (737 total samples), for 115 variables:

- 10 variables were not useful as they only had one value
- 8 variables were not useful as they had completely identical information to that included in other variables
- 32 variables were not useful as they did not have relevance to the model (e.g. file names, sample barcodes, version codes etc.)

An additional 32 clinicopathological variables from another publication (Liu *et al.*, 2018) were also included, the majority of which related to quantification of immune infiltration and various classes of genomic aberrations.

For input into the model, categorical variables were split into columns as dummy variables, with the most common category acting as the reference.

4.2.4 Pathway analysis

Pathway analysis was carried out using single-sample GSEA (ssGSEA) v4 (Barbie *et al.*, 2009), as implemented by GenePattern (The GenePattern Team, 2013), applied to the matrix of features by metavariables, and using the gene sets included in previous analyses of CRC (Guinney *et al.*, 2015).

4.2.5 Average copy number profiles

Genes were mapped to cytogenic bands using *biomaRt* v2.32.1 and human assembly GRCh27, and average copy number was calculated at each band using thresholded copy number data. For stratification into CIN positive and negative, samples' clonal deletion scores were dichotomised using the cutoff 0.0249 from the original publication that introduced this metric (Liu *et al.*, 2018)

4.2.6 Prediction of sample scores in validation dataset

For prediction of the samples' scores on the metavariables in the validation dataset of copy number, gene expression and methylation data, these datasets were first preprocessed and normalised exactly as the training data were (Chapter 4.2.3). They were then unit scaled. Copy number, gene expression and methylation features were then selected from the matrix $\mathbf{W}$, which gives the weighting of the omics features on the metavariables: $\mathbf{W}' = (\mathbf{W}_{\text{copy number}}, \mathbf{W}_{\text{gene expression}}, \mathbf{W}_{\text{methylation}})$. Elements of $\mathbf{W}'$ for features belonging to data types that were not highly weighted on the metavariables in the training data were set to be missing (NA) in order to reduce the noisiness of the metavariables as predictors (gene expression in MV1, MV2, MV6, MV7 and MV8; methylation in MV3, MV4, MV5, MV6, MV7 and MV8). $\mathbf{W}'$ was then used to predict the omics data for these patients using linear regression ($\underline{y}_{validation} = \mathbf{W}'\underline{u}_{validation} + \underline{\xi}_{validation}$), with the resulting coefficients giving the weights of the patients on each metavariable (approximately equivalent to $\underline{u}$ in Equation 4.1).

4.2.7 Survival analysis

Interactions of metavariables with clinical variables in Cox proportional hazards models were predicted and plotted using the R packages *survival* v1.0.3 and *ggeffects* v0.11.0. Kaplan-Meir curves were plotted by dichotomising patients' metavariable scores using the *surv_cutpoint* function from the *survminer* package v0.4.1, with a minimum group size of 20% of the cohort. Curves were plotted using the *ggsurvplot* function, also from the *survminer* package, and with logrank *p*-values and hazard ratios calculated from the *coxph* function from *survival* v2.41-3.

4.3 Results

4.3.1 Selection of input data

To fully model the omics landscape of CRC, matched molecular data from a large cohort of patients is required. The TCGA collaboration has collected data on a large number of CRC patients that encompasses the major omics types that are measurable on high-throughput platforms (The Cancer Genome Atlas Network, 2012), most notably:

- Copy number aberrations
- Gene expression
- Methylation
- microRNA (miRNA) expression
- Somatic mutations
- Protein expression (from The Cancer Proteome Atlas (TCPA))

Clinicopathological and survival data has also been collected on patients during the course of their treatment, and various additional metrics have been calculated based on the omics data in follow-up studies (Liu *et al.*, 2018).

4.3.1.1 Sample-wise data completeness

A total of 631 CRC primary tumour samples had some form of data available for download from TCGA (be they omics data or clinicopathological variables; detailed in Chapter 4.2). However, only 222 of these could be matched between data types. Within these 222 samples, there remained a large number (approximately 40%) of missing clinicopathological observations. All samples were missing at least one clinicopathological observation, while 81/88 clinicopatholigical variables were missing for at least one sample.

Because removing all the samples with missing values is not an option in this case, and to avoid the large loss of data that would result from removing 81 clinicopathological variables from the dataset, another approach was needed. Therefore, I removed any clinicopathological variables with >5% missing values, before removing any samples that still had missing clinicopathological observations. This left a set of 194 samples and 46 clinicopathological variables having complete, matched data across all omics and clinicopathological variables.

4.3.1.2 Pre-selection of clinicopathological variables and features

Due to the still-high number of clinicopathological variables, some of which contained orthogonal estimations of the same value, a final set of 13 variables was selected as being the most independent of each other, and the most likely to be relevant to CRC:

1. Stage (Stage I-III versus Stage IV)
2. MSI (MSS and MSI-L were combined into MSS/MSI-L)
3. Side (derived from anatomic location)
4. Gender
5. Age
6. CIMP
7. CD8+ T cell fraction
8. Resting natural killer (NK) cell fraction
9. M1 macrophage fraction
10. ABSOLUTE purity
11. ABSOLUTE ploidy
12. Fraction of genome with subclonal somatic copy number aberrations (SCNAs)
13. Clonal deletion score

The quality of the standard clinical data (stage, MSI, side, gender, and age) should be high given the standardised process TCGA data collection centres followed. The other variables selected can be estimated in multiple ways, which are not always so consistent with each other (see, for example, the differences in methods immune cell quantification in (Newman *et al.*, 2015)), and as such, they may need to be interpreted more carefully.

4.3.1.3 Feature selection

Due to the high dimensionality of this data set (465,834 total features from all the molecular data types: gene-level SCNAs, gene-level mRNA expression, CpG site-level methylation, mature miRNA expression, gene-level somatic non-synonymous mutations, and protein-level expression), feature selection was employed to ensure the model could run in a reasonable timeframe. The challenge of feature selection in this dataset is the unbalanced size of the different omics datasets — there are approximately 1,700-fold more CpG sites than proteins profiled. It would have been possible to reduce this imbalance by selecting a fixed number of features from each omics type, however this would have meant that only 1,338 features could be used (223 — the number of proteins profiled — from each of 6 omics datasets), severely limiting the potential for the model to discover novel features. Instead, I chose to reduce the number of features from each omics data type m to a number of features P_m proportional to the log of its original number of features P'_m (so $P_m = \alpha \log P'_m$, unless $P'_m < P_m$, in which case all the original features are retained). This strategy has the benefits of maintaining the order of sizes of the datasets, not reducing the number of features of the already small datasets (i.e. protein and miRNA), while reducing the huge magnitude of the largest datasets (in particular, methylation). The relation between the original number of features and the selected number of features for the omics datasets is shown in Figure 4.3 for $\alpha = 50$ (note: mutations with prevalence <5% were also excluded as they are unlikely to be informative).

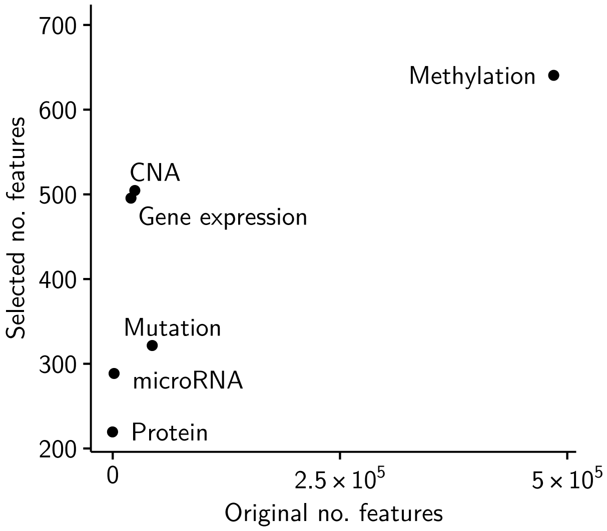


Figure 4.3: Feature selection before modelling allows the model to run faster and with features that are more likely to be of interest in downstream analyses. Scatter plot showing the number of features (e.g. genes, microRNAs, CpG sites) present in the full TCGA dataset, versus the number of features selected for modelling for each data type.

The question then becomes how to select which P'_m features should be included. Two criteria were taken into account to determine a feature's potential relevance, reflecting the two broad frames of reference used in this model: the biological, and the clinical. From the biological point of view, it was determined which features were present in gene sets previously deemed highly relevant to CRC (those included in previous pathway analyses of CRC (Guinney *et al.*, 2015), and those in the KEGG pathway *Colorectal cancer*), as well as which were present in the majority of omics data types ($\geq$ 4). From the clinical perspective, each feature was input into a univariate Cox proportional hazards model of RFS to generate a p-value quantifying its relevance to patients' prognoses (for the 222 patients with complete omics data). The final features chosen for a given omics dataset were then the P'_m that had the lowest p-value *and* were present in the pathways of interest or multiple omics data types.[*]

4.3.2 Metavariables identified from multiomics data

When I applied isBFAC to the TCGA data consisting of 194 patients' omics and clinicopathological profiles (2,473 total omics features; 13 clinicopathological variables), eight metavariables were identified (MV1-8). The weight of patients, features and clinicopathological variables on each metavariable is shown in Figure 4.4.

These metavariables' highest weighted features were often CNAs. However, other types of omics features – particularly methylation and microRNAs, but also gene expression – also had significant weighting.

[*] An exception was made for the miRNA data, for which no straightforward mapping to HGNC symbols/gene identifiers exists, making it impractical to select miRNAs based on gene sets of interest/overlapping genes between omics types. Hence, for the miRNA data, only the Cox p-values were used to select which features should be retained.

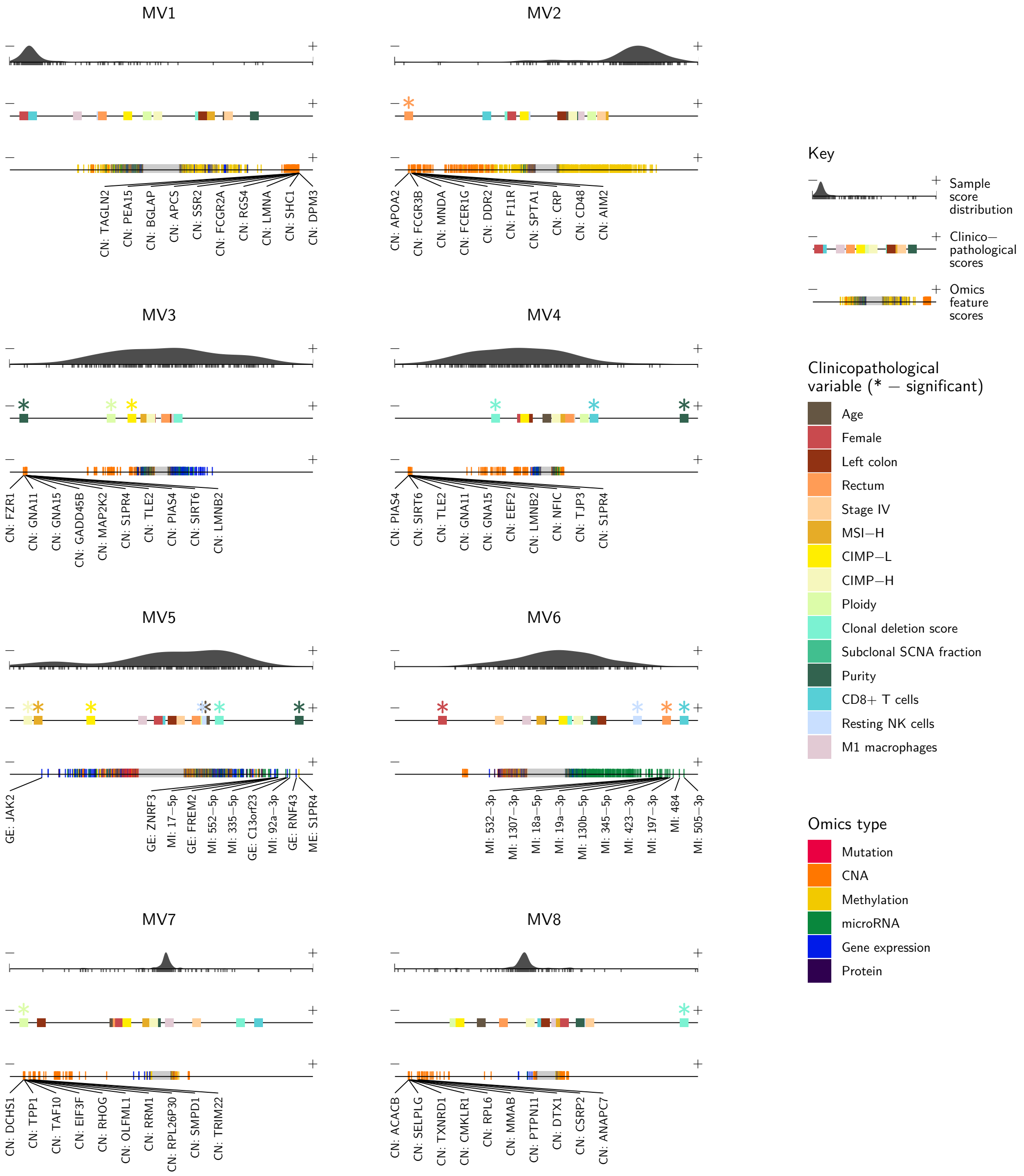


Figure 4.4: Eight metavariables identified by isBFAC have different associations with omics features and clinicopathological factors. Summary of the significant features and clinicopathological variables on each metavariable discovered by modelling TCGA data using isBFAC. The top row of each panel is a density plot showing the distribution of samples' scores on each metavariable; the middle row shows the relative weight of the clinicopathological variables on the metavariable; and the bottom row shows the relative weight and type of features on the metavariable, where the 10 features with the highest magnitude weights are named.

4.3.2.1 Highly-weighted features and clinicopathological variables, including known and novel omics features

Each metavariable was associated with molecular features, as well as clinicopathological variables. These relationships are described below for each metavariable.

4.3.2.1.1 Metavariables 1 and 2. MV1 and MV2 are dominated by CNAs on chromosome 1q21-44 (Figure 4.5). This region encompasses several important members of the cluster of differentiation (CD) family of immune signalling genes, including *CD1A-E*, *CD46*, *CD48*, *CD55*, *CD84*, *CD244*, and *CD247*. Their positive association with MV1 indicates that a subset of patients have increased copy number of these genes relative to the rest of the CRC population.

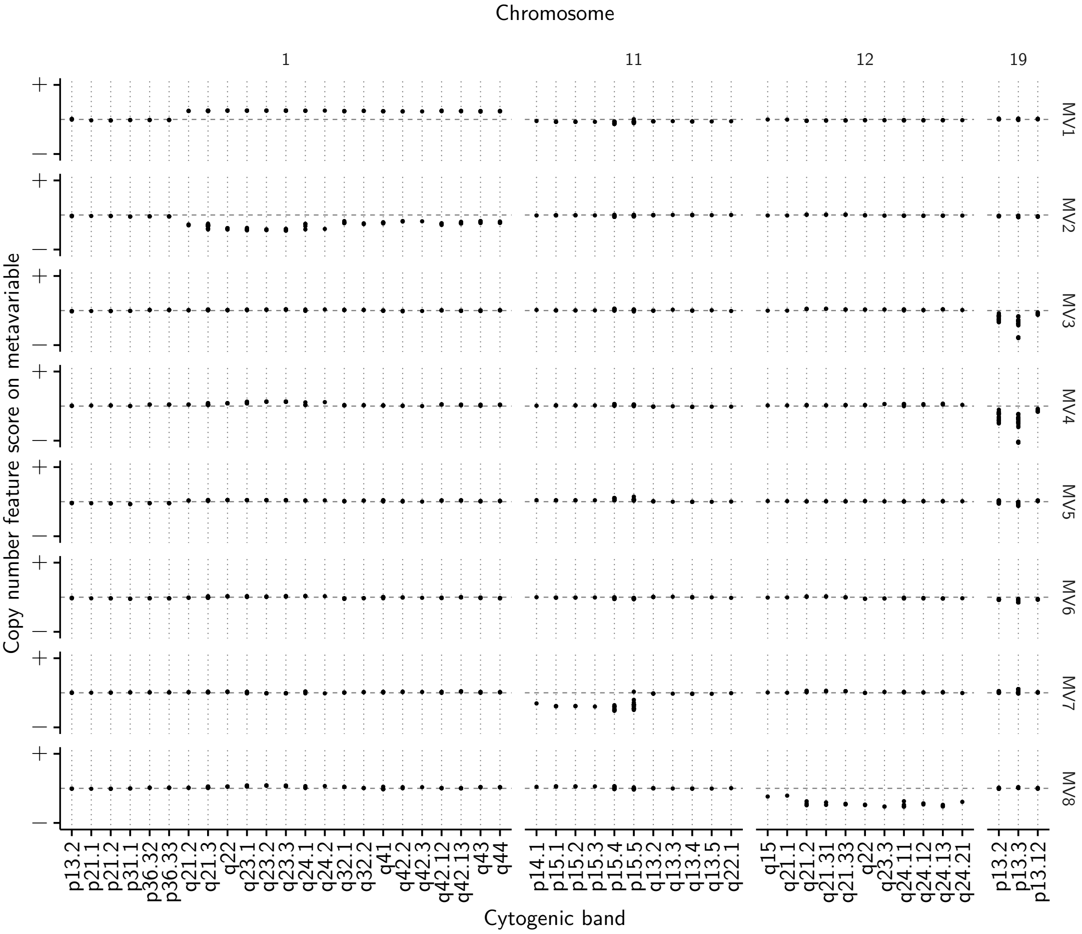


Figure 4.5: The metavariables have different weightings of CNA features that cluster around particular cytogenic bands. Plot showing the weights of CNA features on the metavariables, organised by cytogenic band.

Although not significant, CD8+ T cell infiltration had a low score on MV1, indicating that there is a trend towards higher copy number at these loci in patients with low cytotoxic T lymphocytes. *CD1A-E* are highly expressed in dendritic cells (Leslie *et al.*, 2008), which are involved in the generation of T regulatory cells (Maldonado & Andrian, 2010), potentially explaining this inverse relationship between copy number at these loci and CD8+ T cell infiltration (Figure 4.4).

In the case of MV2, these CNAs and rectal cancers both had a negative association with MV2, implying that rectal cancers have higher copy number of these genes relative to colon cancers. *TP53* was the gene whose mutation was the most strongly weighted on MV2.

Given that the most highly weighted features on MV1 and MV2 were copy number aberrations that came from the same loci, I wanted to understand why both metavariables were chosen by the model, when superficially it would appear that either one alone would be sufficient. When I plotted the average copy number of tumours along chromosome 1q based on the patients' gender and tumour location, it was evident that the effect that tumour location has on copy number is dependent on the patient's gender. Although the weighting of rectal tumours on MV2 implies a higher copy number at bands 1q21-44 relative to colon tumours, this effect was larger for female patients than for males (Figure 4.6). Therefore, the extreme low weighting of female gender on MV1 is likely required for the model to explain the fact that women with colon cancer (a larger group than women with rectal cancer – see Table 4.1 – and therefore having a larger effect on the cohort as a whole) had the lowest copy number at these loci.

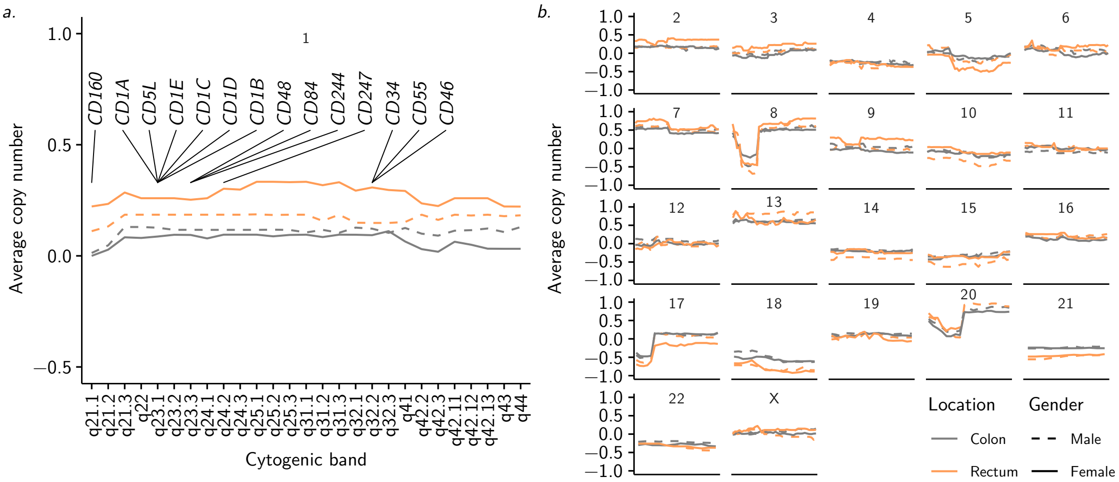


Figure 4.6: Copy number along 1q is influenced by both tumour location and patient gender. Plots showing the average copy number change at each cytogenic band on *(a)* chromosome 1q and *(b)* all other chromosomes for patients with colon or rectal tumours and male or female gender ($n = 194$). Copy number of 0 indicates the normal diploid copy number.

Table 4.1: Women with colon cancer form a larger group than women with rectal cancer, and therefore have a larger effect on the cohort's overall copy number profiles. Cross-table showing the number of female and male patients having colon and rectal tumours in the TCGA cohort.

	Female	Male	*Total*
Colon	63	77	*140*
Rectum	27	27	*54*
Total	*90*	*104*	*194*

4.3.2.1.2 Metavariables 3 and 4. MV3 and MV4 had significant low weightings for CNAs on chromosome 19p13 (Figure 4.5). This cytogenic band carries cancer-associated genes including *BRD4*, *CASP14*, *DNMT1*, *MUC16* and *NOTCH3*. MV3 was also negatively associated with tumour purity, ploidy and CIMP-L. The subset of patients having low scores on MV3 had highly pure tumours with increased ploidy and a CIMP-L phenotype, with relatively high copy number at 19p13 compared to CIMP-0/H, low-ploidy/purity tumours. *TP53* mutation was the mutation with the most negative weight on MV3.

Conversely, on MV4, there was a positive association with tumour purity, and a strong positive weight for *TP53* mutation. Additionally, there was another positive association with CD8+ T cell infiltration, and a negative association with subclonal SCNA fraction. The subset of patients who score highly on this metavariable then represent those who have highly pure tumours with relatively low copy number at 19p13, compared to the rest of the CRC population.

As with MV1 and MV2, there was a strong similarity between the highly-weighted features on MV3 and MV4. In this case, *TP53* mutation switched from being anti-correlated with MV3 scores to correlated with MV4 scores. Moreover, tumour purity switched from being correlated with these features on MV3 to being anti-correlated with them on MV4. This can be explained by MV3's association with CIMP and MV4's association with clonal deletion score, a measure of CIN (Liu *et al.*, 2018). MV3's negative association with tumour purity, CIMP-L and 19p13 implies that highly pure, CIMP-L tumours should have high copy number at these loci. In fact, this is only the case when the tumour is CIN as well as CIMP-L (Figure 4.7). For the majority of patients (i.e. those who do not have both CIMP-L and CIN), high purity is not associated with higher copy number at these loci, as represented by MV4. Hence, MV3 and MV4 were both included by the model so that this interaction between CIMP-L, CIN and tumour purity was accounted for. In summary, the subset of patients who have CIN, CIMP-L, high-purity tumours have the highest copy number along chromosome 19p.

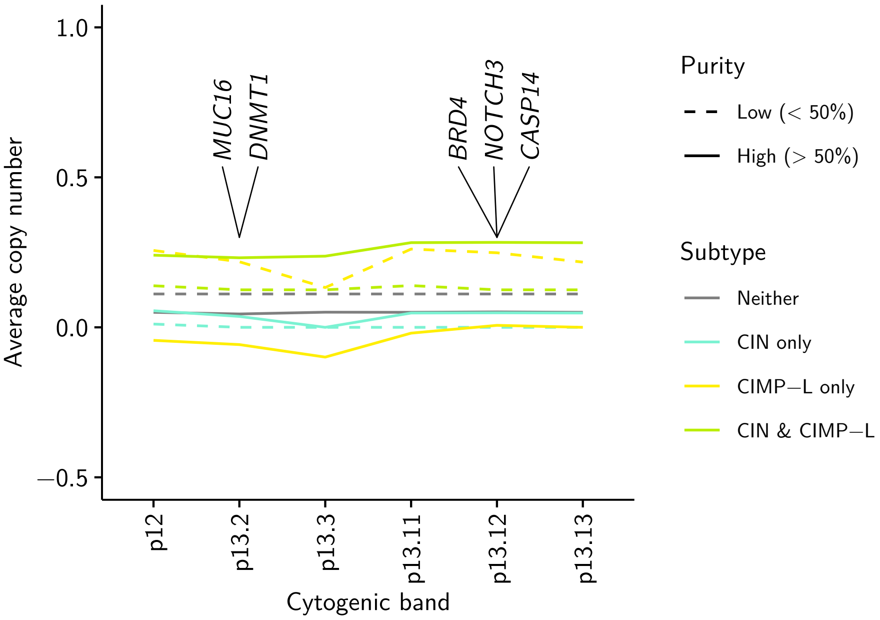


Figure 4.7: Copy number along 19p is influenced by tumour purity, CIMP-L status and CIN status. Plot showing the average copy number change at each cytogenic band on chromosome 19p for patients with CIN or CIMP-L positive or negative tumours (n = 194). Copy number of 0 indicates the normal diploid copy number.

4.3.2.1.3 Metavariable 5. MV5 had significant association with multiple clinicopathological variables and multiple types of omics features. Positive scores on MV5 were associated with tumour purity, clonal deletion score, age and resting NK cell infiltration. Features with positive scores included methylation of the *S1PR4* gene, whose protein expression has been shown to be significantly higher in gastric adenocarcinomas versus benign tissue (Wang *et al.*, 2014). The negative end of MV5 was associated with CIMP-H and MSI-H, and to a lesser extent CIMP-L. Features that had negative scores included *JAK2* gene expression, the key promoter of cell proliferation (Ihle & Gilliland, 2007). This indicates that the minority of patients who are CIMP-H, MSI-H, and CIN negative have high expression of *JAK2* relative to the rest of the CRC population, corroborating recent findings (Peng *et al.*, 2018) of high *JAK2* expression in MSI-H cancers.

4.3.2.1.4 Metavariable 6. MV6 was the only metavariable that was significantly associated with female gender (at the negative end), and also the only metavariable whose significant features were dominated by miRNAs (at the positive end). The highest-weighted miRNA, miR-505-3p, has been found in one study to be a tumour suppressor in lung cancer (Tang *et al.*, 2019), but was upregulated in synovial sarcoma (Fricke *et al.*, 2015). miR-484 expression has been suggested to be attenuated in stage I-II CRC, and to increase at stages III-IV (Lu & Lu, 2015). It has also been shown to have low expression in MSI CRC tumours, and acted as a tumour suppressor in MSI cells *in vitro* and *in vivo* (Mei *et al.*, 2015). miR-18a-5p expression has been associated with better prognosis in CRC (Slattery *et al.*, 2015). However, other miRNAs highly weighted on MV6 have been found to have tumour enhancing effects, such as miR-423-3p (Li *et al.*, 2015), or correlate with poor prognosis, such as miR-345-5p (Yu *et al.*, 2016). miR-130b inhibits the tumour suppressor PTEN's expression (Zhu *et al.*, 2014), while exosomal miR-19a-3p has been proposed as a biomarker for poor prognosis in CRC (Matsumura *et al.*, 2015). Therefore, no conclusion can be drawn as to whether the microRNAs highly weighted on MV6 have a tumour-promoting or suppressive effect overall, and they could play highly context-specific roles in different cancer types. However, the inference that males may have higher expression of certain cancer-associated microRNAs than females could have implications for maturing early-detection approaches based on microRNAs (Ng *et al.*, 2009). The small cluster of CNAs towards the negative end of MV6 were associated with chromosome 19p13

(Figure 4.5), similarly to MV3 and MV4.

4.3.2.1.5 Metavariable 7. The most highly weighted CNAs on MV7 had a negative association with this metavariable, and were clustered around chromosome 11p14-15 (Figure 4.5). This cytogenic band includes several important cancer-associated genes such as *SOX6*, *PIK3C2A*, *ADM*, *WEE1*, and *HRAS*, as well as several mucins, some of which are implicated in CRC (Byrd & Bresalier, 2004) (Figure 4.8). This metavariable was also negatively associated with ploidy. Patients who had low scores on this metavariable therefore had high ploidy and relatively high copy number of genes at these loci compared to other patients, and their tumours also (non-significantly) tended towards left-sidedness.

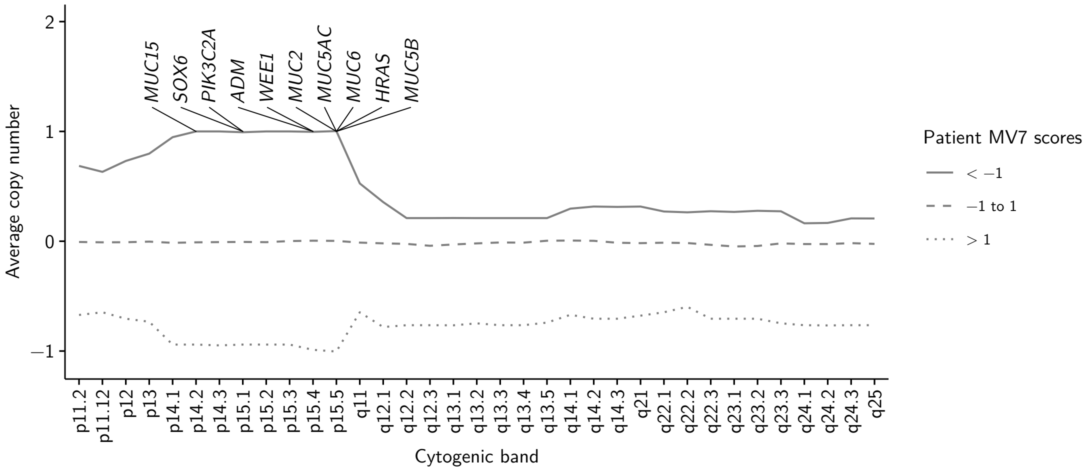


Figure 4.8: Patients with low scores on MV7 have the highest copy number along chromosome 11. Plot showing the average copy number change at each cytogenic band of chromosome 11 for patients with high, medium or low MV7 scores ($n = 194$). Copy number of 0 indicates the normal diploid copy number.

4.3.2.1.6 Metavariable 8. On MV8, CNAs on chromosome 12q15-24 had significant negative weightings (Figure 4.5). Key cancer-associated genes that lie in this region include *IGF1*, *MDM2* and *POLE*, a gene whose mutation is associated with hypermutation (Palles *et al.*, 2013) (although *POLE* mutation did not score highly on any metavariable), but for which copy number aberrations have been less explored. *TP53* was the mutation with the highest positive score on MV8. Clonal deletion score was significantly positively associated with this metavariable. Therefore, patients having CIN tumours (i.e., high clonal deletion score) had relatively low copy number at these loci compared to patients who were CIN negative.

4.3.2.2 Metavariable-specific gene set enrichment and pathway analysis

To discover whether there were distinctive pathways represented by the features highly weighted on each metavariable, I performed single-sample gene set enrichment analysis (ssGSEA) on each data type in each metavariable. Figure 4.9 shows a summary of the results where the scores for each data type have been summed within each metavariable.

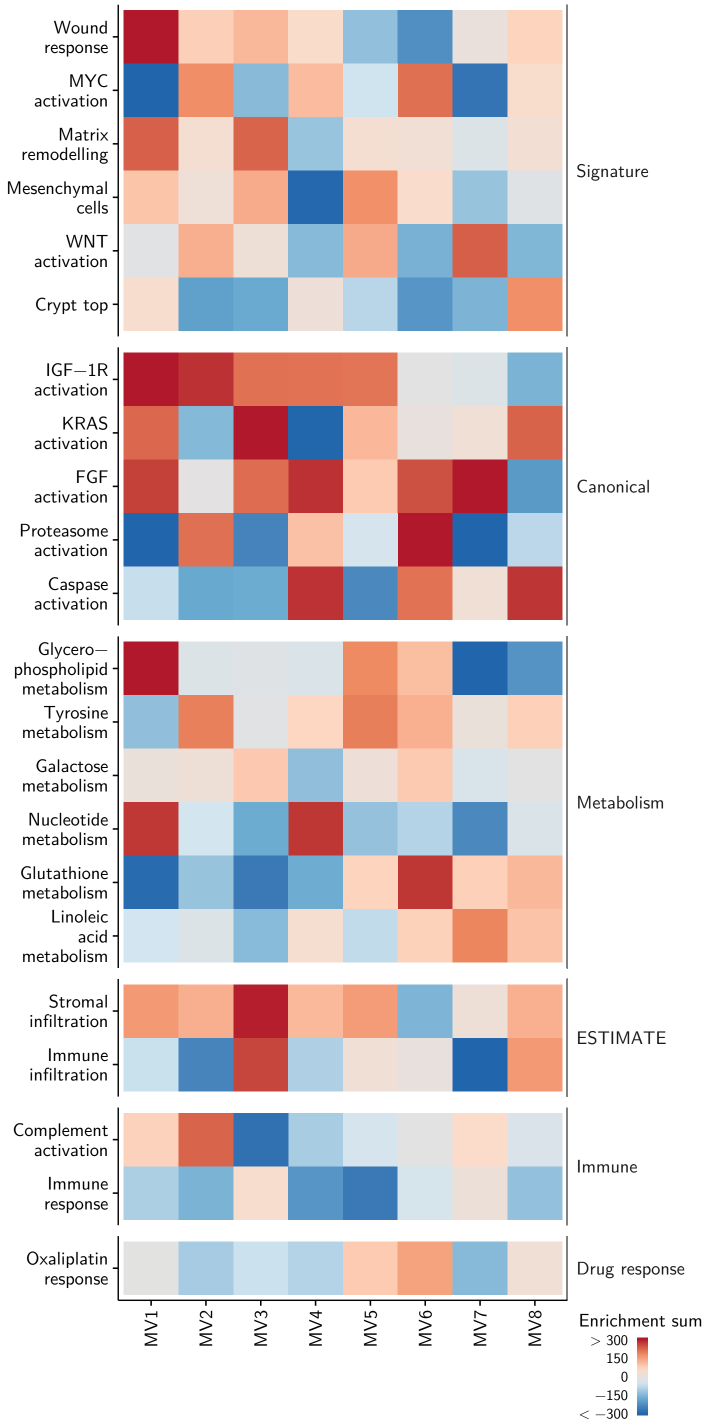


Figure 4.9: Different pathways are active in the features associated with each metavariable. Heatmap of gene set enrichment analysis of features on each metavariable. Enrichment scores for each pathway in each data type were summed to give an overall score for each metavariable. Enrichment scores for methylation features had their sign reversed before summing such that each data type followed the same basis of higher enrichment equating to higher expression downstream (in principle).

On MV1, the high enrichment of wound response, IGF-1R, KRAS and FGF activation, glycerophospholipid and nucleotide metabolism points to a highly proliferative phenotype. MV1 and MV2, which had high scores for the same copy number loci (1q21-44), had the highest enrichment for complement activation gene sets. Dendritic cells, markers for which lie on 1q21-44, (Chapter 4.3.2.1.1) can express certain complement proteins (Lubbers *et al.*, 2017), potentially explaining this activation.

MV3 and MV4 both had negative associations with copy number at chromosome 19p13 (Figure 4.5). However, MV3 had markedly higher scores for stromal and immune infiltration than MV4 (Figure 4.9), which ties in with MV3's negative association with tumour purity (Figure 4.4).

MV5 showed enrichment of gene sets representing mesenchymal cells and WNT activation, which may indicate a phenotype similar to the bottom of the colon crypt. MV6 had the highest enrichment of MYC-associated genes, and an accompanying high score for the antioxidant glutathione's metabolism. Hence, in patients that score highly for this metavariable, tumours driven to proliferate through MYC signalling may protect themselves from the resulting oxidative stress through increased glutathione metabolism, a previously reported mechanism of tumour promotion (Benassi *et al.*, 2006).

MV7 showed the highest enrichment for WNT and FGF pathway activation, suggesting an important role for cancer-associated fibroblasts (CAFs) – which can trigger WNT signalling in tumour cells (Fu *et al.*, 2011; Aizawa *et al.*, 2019) – in patients who score highly on this metavariable. In MV8, there was high enrichment for genes associated with the top of the colon crypt – where more differentiated cells lie – and associated low enrichment of WNT signalling and mesenchymal genes. Caspase and KRAS activation were also both highly enriched, and there is evidence that RAS can promote caspase-mediated apoptosis (Pylayeva-Gupta, Grabocka & Bar-Sagi, 2011). Hence, patients' scores on MV8 could represent an axis of activation or suppression of RAS-triggered cell death.

4.3.3 Prognostic value of the metavariables

I then sought to determine whether these metavariables – which are associated with several distinct molecular features and biological pathways – could provide any prognostic insight. In a multivariable Cox model of patients' RFS that included patients' standardised scores for all the metavariables, high scores on MV1, MV2, MV3, MV5 and MV7 conferred a significantly worse hazard (Figure 4.10). Interestingly, these were the metavariables that had the highest level of WNT activation according to gene set enrichment analysis (Figure 4.9).

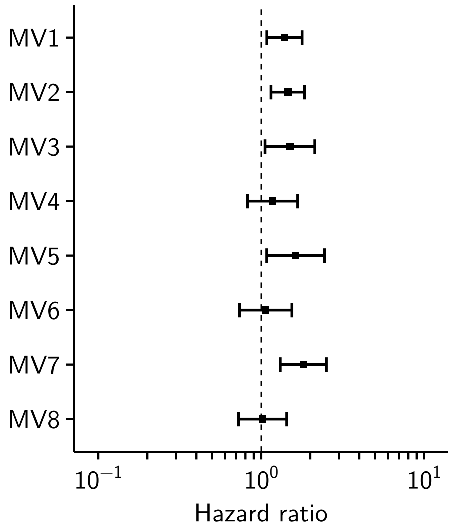


Figure 4.10: Metavariables MV1, MV2, MV3, MV5 and MV7 have significant prognostic value. Forest plot showing the hazard ratios for each standardised metavariable in a multivariable Cox model of RFS in the TCGA training cohort ($n = 180$).

4.3.3.1 Validation of prognostic power of the metavariables

To validate this finding, I then predicted the metavariable scores of 173 patients from TCGA who were not included the data used to fit the isBFAC model due to having missing omics or clinicopathological data, but who had gene expression, copy number and methylation data available (TCGA validation cohort; scores predicted using linear regression, see Chapter 4.2.6). These three data types were dominant in the highly weighted features of the prognostic metavariables. In these patients, and using only gene expression, copy number and methylation features from the metavariables, MV3, MV5 and MV7 remained significantly associated with prognosis (Figure 4.11). However, the hazard ratio of MV7 flipped to indicate a favourable prognosis for patients with high MV7 scores.

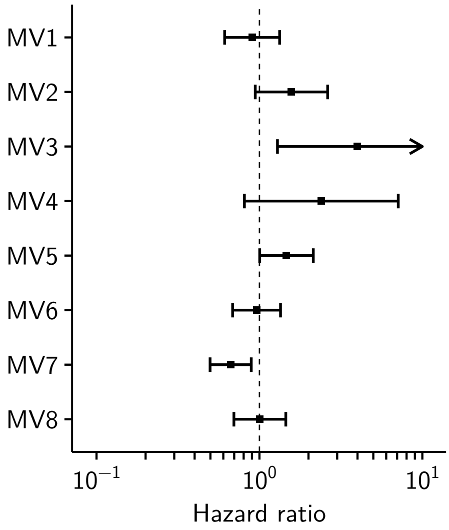


Figure 4.11: The hazard ratio of MV7 flips sign in the validation dataset. Forest plot showing the hazard ratios for each metavariable in a multivariable Cox model of RFS in the TCGA validation cohort ($n = 148$).

To investigate why MV7 could be associated with favourable prognosis in one cohort and unfavourable prognosis in another, I predicted prognosis of the training cohort using a Cox model that included interaction terms between MV7 and the clinical covariates. This revealed a near-significant interaction between MV7 and location of the tumour in predicting survival ($p = 0.059$). Figure 4.12 shows the estimated hazard ratio for patients with colon or rectal tumours at different values of MV7, demonstrating different associations of MV7 with hazard ratio between these two tumour locations.

When patients were dichotomised into high- and low-MV7 groups, patients with colon cancer had worse prognosis when they had low MV7. Conversely, patients with rectal cancer who had low MV7 had significantly better prognosis than those who had high MV7 (Figure 4.13).

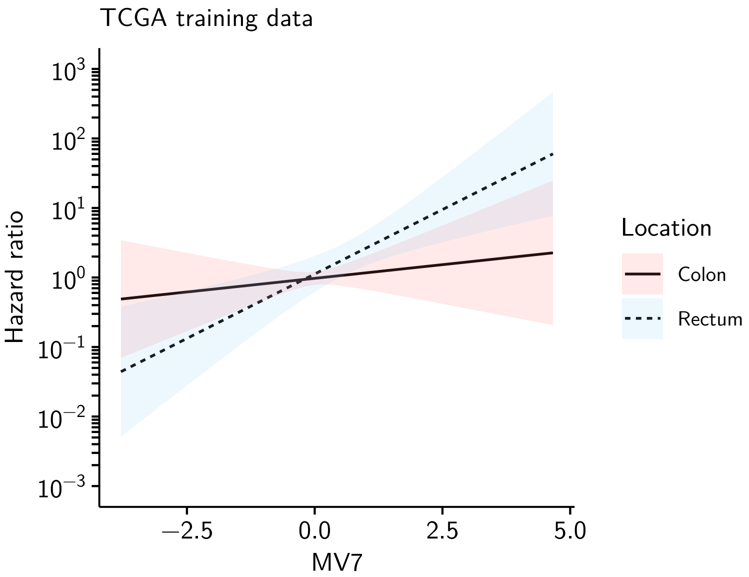


Figure 4.12: MV7 scores have different effects on prognosis for patients with colon or rectal tumours - training cohort. Interaction plot demonstrating the hazard ratio of RFS for TCGA training cohort patients with colon or rectal tumours according to their scores on MV7 ($n = 180$). Coloured areas designate 95% confidence intervals.

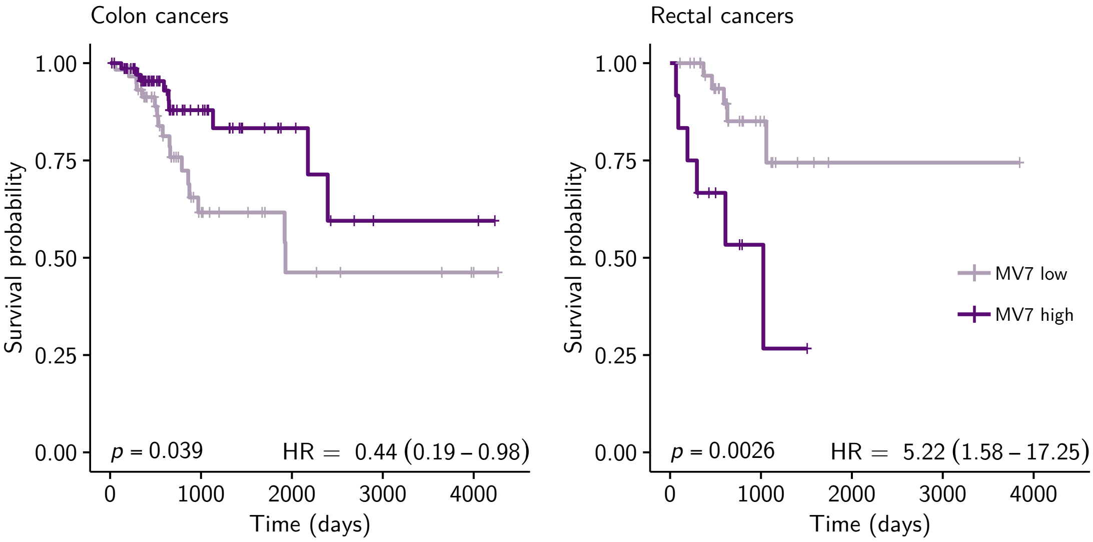


Figure 4.13: High MV7 scores have a positive effect on colon cancer patients' prognosis, and a negative effect on rectal cancer patients' prognosis - training cohort. Kaplan-Meier curves showing the RFS of colon ($n = 73$ high MV7; 59 low MV7) and rectal ($n = 12$ high MV7; 36 low MV7) cancer patients with high or low scores on MV7 in the TCGA training cohort.

When I performed the same analysis in the validation TCGA cohort, colon and rectal cancers again experienced different hazards based on MV7 score (Figure 4.14). Patients with colon cancer had significantly worse prognosis when they had low MV7, and those with rectal cancer fared significantly worse when they had high MV7 (Figure 4.15).

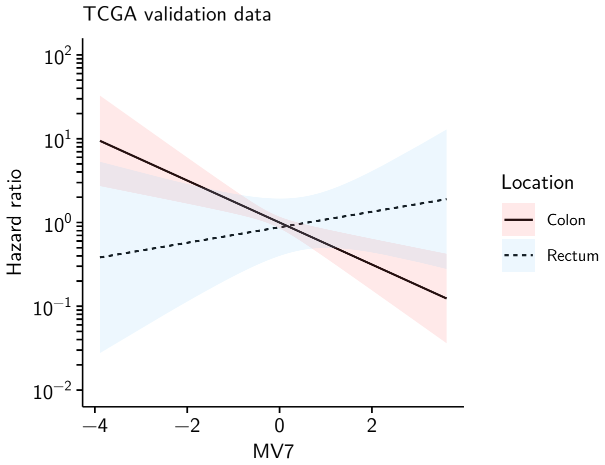


Figure 4.14: MV7 scores have different effects on prognosis for patients with colon or rectal tumours - validation cohort. Interaction plot demonstrating the hazard ratio of RFS for TCGA validation cohort patients with colon or rectal tumours according to their scores on MV7 ($n = 148$). Coloured areas designate 95% confidence intervals.

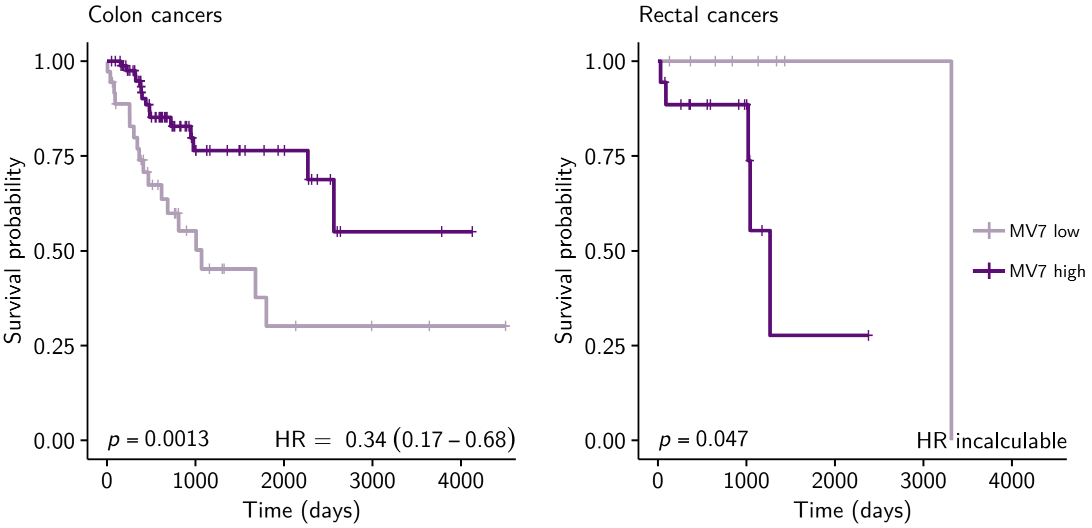


Figure 4.15: High MV7 scores have a positive effect on colon cancer patients' prognosis, and a negative effect on rectal cancer patients' prosnosis - validation cohort. Kaplan-Meir curves showing the RFS of colon ($n = 86$ high MV7; 36 low MV7) and rectal ($n = 18$ high MV7; 8 low MV7) cancer patients with high or low scores on MV7 in the TCGA validation cohort.

Given these concordant results between the training and validation cohorts, the flip in hazard ratio for MV7 between the cohorts is likely explained by the fact that there were significantly fewer rectal cancers in the test cohort (Table 4.2). In the training cohort, the negative effect that high MV7 had on the prognosis of patients with rectal cancer was greater than the positive effect on patients with colon cancer's. Hence, where there were fewer patients with rectal cancer in the validation cohort, the negative effect on the prognosis of the cohort as a whole was attenuated, leaving only the positive effect on the prognosis of patients with colon cancer to dominate the cohort.

Table 4.2: Comparison of clinicopathological covariates between training and validation TCGA datasets. For categorical variables, the number and percentage of patients in each category is shown, with *p*-values calculated using a χ^2 test. For continuous variables, the mean, standard deviation and range is shown, and *p*-values were calculated by unpaired t-test or Mann-Whitney U test where variables cannot be assumed to follow a normal distribution (i.e. proportions).

Variable	Multiomics training dataset ($n = 194$)	Gene expression/copy number/methylation validation dataset ($n = 173$)	P-value
Gender			0.63
Female	90 (46%)	75 (43%)	
Male	104 (53%)	98 (56%)	
Location			0.012*
Left colon	57 (29%)	42 (24%)	0.33
Right colon	83 (42%)	100 (57%)	0.0056*
Rectum	54 (27%)	31 (17%)	0.034*
Stage			0.30
Stage I-III	171 (88%)	145 (83%)	
Stage IV	23 (11%)	28 (16%)	
CIMP			0.45
CIMP-0	93 (47%)	91 (52%)	
CIMP-L	73 (37%)	64 (37%)	
CIMP-H	28 (14%)	18 (10%)	
MSI			0.76
MSS/MSI-L	165 (85%)	150 (86%)	
MSI-H	29 (14%)	23 (13%)	
Age	65 (13, 31-90)	64 (14, 31-90)	0.39
Ploidy	3 (1, 2-6)	3 (1, 2-4)	0.19
Clonal deletion score	9% (7%, 0-30%)	10% (8%, 0-25%)	0.36
Subclonal SCNA fraction	22% (21%, 0-88%)	21% (20%, 0-82%)	0.64

Variable	Multiomics training dataset (n = 194)	Gene expression/copy number/methylation validation dataset (n = 173)	P-value
Purity	64% (17%, 17-95%)	63% (17%, 19-94%)	0.84
CD8+ T cells	11% (8%, 0-61%)	11% (8%, 0-41%)	0.88
Resting NK cells	2% (3%, 0-17%)	3% (3%, 0-15%)	0.15
M1 macrophages	6% (4%, 0-22%)	5% (4%, 0-19%)	0.026 *

4.4 Chapter discussion and conclusions

In this chapter, I have shown that copy number aberrations were the most dominant type of feature in this dataset, but that methylation, gene expression and microRNAs also played important roles. While it could be expected that protein expression should have the highest weighting in explaining molecular variation in cancer, proteins' low weightings on the metavariables might be explained by the known difficulty in properly normalising RPPA data (Neeley *et al.*, 2009; Akbani *et al.*, 2014). Hence, as high-throughput proteomics matures, the weight of proteins on the metavariables could increase if this same model was re-run in the future.

In addition, I have evidenced that the results from latent variable models which include clinical covariates must be interrogated to check for interactions between multiple clinical variables and particular features. Such interactions can lead to spurious conclusions if the metavariables are interpreted separately and not in the context of each other (e.g. MV1/MV2's interaction with tumour location and gender, and MV3/MV4's interaction with tumour purity and CIMP/CIN). The decision to investigate these interactions came from the high weightings of the same molecular features on multiple metavariables which had differently weighted clinicopathological variables, and this highlights the value of including clinical covariates in the modelling process; in a factor analytic model without covariates, there likely would have been only one metavariable with these features having high weighting, and these important interactions would not have been highlighted.

I have also provided evidence that the metavariables give significant prognostic information, but that this information needs to be interpreted carefully in the light of clinical variables to give the full picture, such as MV7's opposite effect on prognosis in colon and rectal tumours. Future validation of these prognostic effects will require the accurate prediction of the metavariables in new datasets, a process which will require extensive validation in itself.

It is likely that the results in this chapter were heavily influenced by the particular features that were chosen as input. Features were chosen based on their known association with CRC, and their prognostic value. While these criteria likely increased the interpretability and prognostic power of the metavariables, it could be preferable for some purposes to run the model using a much broader range of features, for example to discover totally novel drivers in CRC.

Another criticism of this analysis could be the lack of an independent dataset to demonstrate that the results presented are not the result of overfitting. Unfortunately, the scale and diversity of data captured by TCGA has not yet been matched, and so it is not currently possible to run this model on an equivalent dataset. However, several aspects of the model are designed to prevent overfitting, such as priors applied to the weights of features and clinical covariates on the metavariables which shrink all but the largest weights to zero.

In summary, I have demonstrated the application of a novel multiomics and clinicopatholigical integration tool, isBFAC, to CRC. Overall, this new approach to data integration gives a promising avenue for more holistic explorations of tumour biology, which take both molecular and clinical data into account.

Chapter 5

Conclusions, discussion, and future work

5.1 Thesis conclusions and impact

In this thesis, I have presented analyses that have highlighted the extensive inter/intratumoural and intermolecular heterogeneity that exists in CRC. The understanding of CRC's intertumoural heterogeneity that was garnered through the simultaneous efforts of multiple teams from 2012-2013 (Budinska *et al.*, 2013; Marisa *et al.*, 2013; Roepman *et al.*, 2013; De Sousa E Melo *et al.*, 2013; Schlicker *et al.*, 2012; Sadanandam *et al.*, 2013) has matured to the point where standardised and practical subtyping assays are the necessary next step for assessing the value tumour subtyping could have in the clinic. NanoCRCA, the assay whose development is described in Chapter 2 and has been published for use by the wider community (Ragulan *et al.*, 2019), provides this requisite. This assay is faster, more affordable and easier to implement for both research and potential clinical applications, due to its clinically-approved platform (Wallden *et al.*, 2015; Northcott *et al.*, 2012; Scott *et al.*, 2014) and applicability to widely-collected FFPE tissue. Efforts are ongoing within the Sadanandam Lab at the Institute of Cancer Research to collect clinical trial samples to evaluate the predictive potential of the subtypes, particularly the TA subtype's sensitivity to cetuximab (Fontana *et al.*, 2018).

With regards to the intratumoural heterogeneity of CRC, I have demonstrated that the tumour's transcriptomic milieu can be deconvoluted into subpopulations of gene expression subtypes, and that these subpopulations have prognostic and predi ctive value beyond what can be achieved with "bulk" subtyping – labelling a tumour as belonging to a single subtype. In particular, I was able to show that a portion of the previously reported variability in responses to cetuximab from tumours in the TA subtype can be explained by these bulk-TA tumours having different intratumoural subpopulations of TA-subtype cells. In addition, MSI-H tumours appear to be composed primarily of variable proportions of goblet-like and inflammatory subpopulations, which have contrasting prognostic implications and opposite associations with transcriptomic biomarkers of response to anti-PD1 immunotherapy. This finding of transcriptomic heterogeneity within MSI-H tumours could explain why approximately 50-60% of cancers with microsatellite instability do not respond to PD1 blockade (Le *et al.*, 2015, 2017).

In moving beyond transcriptomic analysis to holistic integration of multiple molecular data types with clinicopathological covariates, I was able to demonstrate that there exist “pan-omic” patterns of expression that are prognostic in CRC, and which include both known drivers of cancer and biomarkers which are novel in CRC. One example was copy number changes on chromosome 11p14-15, where the highest-weighted features on the most prognostic metavariable, MV7, lie. Copy number aberrations at this locus have been little-explored in the CRC literature, with reports limited to their apparent appearance during metastasis to the liver (Stange *et al.*, 2010). A further important lesson learned from this analysis was that the associations between clinicopathological variables and omics features cannot be interpreted separately for each metavariable wherever multiple metavariables have high weightings of the same features. Instead, overlapping features between metavariables indicates there is some interaction present between clinicopathological variables that must be carefully evaluated.

5.2 Opportunities for improvement in future work

5.2.1 Maturation of the NanoCRCA assay for clinically-practicable subtyping

Further efforts are needed in order to fully develop the NanoCRCA assay described in Chapter 2 to the point where it could be used routinely for clinical decision making, as discussed below.

One key finding of my efforts to apply the NanoCRCA assay to matched fresh frozen and FFPE tissue was that normal tissue contamination can greatly affect subtyping results (Chapter 2.3.5), an issue that can particularly affect non-macrodissected fresh frozen tissue due to the similarity between normal colon tissue and the enterocyte CRC subtype. While this should not affect the results of subtyping for patients' clinical samples, which are normally FFPE and which can be routinely macrodissected, it could have implications for future work using this assay in the research setting, where fresh frozen tissue is the gold standard. Normal tissue-like subtypes (such as the enterocyte subtype) are not exclusive to CRC, appearing for example in breast cancer (Perou *et al.*, 2000), and the question remains as to whether normal-like malignant tissue could be differentiated from truly normal tissue using gene expression alone.

Furthermore, due to the limited size and clinical annotation of cohorts available for profiling on the NanoCRCA assay during its development, I was not able to validate the prognostic power of the subtypes reported in the original CRCAssigner publication (Sadanandam *et al.*, 2013). This can be rectified with time as follow-up data for those cohorts already profiled becomes mature, and as access becomes available to FFPE tissue from retrospective cohorts that already have mature follow-up.

Finally, data normalisation for my NanoCRCA analysis followed previously published work (Sadanandam *et al.*, 2013) in that datasets were median centred gene-wise, which has the undesirable side-effect of causing the gene expression profile (and hence the subtype) of each sample to be dependent on the gene expression profile of the other samples profiled in that same dataset. In the future, a new algorithm for classification should be adopted which does not have this prerequisite, allowing for truly independent subtyping of individual samples. Such an algorithm could depend on normalisation to an artificial reference RNA control on each nCounter cartridge, as is used by the Prosigna PAM50 assay (NanoString Technologies, 2016), or it could exploit within-sample gene ranks (Tan *et al.*, 2005).

5.2.2 Practical considerations of using transcriptomic subtype subpopulations for patient stratification of prognosis and personalised therapies

While the intratumoural subpopulations of the CRCAssigner subtypes of CRC showed both prognostic and predictive potential (Chapter 3) there are additional analyses that must be performed to fully develop this concept for potential clinical applications.

For the analyses in Chapter 3, wherever patients were dichotomised into groups using their estimated intratumoural subtype subpopulations, optimal cutoffs were calculated that provided the most discrimination between groups. These cutoffs differed between subtypes, but also between cohorts. This was likely due to the differences in platforms used to profile the patients' gene expression, but also differences in the clinical characteristics of the cohorts. For any future clinical applications, cutoffs would have to be trained from large, well-selected groups of patients that have mature follow-up data, and rigorously validated in samples that were unseen during the training step.

In addition, consideration needs to be made for which platform should be used as standard for profiling of subtype subpopulations in a clinical setting. Whether the same SVR algorithm presented in this thesis could be applied to, for example, the data collected using the NanoCRCA assay in Chapter 2, should be assessed, given the lower number of genes profiled on that platform. It has not been tested in this work whether the full set of 786 genes that originally defined the CRCAssigner subtypes (Sadanandam *et al.*, 2013) is required for accurate deconvolution, or whether the smaller 38-gene set could be sufficient for this purpose.

Lastly, the application of deconvolution to gene expression profiles from whole tumours means there is potential for inaccurate results in some tumours due to the infiltration of stromal or immune components of the microenvironment into the tumour. Given the similarity of some of the CRCAssigner subtypes to normal colon tissue (Appendix D) or stroma (Isella *et al.*, 2015), (an issue highlighted in Chapter 2), it is possible that the estimates of the subpopulations for these subtypes could be influenced by low tumour purity. As mentioned in the previous section, it is not yet clear whether these non-malignant and malignant tissue cell types can be distinguished in any practical way, but to do so might require the comparison of these tissues after microdissection (and ideally, single-cell sequencing) in order to refine the gene signatures of the malignant subtypes.

5.2.3 Refinement and further extensions of multiomics modelling of CRC with clinicopathological variables

The results of Chapter 4 included novel muli-omics biomarkers of prognosis in CRC, however, they required careful interrogation in the context of the interactions between clinicopathological variables included in the model in order to provide an accurate interpretation. This then raises the question of whether such interactions could be included in an isBFAC-like model. The barrier to this would be the combinatorial increase in the number of parameters that would then need to be estimated, hence only a limited number of covariates could be used as input.

Furthermore, the prognostic power of the metavariables was likely influenced by the feature selection process, which included univariate survival modelling of each feature as a criterion. If this model were implemented on a computational platform with more power, it could be preferable to instead include all the features from the original data (numbering nearly half a million), and include prognostic data in the model itself. Censored entries could be accounted for using a left-truncated normal distribution extending forward in time from the date of censoring (Ahmad & Fröhlich, 2017). This would allow the metavariables to be composed of features which aren't individually the most prognostically significant, but which together do have power to predict patients' outcomes.

Chapter 3 highlighted the intratumoural transcriptomic heterogeneity of CRC, and previous work has shown that heterogeneity exists in CRC at multiple molecular levels (as discussed in Chapter 1.3). However, this factor was not taken into account as part of this multiomics analysis, because datasets containing high-dimensional, matched molecular data from multiple samples of the same tumour (be those samples from different regions, or even different individual cells) do not yet exist. Once such data does become available, it would be highly interesting to apply this model to understand intratumoural heterogeneity in a multi-omic and clinical context-aware fashion.

Finally, it is not yet clear how the understanding of CRC gleaned from the results from this chapter could be used to personalise therapies in the clinic. In Chapter 2, I showed how it could be possible to subtype patient samples using the nCounter platform, which has a clinically practicable turnaround time. However, this required the profiling of only 38 genes. While only 2,473 features were input into the isBFAC model from the 465,834 available, they come from diverse data types (mutations, copy number, gene and miRNA expression, methylation and protein expression) that would require exome sequencing, RNA sequencing, bisulphite sequencing, and protein arrays to all be performed on the sample. Unless the high-throughput profiling of multiple omics data types becomes affordable and standardised enough to be performed routinely on biopsies and surgical samples, patients' metavariables scores cannot be calculated in a clinical setting. Hence, for the foreseeable future, it could be more prudent to instead take an individual data type from the metavariables and assess if it can act as a surrogate that provide the same or similar information to the multiomics metavariables.

Appendices

A Prognostic power of the iClusterPlus CRC subtypes

I took subtypes from Mo *et al.* (Mo *et al.*, 2013), and plotted RFS of these patients as a function of these subtypes in Figure A.1. This analysis indicated that there was no significant difference in prognosis of these groups of patients ($p = 0.59$).

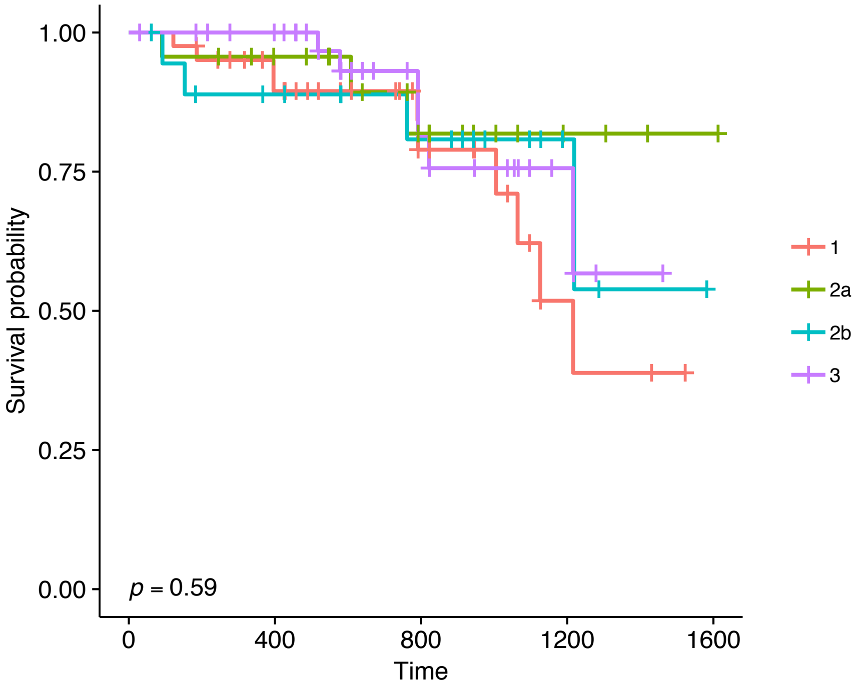


Figure A.1: There is no significant difference in RFS between patients falling in the four iClusterPlus subtypes. Kaplan-Meier survival curves for patients with integrated iClusterPlus subtypes as defined in Mo *et al.* (Mo *et al.*, 2013) (n = 189).

B Initial selection of genes for the nCounter assay

Table B.1: Initial gene selection criteria for 50 genes for the nCounter assay. "qRT-PCR marker" refers to those proposed in Sadanandam *et al.* (Sadanandam *et al.*, 2013) Genes were selected by Dr Anguraj Sadanandam, Team Leader at the Institute of Cancer Research, UK.

50 nCounter genes	Subtype	Note
ACSL6	TA	Top TA gene in CRCA-786
AQP8	Enterocyte	Normal enterocyte marker
AREG	TA	Potential marker of cetuximab response
AXIN2	TA	Top TA gene in CRCA-786; Wnt signalling
BHLHE41	Stem-like	Potential marker of cetuximab resistance
BIRC3	Inflammatory	NF-ϰB signalling
CA1	Enterocyte	Top enterocyte gene in CRCA-786; Normal enterocyte marker
CA4	Enterocyte	Top enterocyte gene in CRCA-786
CEL	TA	Top TA gene in CRCA-786
CFTR	TA	qRT-PCR marker
CLCA4	Enterocyte	Top enterocyte gene in CRCA-786
CLDN8	Enterocyte	Top enterocyte gene in CRCA-786
COL10A1	Stem-like	Top stem-like gene in CRCA-786
CXCL13	Inflammatory	Top inflammatory gene in CRCA-786; Chemokine signalling
CXCL9	Inflammatory	Top inflammatory gene in CRCA-786; Chemokine signalling
CYP1B1	Stem-like	Top stem-like gene in CRCA-786
EREG	TA	Top TA gene in CRCA-786; Potential marker of cetuximab response
FLNA	Stem-like	qRT-PCR marker; Potential marker of cetuximab resistance
GZMA	Inflammatory	Top inflammatory gene in CRCA-786
IDO1	Inflammatory	Top inflammatory gene in CRCA-786

50 nCounter genes	Subtype	Note
IFIT3	Inflammatory	Interferon signalling
KRT20	Enterocyte	Differentiation marker
KRT23	TA	Top TA gene in CRCA-786
LINC00261	NA	Non-coding - not present in CRCA-786 so excluded for CRCA-38
LY6G6D	TA	Top TA gene in CRCA-786
MET	NA	Associated with cetuximab resistance - not present in CRCA-786 so excluded for CRCA-38
MGP	Stem-like	Top stem-like gene in CRCA-786
MS4A12	Enterocyte	Top enterocyte gene in CRCA-786; Normal enterocyte marker
MSRB3	Stem-like	Top stem-like gene in CRCA-786
MUC2	Enterocyte	qRT-PCR marker; Normal goblet cell marker
PCSK1	Goblet-like	Top goblet-like gene in CRCA-786
PLEKHB1	TA	Potential marker of cetuximab resistance
QPRT	TA	Top TA gene in CRCA-786
RARRES3	Inflammatory	qRT-PCR marker; Top inflammatory gene in CRCA-786
REG4	Goblet-like	Top goblet-like gene in CRCA-786
SFRP2	Stem-like	qRT-PCR marker; Top stem-like gene in CRCA-786; Wnt signalling
SFRP4	Stem-like	Top stem-like gene in CRCA-786; Wnt signalling
SLC4A4	Enterocyte	Top enterocyte gene in CRCA-786
SNAI2	Stem-like	EMT marker
SPINK4	Goblet-like	Top goblet-like gene in CRCA-786
STAT1	Inflammatory	Interferon signalling
TAGLN	Stem-like	Myoepithelial marker

50 nCounter genes	Subtype	Note
TCN1	Goblet-like	Top goblet-like gene in CRCA-786
TFF1	Goblet-like	Normal goblet cell marker
TFF3	Goblet-like	qRT-PCR marker
TOX	Goblet-like	Immune differentiation regulator
TWIST1	Stem-like	EMT marker
ZEB1	Stem-like	qRT-PCR marker
ZEB2	Stem-like	EMT marker
ZG16	Enterocyte	Top enterocyte gene in CRCA-786

C Cross tables of sample classifications from different assays

Table C.1: Confusion matrix and statistics showing the concordance in sample classifications between NanoCRCA and CRCA-38.

		NanoCRCA				
		Enterocyte	Goblet-like	Inflammatory	Stem-like	TA
CRCA-38	Enterocyte	11	0	0	0	0
	Goblet-like	1	6	1	0	0
	Inflammatory	1	0	4	1	0
	Stem-like	1	0	0	6	0
	TA	0	0	0	0	6
	Sensitivity	100%	75%	67%	86%	100%
	Specificity	89%	100%	97%	97%	100%
	PPV*	79%	100%	80%	86%	100%
	NPV*	100%	94%	94%	97%	100%
	Balanced accuracy	94%	88%	82%	91%	100%

Table C.2: Confusion matrix and statistics showing the concordance in sample classifications between NanoCRCA and CRCA-786.

		NanoCRCA				
		Enterocyte	Goblet-like	Inflammatory	Stem-like	TA
CRCA-786	Enterocyte	9	0	0	0	0
	Goblet-like	1	6	0	0	0
	Inflammatory	2	0	4	0	0
	Stem-like	2	0	0	6	0
	TA	1	0	0	0	5
	Sensitivity	100%	86%	67%	75%	83%
	Specificity	78%	100%	100%	100%	100%
	PPV*	60%	100%	100%	100%	100%
	NPV*	100%	97%	94%	93%	97%
	Balanced accuracy	89%	93%	83%	88%	92%

Table C.3: Confusion matrix and statistics showing the concordance in sample classifications between CRCA-38 and CRCA-786.

		CRCA-38				
		Enterocyte	Goblet-like	Inflammatory	Stem-like	TA
CRCA-786	Enterocyte	9	0	0	0	0
	Goblet-like	0	6	0	0	0
	Inflammatory	0	1	7	0	0
	Stem-like	1	0	0	8	0
	TA	0	0	0	0	7
	Sensitivity	100%	100%	88%	89%	100%
	Specificity	97%	97%	100%	100%	100%
	PPV*	90%	86%	100%	100%	100%
	NPV*	100%	100%	97%	97%	100%
	Balanced Accuracy	98%	98%	94%	94%	100%

Table C.4: Confusion matrix and statistics showing the concordance in sample classifications between NanoCRCA and CMS.

		CRCA-38			
		Enterocyte/TA	Goblet-like	Inflammatory	Stem-like
CRCA-786	CMS1	2	0	2	0
	CMS2	15	0	0	2
	CMS3	2	3	0	0
	CMS4	2	0	0	4
	Sensitivity	71%	100%	100%	67%
	Specificity	82%	93%	93%	92%
	PPV*	88%	60%	50%	67%
	NPV*	60%	100%	100%	92%
	Balanced Accuracy	77%	97%	97%	79%

*PPV: positive predictive value. NPV: negative predictive value.

D Subtyping of normal colorectal tissue into the CRCAssigner subtypes

To understand how the contamination of tumour samples with normal colorectal tissue could impact on the subtyping of samples into the CRCAssigner subtypes, I classified gene expression profiles from normal and cancerous colorectal tissue samples from TCGA into the five subtypes, using the same procedure as in Chapter 2.2.5 and the 786-gene signature. Rather than median centre the genes based on all the samples – as is the usual approach when all the profiles are from cancerous samples – I instead centred the data using the median expression of the genes in only the tumour samples.

This analysis revealed that all but one of the normal samples fall into the enterocyte subtype, as shown in Table D.1.

Table D.1: Normal colorectal tissue is falls into the enterocyte CRCAssigner subtype. Table showing subtyping of TCGA colorectal normal and tumour tissue samples into the CRCAssigner subtypes.

Subtype	Normal samples	Tumour samples
Enterocyte	50	71
Goblet-like	0	55
Inflammatory	0	67
Stem-like	1	97
TA	0	93

References

Abbas, A. R., Wolslegel, K., Seshasayee, D., Modrusan, Z. & Clark, H. F. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLoS ONE* **4**, e6098 (2009).

Ahmad, A. & Fröhlich, H. Gene expression Towards clinically more relevant dissection of patient heterogeneity via survival-based Bayesian clustering. *Bioinformatics* **33**, 3558–3566 (2017).

Aizawa, T., Karasawa, H., Funayama, R., Shirota, M., Suzuki, T., Maeda, S., Suzuki, H., Yamamura, A., Naitoh, T., Nakayama, K. & Unno, M. Cancer - associated fibroblasts secrete Wnt2 to promote cancer progression in colorectal cancer. *Cancer Medicine* **00**, 1–13 (2019).

Akalin, A. Exploratory data analysis with unsupervised machine learning. *Computational Genomics with R* (2019). Available at: https://compgenomr.github.io/book/.

Akbani, R., Kwok, P., Ng, S., Werner, H. M. J., Shahmoradgoli, M., Zhang, F., Ju, Z., Liu, W., Yang, J.-y., Yoshihara, K., Li, J., Ling, S., Seviour, E. G., Ram, P. T., Minna, J. D., Diao, L., Tong, P., Heymach, J. V., Hill, S. M., Dondelinger, F., Byers, L. A., Meric-bernstam, F., Weinstein, J. N., Broom, B. M., Verhaak, R. G. W., Liang, H., Mukherjee, S., Lu, Y. & Mills, G. B. A pan-cancer proteomic perspective on The Cancer Genome Atlas. *Nature Communications* **5**, 3887 (2014).

Andor, N., Graham, T. A., Jansen, M., Xia, L. C., Aktipis, C. A., Petritsch, C., Ji, H. P. & Maley, C. C. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nature Medicine* **22**, 105–113 (2015).

Andrews, S. FastQC: A quality control tool for high throughput sequence data. *Babraham Bioinformatics* (2016). Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Ayers, M., Lunceford, J., Nebozhyn, M., Murphy, E., Loboda, A., Kaufman, D. R., Albright, A., Cheng, J. D., Kang, S. P., Shankaran, V., Piha-Paul, S. A., Yearley, J., Seiwert, T. Y., Ribas, A. & McClanahan, T. K. IFN-γ-related mRNA profile predicts clinical response to PD-1 blockade. *Journal of Clinical Investigation* **127**, 2930–2940 (2017).

Baldus, S. E., Schaefer, K. L., Engers, R., Hartleb, D., Stoecklein, N. H. & Gabbert, H. E. Prevalence and heterogeneity of KRAS, BRAF, and PIK3CA mutations in primary colorectal adenocarcinomas and their corresponding metastases. *Clinical Cancer Research* **16**, 790–799 (2010).

Barbie, D. A., Tamayo, P., Boehm, J. S., Kim, S. Y., Moody, S. E., Dunn, I. F., Schinzel, A. C., Sandy, P., Meylan, E., Scholl, C., Fröhling, S., Chan, E. M., Sos, M. L., Michel, K., Mermel, C., Silver, S. J., Weir, B. A., Reiling, J. H., Sheng, Q., Gupta, P. B., Wadlow, R. C., Le, H., Hoersch, S., Wittner, B. S., Ramaswamy, S., Livingston, D. M., Sabatini, D. M., Meyerson, M., Thomas, R. K., Lander, E. S., Mesirov, J. P., Root, D. E., Gilliland, D. G., Jacks, T. & Hahn, W. C. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462**, 108–112 (2009).

Bartlett, J. M. S., Bayani, J., Marshall, A., Dunn, J. A., Campbell, A., Cunningham, C., Sobol, M. S., Hall, P. S., Poole, C. J., Cameron, D. A., Earl, H. M., Rea, D. W., Macpherson, I. R., Canney, P., Francis, A., McCabe, C., Pinder, S. E., Hughes-Davies, L., Makris, A. & Stein, R. C. Comparing breast cancer multiparameter tests in the OPTIMA Prelim trial: no test is more equal than the others. *Journal of the National Cancer Institute* **108**, djw050 (2016).

Baudis, M. Genomic imbalances in 5918 malignant epithelial tumors: An explorative meta-analysis of chromosomal CGH data. *BMC Cancer* **7**, 1–15 (2007).

Benassi, B., Fanciulli, M., Fiorentino, F., Porrello, A., Chiorino, G., Loda, M., Zupi, G., Biroccio, A. & Carolina, N. c-Myc Phosphorylation Is Required for Cellular Response to Oxidative Stress. *Molecular Cell* **21**, 509–519 (2006).

Betts, J. G., DeSaix, P., Johnson, E., Johnson, J. E., Korol, O., Kruse, D. H., Poe, B., Wise, J. A. & Young, K. A. The small and large intestines. In *Anatomy & Physiology* (OpenStax CNX, 2016).

Boland, C. R. & Goel, A. Microsatellite instability in colorectal cancer. *Gastroenterology* **138**, 2073–2087 (2010).

Breiman, L. Random forests. *Machine Learning* **45**, 5–32 (2001).

Broad Institute. Picard. *Broad Institute* (2015). Available at: http://broadinstitute.github.io/picard/.

Brunet, J.-P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences* **101**, 4164–4169 (2004).

Budinska, E., Popovici, V., Tejpar, S., D'Ario, G., Lapique, N., Sikora, K. O., Di Narzo, A. F., Yan, P., Graeme Hodgson, J., Weinrich, S., Bosman, F., Roth, A. & Delorenzi, M. Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. *Journal of Pathology* **231**, 63–76 (2013).

Byrd, J. C. & Bresalier, R. S. Mucins and mucin binding proteins in colorectal cancer. *Cancer Metastasis Reviews* **23**, 77–99 (2004).

Cancer Research UK. Bowel cancer mortality statistics. *Cancer Research UK* (2016). Available at: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer/mortality.

Chen, D. S. & Mellman, I. Elements of cancer immunity and the cancer-immune set point. *Nature* **541**, 321–330 (2017).

Cherkassky, V. & Ma, Y. Practical selection of SVM parameters and noise estimation for SVM regression. *Neural Networks* **17**, 113–126 (2004).

Christie, M. & Sieber, O. Pathways of Carcinogenesis. In *ABC of Colorectal Cancer* (eds. Young, A., Hobbs, R. & Kerr, D.) (BMJ Books, 2011).

Cogdill, A. P., Andrews, M. C. & Wargo, J. A. Hallmarks of response to immune checkpoint blockade. *British Journal of Cancer* **117**, 1–7 (2017).

Cortes, C. & Vapnik, V. Support-vector networks. *Machine Learning* **20**, 273–297 (1995).

Cronin, M., Sangli, C., Liu, M. L., Pho, M., Dutta, D., Nguyen, A., Jeong, J., Wu, J., Langone, K. C. & Watson, D. Analytical validation of the Oncotype DX genomic diagnostic test for recurrence prognosis and therapeutic response prediction in node-negative, estrogen receptor-positive breast cancer. *Clinical Chemistry* **53**, 1084–1091 (2007).

Dalerba, P., Kalisky, T., Sahoo, D., Rajendran, P. S., Rothenberg, M. E., Leyrat, A. A., Sim, S., Okamoto, J., Johnston, D. M., Qian, D., Zabala, M., Bueno, J., Neff, N. F., Wang, J., Shelton, A. A., Visser, B., Hisamori, S., Shimono, Y., Wetering, M. van de, Clevers, H., Clarke, M. F. & Quake, S. R. Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nature Biotechnology* **29**, 1120–1127 (2011).

Day, A. heatmap.plus: Heatmap with more sensible behavior. (2012). Available at: https://cran.r-project.org/package=heatmap.plus.

Del Rio, M., Molina, F., Bascoul-Mollevi, C., Copois, V., Bibeau, F. F., Chalbos, P., Bareil, C., Kramar, A., Salvetat, N., Fraslon, C., Conseiller, E., Granci, V., Leblanc, B., Pau, B., Martineau, P. & Ychou, M. Gene expression signature in advanced colorectal cancer patients select drugs and response for the use of leucovorin, fluorouracil, and irinotecan. *Journal of Clinical Oncology* **25**, 773–780 (2007).

De Palma, F. D. E., D'argenio, V., Pol, J., Kroemer, G., Maiuri, M. C. & Salvatore, F. The molecular hallmarks of the serrated pathway in colorectal cancer. *Cancers* **11**, 3–5 (2019).

De Sousa E Melo, F., Wang, X., Jansen, M., Fessler, E., Trinh, A., Rooij, L. P. M. H. de, Jong, J. H. de, Boer, O. J. de, Leersum, R. van, Bijlsma, M. F., Rodermond, H., Heijden, M. van der, Noesel, C. J. M. van, Tuynman, J. B., Dekker, E., Markowetz, F., Medema, J. P. & Vermeulen, L. Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nature Medicine* **19**, 614–618 (2013).

Diergaarde, B., Tiemersma, E. W., Braam, H., Van Muijen, G. N. P., Nagengast, F. M., Kok, F. J. & Kampman, E. Dietary factors and truncating APC mutations in sporadic colorectal adenomas. *International Journal of Cancer* **113**, 126–132 (2005).

Dowsett, M., Sestak, I., Lopez-Knowles, E., Sidhu, K., Dunbier, A. K., Cowens, J. W., Ferree, S., Storhoff, J., Schaper, C. & Cuzick, J. Comparison of PAM50 risk of recurrence score with Oncotype DX and IHC4 for predicting risk of distant recurrence after endocrine therapy. *Journal of Clinical Oncology* **31**, 2783–2790 (2013).

Dudoit, S., Fridlyand, J. & Speed, T. P. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* **97**, 77–87 (2002).

Duenwald, S., Zhou, M., Wang, Y., Lejnine, S., Kulkarni, A., Graves, J., Smith, R., Castle, J., Tokiwa, G., Fine, B., Dai, H., Fare, T. & Marton, M. Development of a microarray platform for FF-PET profiling: Application to the classification of human tumors. *Journal of Translational Medicine* **7**, 65 (2009).

Dunne, P. D., McArt, D. G., Bradley, C. A., O'Reilly, P. G., Barrett, H. L., Cummins, R., O'Grady, T., Arthur, K., Loughrey, M., Allen, W. L., McDade, S., Waugh, D. J., Hamilton, P. W., Longley, D. B., Kay, E. W., Johnston, P. G., Lawler, M., Salto-Tellez, M. & Van Schaeybroeck, S. Challenging the cancer molecular stratification dogma: Intratumoral heterogeneity undermines consensus molecular subtypes and potential diagnostic value in colorectal cancer. *Clinical Cancer Research* **22**, 4095–4104 (2016).

Eason, K., Nyamundanda, G. & Sadanandam, A. polyClustR : defining communities of reconciled cancer subtypes with biological and prognostic significance. *BMC Bioinformatics* **19**, 182 (2018).

Engelhardt, B. E. & Stephens, M. Analysis of population structure: A unifying framework and novel methods based on sparse factor analysis. *PLoS Genetics* **6**, e1001117 (2010).

Etienne, W., Meyer, M. H., Peppers, J. & Meyer, R. A. Comparison of mRNA gene expression by RT-PCR and DNA microarray. *BioTechniques* **36**, 618–626 (2004).

Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–767 (1990).

Fedorowicz, G., Guerrero, S., Wu, T. D. & Modrusan, Z. Microarray analysis of RNA extracted from formalin-fixed, paraffin-embedded and matched fresh-frozen ovarian adenocarcinomas. *BMC Medical Genomics* **2**, 1–11 (2009).

Finotello, F. & Trajanoski, Z. Quantifying tumor-infiltrating immune cells from transcriptomics data. *Cancer Immunology, Immunotherapy* **67**, 1031–1040 (2018).

Fontana, E., Nyamundanda, G., Cunningham, D., Ragulan, C., Sclafani, F., Eason, K., Bali, M., Vendrell, I., Patil, Y., Wilson, S., Moorcraft, J., Begum, R., Chau, I., Starling, N. & Sadanandam, A. Molecular subtype assay to reveal anti-EGFR response sub-clones in colorectal cancer (CRC). In *Journal of Clinical Oncology* **36**, 658 (2018).

Frank, M., Doring, C., Metzler, D., Eckerle, S. & Hansmann, M.-L. Global gene expression profiling of formalin-fixed paraffin-embedded tumor samples: a comparison to snap-frozen material using oligonucleotide microarrays. *Virchows Archiv* **450**, 699–711 (2007).

Fricke, A., Ullrich, P. V., Heinz, J., Pfeifer, D., Scholber, J., Herget, G. W., Hauschild, O., Bronsert, P., Stark, G. B., Bannasch, H., Eisenhardt, S. U. & Braig, D. Identification of a blood-borne miRNA signature of synovial sarcoma. *Molecular Cancer* **14**, 1–13 (2015).

Fu, L., Zhang, C., Zhang, L.-Y., Dong, S.-S., Lu, L.-H., Chen, J., Dai, Y., Li, Y., Kong, K. L., Kwong, D. L. & Guan, X.-Y. Wnt2 secreted by tumour fibroblasts promotes tumour progression in oesophageal cancer by activation of the Wnt/β-catenin signalling pathway. *Gut* **60**, 1635 LP–1643 (2011).

Gao, H., Korn, J. M., Ferretti, S., Monahan, J. E., Wang, Y., Singh, M., Zhang, C., Schnell, C., Yang, G., Zhang, Y., Balbin, O. A., Barbe, S., Cai, H., Casey, F., Chatterjee, S., Chiang, D. Y., Chuai, S., Cogan, S. M., Collins, S. D., Dammassa, E., Ebel, N., Embry, M., Green, J., Kauffmann, A., Kowal, C., Leary, R. J., Lehar, J., Liang, Y., Loo, A., Lorenzana, E., Robert McDonald, E., McLaughlin, M. E., Merkin, J., Meyer, R., Naylor, T. L., Patawaran, M., Reddy, A., Röelli, C., Ruddy, D. A., Salangsang, F., Santacroce, F., Singh, A. P., Tang, Y., Tinetto, W., Tobler, S., Velazquez, R., Venkatesan, K., Von Arx, F., Wang, H. Q., Wang, Z., Wiesmann, M., Wyss, D., Xu, F., Bitter, H., Atadja, P., Lees, E., Hofmann, F., Li, E., Keen, N., Cozens, R., Jensen, M. R., Pryer, N. K., Williams, J. A. & Sellers, W. R. High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nature Medicine* **21**, 1318–1325 (2015).

Geiss, G. K., Bumgarner, R. E., Birditt, B., Dahl, T., Dowidar, N., Dunaway, D. L., Fell, H. P., Ferree, S., George, R. D., Grogan, T., James, J. J., Maysuria, M., Mitton, J. D., Oliveri, P., Osborn, J. L., Peng, T., Ratcliffe, A. L., Webster, P. J., Davidson, E. H., Hood, L. & Dimitrov, K. Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nature Biotechnology* **26**, 317–25 (2008).

Goldman, M., Craft, B., Kamath, A., Brooks, A. N., Zhu, J. & Haussler, D. The UCSC Xena Platform for cancer genomics data visualization and interpretation. *bioRxiv* (2018). doi:10.1101/326470

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. & Lander, E. S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).

Gong, T., Hartmann, N., Kohane, I. S., Brinkmann, V., Staedtler, F., Letzkus, M., Bongiovanni, S. & Szustakowski, J. D. Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. *PLoS ONE* **6**, e27156 (2011).

Gong, T. & Szustakowski, J. D. DeconRNASeq: A statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics* **29**, 1083–1085 (2013).

Gorbach, S. L. Microbiology of the gastrointestinal tract. In *Medical Microbiology* (ed. Baron, S.) (University of Texas Medical Branch at Galveston, 1996).

Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).

Guénin, S., Mauriat, M., Pelloux, J., Van Wuytswinkel, O., Bellini, C. & Gutierrez, L. Normalization of qRT-PCR data: The necessity of adopting a systematic, experimental conditions-specific, validation of references. *Journal of Experimental Botany* **60**, 487–493 (2009).

Guinney, J., Dienstmann, R., Wang, X., Reyniès, A. de, Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P., Bot, B. M., Morris, J. S., Simon, I. M., Gerster, S., Fessler, E., De Sousa E Melo, F., Missiaglia, E., Ramay, H., Barras, D., Homicsko, K., Maru, D., Manyam, G. C., Broom, B., Boige, V., Perez-Villamil, B., Laderas, T., Salazar, R., Gray, J. W., Hanahan, D., Tabernero, J., Bernards, R., Friend, S. H., Laurent-Puig, P., Medema, J. P., Sadanandam, A., Wessels, L., Delorenzi, M., Kopetz, S., Vermeulen, L. & Tejpar, S. The consensus molecular subtypes of colorectal cancer. *Nature Medicine* **21**, 1350–1356 (2015).

Hoadley, K. a., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., Ng, S., Leiserson, M. D. M., Niu, B., McLellan, M. D., Uzunangelov, V., Zhang, J., Kandoth, C., Akbani, R., Shen, H., Omberg, L., Chu, A., Margolin, A. a., Veer, L. J. van't, Lopez-Bigas, N., Laird, P. W., Raphael, B. J., Ding, L., Robertson, a. G., Byers, L. a., Mills, G. B., Weinstein, J. N., Van Waes, C., Chen, Z., Collisson, E. a., Benz, C. C., Perou, C. M. & Stuart, J. M. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929–944 (2014).

Hoerl, A. E. & Kennard, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **12**, 55–67 (1970).

Ihle, J. N. & Gilliland, D. G. Jak2 : normal function and role in hematopoietic disorders. *Current Opinion in Genetics & Development 2007,* **17**, 8–14 (2007).

Iles, R. K. & Butler, S. A. Cellular Pathology Part II: Clinical application and laboratory techniques. In *Biomedical Sciences: Essential Laboratory Medicine* (eds. Iles, R. K. & Docherty, S.) 175–203 (John Wiley & Sons, Incorporated, 2012).

Isella, C., Brundu, F., Bellomo, S. E., Galimi, F., Zanella, E., Porporato, R., Petti, C., Fiori, A., Orzan, F., Senetta, R., Boccaccio, C., Ficarra, E., Marchionni, L., Trusolino, L., Medico, E. & Bertotti, A. Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer. *Nature Communications* **8**, 15107 (2017).

Isella, C., Terrasi, A., Bellomo, S. E., Petti, C., Galatola, G., Muratore, A., Mellano, A., Senetta, R., Cassenti, A., Sonetto, C., Inghirami, G., Trusolino, L., Fekete, Z., De Ridder, M., Cassoni, P., Storme, G., Bertotti, A. & Medico, E. Stromal contribution to the colorectal cancer transcriptome. *Nature Genetics* **47**, 312–319 (2015).

Jorissen, R. N., Gibbs, P., Christie, M., Prakash, S., Lipton, L., Desai, J., Kerr, D., Aaltonen, L. A., Arango, D., Kruhøffer, M., Ørntoft, T. F., Andersen, C. L., Gruidl, M., Kamath, V. P., Eschrich, S., Yeatman, T. J. & Sieber, O. M. Metastasis-associated gene expression changes predict poor outcomes in patients with Dukes stage B and C colorectal cancer. *Clinical Cancer Research* **15**, 7642–7651 (2009).

Joung, J. G., Oh, B. Y., Hong, H. K., Al-Khalidi, H., Al-Alem, F., Lee, H. O., Bae, J. S., Kim, J., Cha, H. U., Alotaibi, M., Cho, Y. B., Hassanain, M., Park, W. Y. & Lee, W. Y. Tumor heterogeneity predicts metastatic potential in colorectal cancer. *Clinical Cancer Research* **23**, 7209–7216 (2017).

Juárez, M., Egoavil, C., Rodríguez-Soler, M., Hernández-Illán, E., Guarinos, C., García-Martínez, A., Alenda, C., Giner-Calabuig, M., Murcia, O., Mangas, C., Payá, A., Aparicio, J. R., Ruiz, F. A., Martínez, J., Casellas, J. A., Soto, J. L., Zapater, P. & Jover, R. KRAS and BRAF somatic mutations in colonic polyps and the risk of metachronous neoplasia. *PLoS ONE* **12**, e0184937 (2017).

Kawakami, H., Zaanan, A. & Sinicrope, F. A. Microsatellite instability testing and its role in the management of colorectal cancer. *Current Treatment Options in Oncology* **16**, (2015).

Kelley, R. K. & Venook, A. P. Prognostic and Predictive Markers in Stage II Colon Cancer : Is There a Role for Gene Expression Profiling? *Clinical Colorectal Cancer* **10**, 73–80 (2011).

Kerr, D., Gray, R., Quirke, P., Watson, D., Yothers, G., Lavery, I. C., Lee, M., O'Connell, M. J., Shak, S. & Wolmark, N. A quantitative multigene RT-PCR assay for prediction of recurrence in stage II colon cancer: Selection of the genes in four large studies and results of the independent, prospectively designed QUASAR validation study. *Journal of Clinical Oncology* **27**, 4000 (2009).

Khambata-Ford, S., Garrett, C. R., Meropol, N. J., Basik, M., Harbison, C. T., Wu, S., Wong, T. W., Huang, X., Takimoto, C. H., Godwin, A. K., Tan, B. R., Krishnamurthi, S. S., Burris, H. A., Poplin, E. A., Hidalgo, M., Baselga, J., Clark, E. A. & Mauro, D. J. Expression of epiregulin and amphiregulin and K-ras mutation status predict disease control in metastatic colorectal cancer patients treated with cetuximab. *Journal of Clinical Oncology* **25**, 3230–3237 (2007).

Kim, R., Sarker, D., Macarulla, T., Yau, T., Choo, S. P., Meyer, T., Hollebecque, A., Whisenant, J., Sung, M., Yoon, J.-H., Lim, H. Y., Zhu, A., Park, J.-W., Faivre, S., Mazzaferro, V., Shi, H., Schmidt-Kittler, O., Clifford, C., Wolf, B. & Kang, Y.-K. Efficacy of pembrolizumab in phase 2 KEYNOTE-164 and KEYNOTE-158 studies of microsatellite instability high cancers. *Annals of Oncology* **28**, 128–129 (2017).

Kim, T. M., Jung, S. H., An, C. H., Lee, S. H., Baek, I. P., Kim, M. S., Park, S. W., Rhee, J. K., Lee, S. H. & Chung, Y. J. Subclonal genomic architectures of primary and metastatic colorectal cancer based on intratumoral genetic heterogeneity. *Clinical Cancer Research* **21**, 4461–4472 (2015).

Kiselev, V. Y., Andrews, T. S. & Hemberg, M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics* **20**, 273–282 (2019).

Klopfleisch, R., Weiss, A. T. A. & Gruber, A. D. Excavation of a buried treasure - DNA, mRNA, miRNA and protein analysis in formalin fixed, paraffin embedded tissues. *Histology and Histopathology* **26**, 797–810 (2011).

Kolbert, C. P., Feddersen, R. M., Rakhshan, F., Grill, D. E., Simon, G., Middha, S., Jang, J. S., Simon, V., Schultz, D. A., Zschunke, M., Lingle, W., Carr, J. M., Thompson, E. A., Oberg, A. L., Eckloff, B. W., Wieben, E. D., Li, P., Yang, P. & Jen, J. Multi-platform analysis of microRNA expression measurements in RNA from fresh frozen and FFPE Tissues. *PLoS ONE* **8**, (2013).

Kong, H., Zhu, M., Cui, F., Wang, S., Gao, X., Lu, S., Wu, Y. & Zhu, H. Quantitative assessment of short amplicons in FFPE-derived long-chain RNA. *Scientific Reports* **4**, 7246 (2014).

Le, D. T., Durham, J. N., Smith, K. N., Wang, H., Bartlett, B. R., Aulakh, L. K., Lu, S., Kemberling, H., Wilt, C., Luber, B. S., Wong, F., Azad, N. S., Rucki, A. A., Laheru, D., Donehower, R., Zaheer, A., Fisher, G. A., Crocenzi, T. S., Lee, J. J., Greten, T. F., Duffy, A. G., Ciombor, K. K., Eyring, A. D., Lam, B. H., Joe, A., Kang, S. P., Holdhoff, M., Danilova, L., Cope, L., Meyer, C., Zhou, S., Goldberg, R. M., Armstrong, D. K., Bever, K. M., Fader, A. N., Taube, J., Housseau, F., Spetzler, D., Xiao, N., Pardoll, D. M., Papadopoulos, N., Kinzler, K. W., Eshleman, J. R., Vogelstein, B., Anders, R. A. & Diaz, L. A. Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science* **357**, 409–413 (2017).

Le, D. T., Uram, J. N., Wang, H., Bartlett, B., Kemberling, H., Eyring, A., Skora, A., Azad, N. S., Laheru, D. A., Donehower, R. C., Luber, B., Crocenzi, T. S., Fisher, G. A., Duffy, S. M., Lee, J. J., Koshiji, M., Eshleman, J. R., Anders, R. A., Vogelstein, B. & Diaz, L. A. PD-1 blockade in tumors with mismatch repair deficiency. *Journal of Clinical Oncology* **372**, 2509–2520 (2015).

Lenz, H.-J., Ou, F.-S., Venook, A. P., Hochster, H. S., Niedzwiecki, D., Goldberg, R. M., Mayer, R. J., Bertagnolli, M. M., Blanke, C. D., Zemla, T., Qu, X., Innocenti, F. & Kabbarah, O. Impact of consensus molecular subtyping (CMS) on overall survival (OS) and progression free survival (PFS) in patients (pts) with metastatic colorectal cancer (mCRC): Analysis of CALGB/SWOG 80405 (Alliance). *Journal of Clinical Oncology* **35**, 3511–3511 (2018).

Leslie, D. S., Dascher, C. C., Cembrola, K., Townes, M. A., Hava, D. L., Hugendubler, L. C., Mueller, E., Fox, L., Roura-Mir, C., Moody, D. B., Vincent, M. S., Gumperz, J. E., Illarionov, P. A., Besra, G. S., Reynolds, C. G. & Brenner, M. B. Serum lipids regulate dendritic cell CD1 expression and function. *Immunology* **125**, 289–301 (2008).

Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

Li, B., Severson, E., Pignon, J. C., Zhao, H., Li, T., Novak, J., Jiang, P., Shen, H., Aster, J. C., Rodig, S., Signoretti, S., Liu, J. S. & Liu, X. S. Comprehensive analyses of tumor immunity: Implications for cancer immunotherapy. *Genome Biology* **17**, 174 (2016).

Li, H., Courtois, E. T., Sengupta, D., Tan, Y., Chen, K. H., Goh, J. J. L., Kong, S. L., Chua, C., Hon, L. K., Tan, W. S., Wong, M., Choi, P. J., Wee, L. J. K., Hillmer, A. M., Tan, I. B., Robson, P. & Prabhakar, S. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nature Genetics* **49**, 708–718 (2017a).

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

Li, H., Zhang, H., Chen, Y., Liu, X. & Qian, J. MiR-423-3p Enhances Cell Growth Through Inhibition of p21Cip1/Waf1 in Colorectal Cancer. *Cellular Physiology and Biochemistry* **37**, 1044–1054 (2015).

Li, J., Akbani, R., Zhao, W., Lu, Y., Weinstein, J. N., Mills, G. B. & Liang, H. Explore, visualize, and analyze functional cancer proteomic data using The Cancer Proteome Atlas. *Cancer Research* **77**, e51–e54 (2017b).

Li, J., Lu, Y., Akbani, R., Ju, Z., Roebuck, P. L., Liu, W., Yang, J. Y., Broom, B. M., Verhaak, R. G. W., Kane, D. W., Wakefield, C., Weinstein, J. N., Mills, G. B. & Liang, H. TCPA: A resource for cancer functional proteomics data. *Nature Methods* **10**, 1046–1047 (2013).

Liu, Y., Sethi, N. S., Hinoue, T., Schneider, B. G., Cherniack, A. D., Sanchez-Vega, F., Seoane, J. A., Farshidfar, F., Bowlby, R., Islam, M., Kim, J., Chatila, W., Akbani, R., Kanchi, R. S., Rabkin, C. S., Willis, J. E., Wang, K. K., McCall, S. J., Mishra, L., Ojesina, A. I., Bullman, S., Pedamallu, C. S., Lazar, A. J., Sakai, R., Caesar-Johnson, S. J., Demchok, J. A., Felau, I., Kasapi, M., Ferguson, M. L., Hutter, C. M., Sofia, H. J., Tarnuzzer, R., Wang, Z., Yang, L., Zenklusen, J. C., Zhang, J. (., Chudamani, S., Liu, J., Lolla, L., Naresh, R., Pihl, T., Sun, Q., Wan, Y., Wu, Y., Cho, J., DeFreitas, T., Frazer, S., Gehlenborg, N., Getz, G., Heiman, D. I., Kim, J., Lawrence, M. S., Lin, P., Meier, S., Noble, M. S., Saksena, G., Voet, D., Zhang, H., Bernard, B., Chambwe, N., Dhankani, V., Knijnenburg, T., Kramer, R., Leinonen, K., Liu, Y., Miller, M., Reynolds, S., Shmulevich, I., Thorsson, V., Zhang, W., Akbani, R., Broom, B. M., Hegde, A. M., Ju, Z., Kanchi, R. S., Korkut, A., Li, J., Liang, H., Ling, S., Liu, W., Lu, Y., Mills, G. B., Ng, K. S., Rao, A., Ryan, M., Wang, J., Weinstein, J. N., Zhang, J., Abeshouse, A., Armenia, J., Chakravarty, D., Chatila, W. K., Bruijn, I., Gao, J., Gross, B. E., Heins, Z. J., Kundra, R., La, K., Ladanyi, M., Luna, A., Nissan, M. G., Ochoa, A., Phillips, S. M., Reznik, E., Sanchez-Vega, F., Sander, C., Schultz, N., Sheridan, R., Sumer, S. O., Sun, Y., Taylor, B. S., Wang, J., Zhang, H., Anur, P., Peto, M., Spellman, P., Benz, C., Stuart, J. M., Wong, C. K., Yau, C., Hayes, D. N., Parker, J. S., Wilkerson, M. D., Ally, A., Balasundaram, M., Bowlby, R., Brooks, D., Carlsen, R., Chuah, E., Dhalla, N., Holt, R., Jones, S. J. M., Kasaian, K., Lee, D., Ma, Y., Marra, M. A., Mayo, M., Moore, R. A., Mungall, A. J., Mungall, K., Robertson, A. G., Sadeghi, S., Schein, J. E., Sipahimalani, P., Tam, A., Thiessen, N., Tse, K., Wong, T., Berger, A. C., Beroukhim, R., Cherniack, A. D., Cibulskis, C., Gabriel, S. B., Gao, G. F., Ha, G., Meyerson, M., Schumacher, S. E., Shih, J., Kucherlapati, M. H., Kucherlapati, R. S., Baylin, S., Cope, L., Danilova, L., Bootwalla, M. S., Lai, P. H., Maglinte, D. T., Van Den Berg, D. J., Weisenberger, D. J., Auman, J. T., Balu, S., Bodenheimer, T., Fan, C., Hoadley, K. A., Hoyle, A. P., Jefferys, S. R., Jones, C. D., Meng, S., Mieczkowski, P. A., Mose, L. E., Perou, A. H., Perou, C. M., Roach, J., Shi, Y., Simons, J. V., Skelly, T., Soloway, M. G., Tan, D., Veluvolu, U., Fan, H., Hinoue, T., Laird, P. W., Shen, H., Zhou, W., Bellair, M., Chang, K., Covington, K., Creighton, C. J., Dinh, H., Doddapaneni, H. V., Donehower, L. A., Drummond, J., Gibbs, R. A., Glenn, R., Hale, W., Han, Y., Hu, J., Korchina, V., Lee, S., Lewis, L., Li, W., Liu, X., Morgan, M., Morton, D., Muzny, D., Santibanez, J., Sheth, M., Shinbrot, E., Wang, L., Wang, M., Wheeler, D. A., Xi, L., Zhao, F., Hess, J., Appelbaum, E. L., Bailey, M., Cordes, M. G., Ding, L., Fronick, C. C., Fulton, L. A., Fulton, R. S., Kandoth, C., Mardis, E. R., McLellan, M. D., Miller, C. A., Schmidt, H. K., Wilson, R. K., Crain, D., Curley, E., Gardner, J., Lau, K., Mallery, D., Morris, S., Paulauskis, J., Penny, R., Shelton, C., Shelton, T., Sherman, M., Thompson, E., Yena, P., Bowen, J., Gastier-Foster, J. M., Gerken, M., Leraas, K. M., Lichtenberg, T. M., Ramirez, N. C., Wise, L., Zmuda, E., Corcoran, N., Costello, T., Hovens, C., Carvalho, A. L., Carvalho, A. C. de, Fregnani, J. H., Longatto-Filho, A., Reis, R. M., Scapulatempo-Neto, C., Silveira, H. C. S., Vidal, D. O., Burnette, A., Eschbacher, J., Hermes, B., Noss, A., Singh, R., Anderson, M. L., Castro, P. D., Ittmann, M., Huntsman, D., Kohl, B., Le, X., Thorp, R., Andry, C., Duffy, E. R., Lyadov, V., Paklina, O., Setdikova, G., Shabunin, A., Tavobilov, M., McPherson, C., Warnick, R., Berkowitz, R., Cramer, D., Feltmate, C., Horowitz, N., Kibel, A., Muto, M., Raut, C. P., Malykh, A., Barnholtz-Sloan, J. S., Barrett, W., Devine, K., Fulop, J., Ostrom, Q. T., Shimmel, K., Wolinsky, Y., Sloan, A. E., De Rose, A., Giuliante, F., Goodman, M., Karlan, B. Y., Hagedorn, C. H., Eckman, J., Harr, J., Myers, J., Tucker, K., Zach, L. A., Deyarmin, B., Hu, H., Kvecher, L., Larson, C., Mural, R. J., Somiari, S., Vicha, A., Zelinka, T., Bennett, J., Iacocca, M., Rabeno, B., Swanson, P., Latour, M., Lacombe, L., Têtu, B., Bergeron, A., McGraw, M., Staugaitis, S. M., Chabot, J., Hibshoosh, H., Sepulveda, A., Su, T., Wang, T., Potapova, O., Voronina, O., Desjardins, L., Mariani, O., Roman-Roman, S., Sastre, X., Stern, M. H., Cheng, F., Signoretti, S., Berchuck, A., Bigner, D., Lipp, E., Marks, J., McCall, S., McLendon, R., Secord, A., Sharp, A., Behera, M., Brat, D. J., Chen, A., Delman, K., Force, S., Khuri, F., Magliocca, K., Maithel, S., Olson, J. J., Owonikoko, T., Pickens, A., Ramalingam, S., Shin, D. M., Sica, G., Van Meir, E. G., Zhang, H., Eijckenboom, W., Gillis, A., Korpershoek, E., Looijenga, L., Oosterhuis, W., Stoop, H., Kessel, K. E. van, Zwarthoff, E. C., Calatozzolo, C., Cuppini, L., Cuzzubbo, S., DiMeco, F., Finocchiaro, G., Mattei, L., Perin, A., Pollo, B., Chen, C., Houck, J., Lohavanichbutr, P., Hartmann, A., Stoehr, C., Stoehr, R., Taubert, H., Wach, S., Wullich, B., Kycler, W., Murawa, D., Wiznerowicz, M., Chung, K., Edenfield, W. J., Martin, J., Baudin, E., Bubley, G., Bueno, R., De Rienzo, A., Richards, W. G., Kalkanis, S., Mikkelsen, T., Noushmehr, H., Scarpace, L., Girard, N., Aymerich, M., Campo, E., Giné, E., Guillermo, A. L., Van Bang, N., Hanh, P. T., Phu, B. D., Tang, Y., Colman, H., Evason, K., Dottino, P. R., Martignetti, J. A., Gabra, H., Juhl, H., Akeredolu, T., Stepa, S., Hoon, D., Ahn, K., Kang, K. J., Beuschlein, F., Breggia, A., Birrer, M., Bell, D., Borad, M., Bryce, A. H., Castle, E., Chandan, V., Cheville, J., Copland, J. A., Farnell, M., Flotte, T., Giama, N., Ho, T., Kendrick, M., Kocher, J. P., Kopp, K., Moser, C., Nagorney, D., O'Brien, D., O'Neill, B. P., Patel, T., Petersen, G., Que, F., Rivera, M., Roberts, L., Smallridge, R., Smyrk, T., Stanton, M., Thompson, R. H., Torbenson, M., Yang, J. D., Zhang, L., Brimo, F., Ajani, J. A., Gonzalez, A. M. A., Behrens, C., Bondaruk, J., Broaddus, R., Czerniak, B., Esmaeli, B., Fujimoto, J., Gershenwald, J., Guo, C., Logothetis, C., Meric-Bernstam, F., Moran, C., Ramondetta, L., Rice, D., Sood, A., Tamboli, P., Thompson, T., Troncoso, P., Tsao, A., Wistuba, I., Carter, C., Haydu, L., Hersey, P., Jakrot, V., Kakavand, H., Kefford, R., Lee, K., Long, G., Mann, G., Quinn, M., Saw, R., Scolyer, R., Shannon, K., Spillane, A., Stretch, J., Synott, M., Thompson, J., Wilmott, J., Al-Ahmadie, H., Chan, T. A., Ghossein, R., Gopalan, A., Levine, D. A., Reuter, V., Singer, S., Singh, B., Tien, N. V., Broudy, T., Mirsaidi, C., Nair, P., Drwiega, P., Miller, J., Smith, J., Zaren, H., Park, J. W., Hung, N. P., Kebebew, E., Linehan, W. M., Metwalli, A. R., Pacak, K., Pinto, P. A., Schiffman, M., Schmidt, L. S., Vocke, C. D., Wentzensen, N., Worrell, R., Yang, H., Moncrieff, M., Goparaju, C., Melamed, J., Pass, H., Botnariuc, N., Caraman, I., Cernat, M., Chemencedji, I., Clipca, A., Doruc, S., Gorincioi, G., Mura, S., Pirtac, M., Stancul, I., Tcaciuc, D., Albert, M., Alexopoulou, I., Arnaout, A., Bartlett, J., Engel, J., Gilbert, S., Parfitt, J., Sekhon, H., Thomas, G., Rassl, D. M., Rintoul, R. C., Bifulco, C., Tamakawa, R., Urba, W., Hayward, N., Timmers, H., Antenucci, A., Facciolo, F., Grazi, G., Marino, M., Merola, R., Krijger, R. de, Gimenez-Roqueplo, A. P., Piché, A., Chevalier, S., McKercher, G., Birsoy, K., Barnett, G., Brewer, C., Farver, C., Naska, T., Pennell, N. A., Raymond, D., Schilero, C., Smolenski, K., Williams, F., Morrison, C., Borgia, J. A., Liptay, M. J., Pool, M., Seder, C. W., Junker, K., Omberg, L., Dinkin, M., Manikhas, G., Alvaro, D., Bragazzi, M. C., Cardinale, V., Carpino, G., Gaudio, E., Chesla, D., Cottingham, S., Dubina, M., Moiseenko, F., Dhanasekaran, R., Becker, K. F., Janssen, K. P., Slotta-Huspenina, J., Abdel-Rahman, M. H., Aziz, D., Bell, S., Cebulla, C. M., Davis, A., Duell, R., Elder, J. B., Hilty, J., Kumar, B., Lang, J., Lehman, N. L., Mandt, R., Nguyen, P., Pilarski, R., Rai, K., Schoenfield, L., Senecal, K., Wakely, P., Hansen, P., Lechan, R., Powers, J., Tischler, A., Grizzle, W. E., Sexton, K. C., Kastl, A., Henderson, J., Porten, S., Waldmann, J., Fassnacht, M., Asa, S. L., Schadendorf, D., Couce, M., Graefen, M., Huland, H., Sauter, G., Schlomm, T., Simon, R., Tennstedt, P., Olabode, O., Nelson, M., Bathe, O., Carroll, P. R., Chan, J. M., Disaia, P., Glenn, P., Kelley, R. K., Landen, C. N., Phillips, J., Prados, M., Simko, J., Smith-McCune, K., VandenBerg, S., Roggin, K., Fehrenbach, A., Kendler, A., Sifri, S., Steele, R., Jimeno, A., Carey, F., Forgie, I., Mannelli, M., Carney, M., Hernandez, B., Campos, B., Herold-Mende, C., Jungk, C., Unterberg, A., Deimling, A. von, Bossler, A., Galbraith, J., Jacobus, L., Knudson, M., Knutson, T., Ma, D., Milhem, M., Sigmund, R., Godwin, A. K., Madan, R., Rosenthal, H. G., Adebamowo, C., Adebamowo, S. N., Boussioutas, A., Beer, D., Giordano, T., Mes-Masson, A. M., Saad, F., Bocklage, T., Landrum, L., Mannel, R., Moore, K., Moxley, K., Postier, R., Walker, J., Zuna, R., Feldman, M., Valdivieso, F., Dhir, R., Luketich, J., Pinero, E. M., Quintero-Aguilo, M., Carlotti, C. G., Dos Santos, J. S., Kemp, R., Sankarankuty, A., Tirapelli, D., Catto, J., Agnew, K., Swisher, E., Creaney, J., Robinson, B., Shelley, C. S., Godwin, E. M., Kendall, S., Shipman, C., Bradford, C., Carey, T., Haddad, A., Moyer, J., Peterson, L., Prince, M., Rozek, L., Wolf, G., Bowman, R., Fong, K. M., Yang, I., Korst, R., Rathmell, W. K., Fantacone-Campbell, J. L., Hooke, J. A., Kovatich, A. J., Shriver, C. D., DiPersio, J., Drake, B., Govindan, R., Heath, S., Ley, T., Van Tine, B., Westervelt, P., Rubin, M. A., Lee, J. I., Aredes, N. D., Mariamidze, A., Thorsson, V., Bass, A. J. & Laird, P. W. Comparative Molecular Analysis of Gastrointestinal Adenocarcinomas. *Cancer Cell* **33**, 721–735 (2018).

Losi, L., Baisse, B., Bouzourene, H. & Benhattar, J. Evolution of intratumoral genetic heterogeneity during colorectal cancer progression. *Carcinogenesis* **26**, 916–922 (2005).

Lu, X. & Lu, J. The significance of detection of serum miR-423-5p and miR-484 for diagnosis of colorectal cancer. *Clinical Laboratory* **61**, 187–190 (2015).

Lubbers, R., Essen, M. F. van, Kookten, C. van & Trouw, L. A. Production of complement components by cells of the immune system. *The Journal of Translational Immunology* **188**, 183–194

(2017).

Maldonado, R. A. & Andrian, U. H. von. How tolerogenic dendritic cells induce regulatory T cells. *Advances in Immunology* **108**, 111–165 (2010).

Marisa, L., Reyniès, A. de, Duval, A., Selves, J., Gaub, M. P., Vescovo, L., Etienne-Grimaldi, M. C., Schiappa, R., Guenot, D., Ayadi, M., Kirzin, S., Chazal, M., Fléjou, J. F., Benchimol, D., Berger, A., Lagarde, A., Pencreach, E., Piard, F., Elias, D., Parc, Y., Olschwang, S., Milano, G., Laurent-Puig, P. & Boige, V. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS Medicine* **10**, (2013).

Martins, M., Mansinho, A., Cruz-Duarte, R., Martins, S. L. & Costa, L. Anti-EGFR therapy to treat metastatic colorectal cancer: Not for all. *Advances in Experimental Medicine and Biology* **1110**, 113–131 (2018).

Masuda, N., Ohnishi, T., Kawamoto, S., Monden, M. & Okubo, K. Analysis of chemical modification of RNA from formalin-fixed samples and optimization of molecular biology applications for such samples. *Nucleic Acids Research* **27**, 4436–4443 (1999).

Matsumura, T., Sugimachi, K., Iinuma, H., Takahashi, Y., Kurashige, J., Sawada, G., Ueda, M., Uchi, R., Ueo, H., Takano, Y., Shinden, Y., Eguchi, H., Yamamoto, H., Doki, Y., Mori, M., Ochiya, T. & Mimori, K. Exosomal microRNA in serum is a novel biomarker of recurrence in human colorectal cancer. *British Journal of Cancer* **113**, 275 (2015).

Medico, E., Russo, M., Picco, G., Cancelliere, C., Valtorta, E., Corti, G., Buscarino, M., Isella, C., Lamba, S., Martinoglio, B., Veronese, S., Siena, S., Sartore-Bianchi, A., Beccuti, M., Mottolese, M., Linnebacher, M., Cordero, F., Di Nicolantonio, F. & Bardelli, A. The molecular landscape of colorectal cancer cell lines unveils clinically actionable kinase targets. *Nature Communications* **6**, 7002 (2015).

Mei, Q., Xue, G., Li, X., Wu, Z., Li, X., Yan, H., Guo, M., Sun, S. & Han, W. Methylation-induced loss of miR-484 in microsatellite-unstable colorectal cancer promotes both viability and IL-8 production via CD137L. *The Journal of Pathology* **236**, 165–174 (2015).

Miller, A. B., Hoogstraten, B., Staquet, M. & Winkler, A. Reporting results of cancer treatment. *Cancer* **47**, 207–214 (1981).

Mitra, A. K., Mukherjee, U. K., Harding, T., Jang, J. S., Stessman, H., Li, Y., Abyzov, A., Jen, J., Kumar, S., Rajkumar, V. & Van Ness, B. Single-cell analysis of targeted transcriptome predicts drug sensitivity of single cells within human myeloma tumors. *Leukemia* **30**, 1094–1102 (2016).

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

Mooi, J. K., Wirapati, P., Asher, R., Lee, C. K., Savas, P., Price, T. J., Townsend, A., Hardingham, J., Buchanan, D., Williams, D., Tejpar, S., Mariadason, J. M. & Tebbutt, N. C. The prognostic impact of consensus molecular subtypes (CMS) and its predictive effects for bevacizumab benefit in metastatic colorectal cancer: Molecular analysis of the AGITG MAX clinical trial. *Annals of Oncology* **29**, 2240–2246 (2018).

Mook, S., Van't Veer, L. J., Rutgers, E. J. T., Piccart-Gebhart, M. J. & Cardoso, F. Individualization of therapy using mammaprint: From development to the MINDACT trial. *Cancer Genomics and Proteomics* **4**, 147–156 (2007).

Morris, L. G. T., Riaz, N., Desrichard, A., Senbabaoglu, Y., Hakimi, A. A., Makarov, V., Reis-Filho, J. S. & Chan, T. A. Pan-cancer analysis of intratumor heterogeneity as a prognostic determinant of survival. *Oncotarget* **7**, 10051–10063 (2016).

Mowat, A. M. & Agace, W. W. Regional specialization within the intestinal immune system. *Nature Reviews Immunology* **14**, 667–685 (2014).

Mueller, M. M. & Fusenig, N. E. Friends or foes - Bipolar effects of the tumour stroma in cancer. *Nature Reviews Cancer* **4**, 839–849 (2004).

Mullins, M., Perreard, L., Quackenbush, J. F., Gauthier, N., Bayer, S., Ellis, M., Parker, J., Perou, C. M., Szabo, A. & Bernard, P. S. Agreement in breast cancer classification between microarray and quantitative reverse transcription PCR from fresh-frozen and formalin-fixed, paraffin-embedded tissues. *Clinical Chemistry* **53**, 1273–1279 (2007).

NanoString Technologies. Prosigna Breast Cancer Prognostic Gene Signature Assay Package Insert. (2016).

National Institute for Health and Care Excellence. Clinical guideline [CG131] Colorectal cancer: diagnosis and management. *NICE Guidance* (2014). Available at: https://www.nice.org.uk/guidance/cg131/chapter/1-Recommendations.

Neeley, E. S., Kornblau, S. M., Coombes, K. R. & Baggerly, K. A. Variable slope normalization of reverse phase protein arrays. *Bioinformatics* **25**, 1384–1389 (2009).

Neugut, A. I., Lin, A., Raab, G. T., Hillyer, G. C., Keller, D., O'Neil, D. S., Accordino, M. K., Kiran, R. P., Wright, J. & Hershman, D. L. FOLFOX and FOLFIRI Use in Stage IV Colon Cancer: Analysis of SEER-Medicare Data. *Clinical Colorectal Cancer* **18**, 133–140 (2019).

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., Hoang, C. D., Diehn, M. & Alizadeh, A. A. Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods* **12**, 453–457 (2015).

Ng, E. K. O., Chong, W. W. S., Jin, H., Lam, E. K. Y., Shin, V. Y., Yu, J., Poon, T. C. W., Ng, S. S. M. & Sung, J. J. Y. Differential expression of microRNAs in plasma of patients with colorectal cancer: a potential marker for colorectal cancer screening. *Gut* **58**, 1375–1381 (2009).

Nielsen, T., Wallden, B., Schaper, C., Ferree, S., Liu, S., Gao, D., Barry, G., Dowidar, N., Maysuria, M., Storhoff, J., Henry, N., Hayes, D., Dietel, M., Johrens, K., Laffert, M., Hummel, M., Blaker, H., Muller, B., Lehmann, A., Denkert, C., Heppner, F., Koch, A., Sers, C., Anagnostopoulos, I., Duffy, M., Crown, J., Gown, A., Wolff, A., Hammond, M., Schwartz, J., Hagerty, K., Allred, D., Cote, R., Dowsett, M., Fitzgibbons, P., Hanna, W., Langer, A., McShane, L., Paik, S., Pegram, M., Perez, E., Press, M., Rhodes, A., Sturgeon, C., Taube, S., Tubbs, R., Vance, G., Vijver, M. van de, Wheeler, T., Hayes, D., Sholl, L., Xiao, Y., Joshi, V., Yeap, B., Cioffredi, L., Jackman, D., Lee, C., Janne, P., Lindeman, N., Weichert, W., Schewe, C., Lehmann, A., Sers, C., Denkert, C., Budczies, J., Stenzinger, A., Joos, H., Landt, O., Heiser, V., Rocken, C., Dietel, M., Perou, C., Sorlie, T., Eisen, M., Rijn, M. van de, Jeffrey, S., Rees, C., Pollack, J., Ross, D., Johnsen, H., Akslen, L., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S., Lonning, P., Borresen-Dale, A., Brown, P., Botstein, D., Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F., Walker, M., Watson, D., Park, T., Hiller, W., Fisher, E., Wickerham, D., Bryant, J., Wolmark, N., Vijver, M. van de, He, Y., Veer, L. van't, Dai, H., Hart, A., Voskuil, D., Schreiber, G., Peterse, J., Roberts, C., Marton, M., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., Velde, T. van der, Bartelink, H., Rodenhuis, S., Rutgers, E., Friend, S., Bernards, R., Arpino, G., Generali, D., Sapino, A., Lucia, M. D., Frassoldati, A., Laurentis, M. D., Paolo, P., Mustacchi, G., Cazzaniga, M., Placido, S. D., Conte, P., Cappelletti, M., Zanoni, V., Antonelli, A., Martinotti, M., Puglisi, F., Berruti, A., Bottini, A., Dogliotti, L., Harris, L., Fritsche, H., Mennel, R., Norton, L., Ravdin, P., Taube, S., Somerfield, M., Hayes, D., Bast, R., Zujewski, J., Kamin, L., Cardoso, F., Veer, L. V., Rutgers, E., Loi, S., Mook, S., Piccart-Gebhart, M., Parker, J., Mullins, M., Cheang, M., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenbush, J., Stijleman, I., Palazzo, J., Marron, J., Nobel, A., Mardis, E., Nielsen, T., Ellis, M., Perou, C., Bernard, P., Nielsen, T., Parker, J., Leung, S., Voduc, D., Ebbert, M., Vickery, T., Davies, S., Snider, J., Stijleman, I., Reed, J., Cheang, M., Mardis, E., Perou, C., Bernard, P., Ellis, M., Geiss, G., Bumgarner, R., Birditt, B., Dahl, T., Dowidar, N., Dunaway, D., Fell, H., Ferree, S., George, R., Grogan, T., James, J., Maysuria, M., Mitton, J., Oliveri, P., Osborn, J., Peng, T., Ratcliffe, A., Webster, P., Davidson, E., Hood, L., Dimitrov, K., Reis, P., Waldron, L., Goswami, R., Xu, W., Xuan, Y., Perez-Ordonez, B., Gullane, P., Irish, J., Jurisica, I., Kamel-Reid, S., Dowsett, M., Sestak, I., Lopez-Knowles, E., Sidhu, K., Dunbier, A., Cowens, J., Ferree, S., Storhoff, J., Schaper, C., Cuzick, J., Gnant, M., Filipits, M., Mlineritsch, B., Dubsky, P., Jakesz, R., Kwasny, W., Fitzal, F., Rudas, M., Knauer, M., Singer, C., Greil, R., Ferree, S., Storhoff, J., Cowens, J., Schaper, C., Liu, S., Nielsen, T., Gnant, M., Filipits, M., Dubsky, P., Rudas, M., Balic, M., Greil, R., Ferree, S., Cowens, J., Schaper, C., Nielsen, T., Simon, R., Paik, S., Hayes, D., Teutsch, S., Bradley, L., Palomaki, G., Haddow, J., Piper, M., Calonge, N., Dotson, W., Douglas, M., Berg, A., Gnant, M., Dowsett, M., Filipits, M., Lopez-Knowles, E., Greil, R., Balic, M., Cowens, J., Nielsen, T., Shaper, C., Sestak, I., Fesl, C., Cuzick, J., Tholen, D., Kallner, A., Kennedy, J., Krouwer, J., Meier, K., Majidzadeh-A, K., Esmaeili, R., Abdoli, N., Szabo, A., Perou, C., Karaca, M., Perreard, L., Palais, R., Quackenbush, J., Bernard, P., Baker, S., Bauer, S., Beyer, R., Brenton, J., Bromley, B., Burrill, J., Causton, H., Conley, M., Elespuru, R., Fero, M., Foy, C., Fuscoe, J., Gao, X., Gerhold, D., Gilles, P., Goodsaid, F., Guo, X., Hackett, J., Hockett, R., Ikonomi, P., Irizarry, R., Kawasaki, E., Kaysser-Kranich, T., Kerr, K., Kiser, G., Koch, W., Lee, K., Liu, C., Liu, Z., Lucas, A., Manohar, C., Miyada, G., Modrusan, Z., Parkes, H., Puri, R., Reid, L., Ryder, T., Salit, M., Samaha, R., Scherf, U., Sendera, T., Setterquist, R., Shi, L., Shippy, R., Soriano, J., Wagar, E., Warrington, J., Williams, M., Wilmer, F., Wilson, M., Wolber, P., Wu, X., Zadro, R., Warrington, J., Corbisier, P., Feilotter, H., Hackett, J., Reid, L., Salit, M., Wagar, E., Williams, P., Wolber, P., Berger, R., Hsu, J., Cronin, M., Sangli, C., Liu, M., Pho, M., Dutta, D., Nguyen, A., Jeong, J., Wu, J., Langone, K., Watson, D., Nuyten, D., Hastie, T., Chi, J., Chang, H., Vijver, M. van de, Chia, S., Bramwell, V., Tu, D., Shepherd, L., Jiang, S., Vickery, T., Mardis, E., Leung, S., Ung, K., Pritchard, K., Parker, J., Bernard, P., Perou, C., Ellis, M., Nielsen, T., Cheang, M., Voduc, K., Tu, D., Jiang, S., Leung, S., Chia, S., Shepherd, L., Levine, M., Pritchard, K., Davies, S., Stijleman, I., Davis, C., Ebbert, M., Parker, J., Ellis, M., Bernard, P., Perou, C., Nielsen, T., Jorgensen, C., Nielsen, T., Bjerre, K., Liu, S., Wallden, B., Balslev, E., Nielsen, D., Ejlertsen, B., Filipits, M., Rudas, M., Jakesz, R., Dubsky, P., Fitzal, F., Singer, C., Dietze, O., Greil, R., Jelen, A., Sevelda, P., Freibauer, C., Muller, V., Janicke, F., Schmidt, M., Kolbl, H., Rody, A., Kaufmann, M., Schroth, W., Brauch, H., Schwab, M., Fritz, P., Weber, K., Feder, I., Hennig, G., Kronenwett, R., Gehrmann, M., Gnant, M., Kronenwett, R., Bohmann, K., Prinzler, J., Sinn, B., Haufe, F., Roth, C., Averdick, M., Ropers, T., Windbergs, C., Brase, J., Weber, K., Fisch, K., Muller, B., Schmidt, M., Filipits, M., Dubsky, P., Petry, C., Dietel, M., Denkert, C., Elloumi, F., Hu, Z., Li, Y., Parker, J., Gulley, M., Amos, K., Troester, M., Mee, B., Carroll, P., Donatello, S., Connolly, E., Griffin, M., Dunne, B., Burke, L., Flavin, R., Rizkalla, H., Ryan, C., Hayes, B., D'Adhemar, C., Banville, N., Faheem, N., Muldoon, C., Gaffney, E., Graham, K., Ge, X., Las, M. D., Tripathi, A., Rosenberg, C., Clare, S., Pardo, I., Mathieson, T., Lillemoe, H., Blosser, R., Choi, M., Sauder, C., Doxey, D., Badve, S., Storniolo, A., Atale, R. & Radovich, M. Analytical validation of the PAM50-based Prosigna Breast Cancer Prognostic Gene Signature Assay and nCounter Analysis System using formalin-fixed paraffin-embedded breast tumor specimens. *BMC Cancer* **14**, 177 (2014).

Northcott, P. A., Shih, D. J. H., Remke, M., Cho, Y. J., Kool, M., Hawkins, C., Eberhart, C. G., Dubuc, A., Guettouche, T., Cardentey, Y., Bouffet, E., Pomeroy, S. L., Marra, M., Malkin, D., Rutka, J. T., Korshunov, A., Pfister, S. & Taylor, M. D. Rapid, reliable, and reproducible molecular sub-grouping of clinical medulloblastoma samples. *Acta Neuropathologica* **123**, 615–626 (2012).

Norton, N., Sun, Z., Asmann, Y. W., Serie, D. J., Necela, B. M., Bhagwate, A., Jen, J., Eckloff, B. W., Kalari, K. R., Thompson, K. J., Carr, J. M., Kachergus, J. M., Geiger, X. J., Perez, E. A. & Thompson, E. A. Gene expression, single nucleotide variant and fusion transcript discovery in archival material from breast tumors. *PLoS ONE* **8**, e81925 (2013).

Nyamundanda, G., Eason, K. & Sadanandam, A. A next generation clustering tool enables identification of functional cancer subtypes with associated biological phenotypes. *bioRxiv* (2017). doi:10.1101/175307

Ogino, S. & Goel, A. Molecular classification and correlates in colorectal cancer. *Journal of Molecular Diagnostics* **10**, 13–27 (2008).

Overman, M. J., Lonardi, S., Wong, K. Y. M., Lenz, H. J., Gelsomino, F., Aglietta, M., Morse, M. A., Van Cutsem, E., McDermott, R., Hill, A., Sawyer, M. B., Hendlisz, A., Neyns, B., Svrcek, M., Moss, R. A., Ledeine, J. M., Cao, Z. A., Kamble, S., Kopetz, S. & André, T. Durable clinical benefit with nivolumab plus ipilimumab in DNA mismatch repair-deficient/microsatellite instability-high metastatic colorectal cancer. *Journal of Clinical Oncology* **36**, 773–779 (2018).

Overman, M. J., McDermott, R., Leach, J. L., Lonardi, S., Lenz, H. J., Morse, M. A., Desai, J., Hill, A., Axelson, M., Moss, R. A., Goldberg, M. V., Cao, Z. A., Ledeine, J. M., Maglinte, G. A., Kopetz, S. & André, T. Nivolumab in patients with metastatic DNA mismatch repair-deficient or microsatellite instability-high colorectal cancer (CheckMate 142): an open-label, multicentre, phase 2 study. *The Lancet Oncology* **18**, 1182–1191 (2017).

Palles, C., Cazier, J.-b., Howarth, K. M., Domingo, E., Jones, A. M., Broderick, P., Kemp, Z., Guarino, E., Salguero, I., Sherborne, A., Chubb, D., Carvajal-carmona, L. G., Ma, Y., Kaur, K., Dobbins, S., Barclay, E., Gorman, M., Martin, L., Kovac, M. B., Humphray, S., Consortium, T. C., Consortium, T. W. G. S., Lucassen, A., Holmes, C. C., Bentley, D., Donnelly, P., Taylor, J., Petridis, C., Roylance, R., Sawyer, E. J., Kerr, D. J., Clark, S., Grimes, J., Kearsey, S. E., Thomas, H. J. W., Mcvean, G., Houlston, R. S. & Tomlinson, I. Articles Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nature Genetics* **45**, 136–144 (2013).

Pawlik, T. M., Raut, C. P. & Rodriguez-Bigas, M. A. Colorectal carcinogenesis: MSI-H versus MSI-L. *Disease Markers* **20**, 199–206 (2004).

Peng, J., Xiao, L. S., Dong, Z. Y., Li, W. W., Wang, K. Y., Wu, D. H. & Liu, L. Potential predictive value of JAK2 expression for Pan-cancer response to PD-1 blockade immunotherapy. *Translational Cancer Research* **7**, 462–471 (2018).

Perou, C. M., Sørlie, T., Eisen, M. B., Rijn, M. van de, Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lønning, P. E., Børresen-Dale, A. L., Brown, P. O. & Botstein, D. Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).

Pond, N., Piccart-gebhart, M. & Brand, M. Mammaprint: a comprehensive review. *Future Oncology* **15**, 207–224 (2019).

Pylayeva-Gupta, Y., Grabocka, E. & Bar-Sagi, D. RAS oncogenes: weaving a tumorigenic web. *Nature Reviews Cancer* **11**, 761–774 (2011).

Racle, J., Jonge, K. de, Baumgaertner, P., Speiser, D. E. & Gfeller, D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *eLife* **6**, e26476 (2017).

Ragulan, C., Eason, K., Fontana, E., Nyamundanda, G., Poudel, P., Lawlor, R. T., Rio, M. D., Si-lin, K., Siew, T. W., Sclafani, F., Begum, R., Mendes, L. S. T., Martineau, P., Tan, I. B., Cunningham, D. & Sadanandam, A. Analytical Validation of Multiplex Biomarker Assay to Stratify Colorectal Cancer into Molecular Subtypes. *Scientific Reports* **9**, 7665 (2019).

Rao, J. N. & Wang, J.-Y. Intestinal stem cells. In *Regulation of Gastrointestinal Mucosal Growth* (eds. Rao, J. N. & Wang, J.-Y.) (Morgan & Claypool Life Sciences, 2010).

Raynaud, F., Mina, M., Tavernari, D. & Ciriello, G. Pan-cancer inference of intra-tumor heterogeneity reveals associations with different forms of genomic instability. *PLoS Genetics* **14**, 1–18 (2018).

R Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing* (2017). Available at: https://www.r-project.org/.

Reis, P. P., Waldron, L., Goswami, R. S., Xu, W., Xuan, Y., Perez-Ordonez, B., Gullane, P., Irish, J., Jurisica, I. & Kamel-Reid, S. mRNA transcript quantification in archival samples using multiplexed, color-coded probes. *BMC Biotechnology* **11**, 46 (2011).

Reynies, A. de & Guinney, J. CMSclassifier: prediction of the Consensus Molecular Subtype (CMS) of colorectal carcinomas based on log2-scaled Gene Expression Profiles (GEP). *Sage Bionetworks* (2015). Available at: https://github.com/Sage-Bionetworks/CMSclassifier.

Roepman, P., Schlicker, A., Tabernero, J., Majewski, I., Tian, S., Moreno, V., Snel, M. H., Chresta, C. M., Rosenberg, R., Nitsche, U., Macarulla, T., Capella, G., Salazar, R., Orphanides, G., Wessels, L. F. A., Bernards, R. & Simon, I. M. Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition. *International Journal of Cancer* **134**, 552–562 (2013).

Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., Ha, G., Aparicio, S., Bouchard-Côté, A. & Shah, S. P. PyClone: Statistical inference of clonal population structure in cancer. *Nature Methods* **11**, 396–398 (2014).

Ryan, E., Sheahan, K., Creavin, B., Mohan, H. M. & Winter, D. C. The current value of determining the mismatch repair status of colorectal cancer: A rationale for routine testing. *Critical Reviews in Oncology/Hematology* **116**, 38–57 (2017).

Sadanandam, A., Lyssiotis, C. A., Homicsko, K., Collisson, E. A., Gibb, W. J., Wullschleger, S., Ostos, L. C. G., Lannon, W. A., Grotzinger, C., Del Rio, M., Lhermitte, B., Olshen, A. B., Wiedenmann, B., Cantley, L. C., Gray, J. W. & Hanahan, D. A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nature Medicine* **19**, 619–625 (2013).

Sadanandam, A., Wullschleger, S., Lyssiotis, C. A., Grotzinger, C., Barbi, S., Bersani, S., Korner, J., Wafy, I., Mafficini, A., Lawlor, R. T., Simbolo, M., Asara, J. M., Blaker, H., Cantley, L. C., Wiedenmann, B., Scarpa, A. & Hanahan, D. A cross-species analysis in pancreatic neuroendocrine tumors reveals molecular subtypes with distinctive clinical, metastatic, developmental, and metabolic characteristics. *Cancer Discovery* **5**, 1296–1313 (2015).

Salazar, R., Roepman, P., Capella, G., Moreno, V., Simon, I., Dreezen, C., Lopez-doriga, A., Santos, C., Marijnen, C., Westerga, J., Bruin, S. & Kerr, D. Gene Expression Signature to Improve Prognosis Prediction of Stage II and III Colorectal Cancer. *Journal of Clinical Oncology* **29**, 17–24 (2011).

Sandmeier, D., Benhattar, J., Martin, P. & Bouzourene, H. Serrated polyps of the large intestine: A molecular study comparing sessile serrated adenomas and hyperplastic polyps. *Histopathology* **55**, 206–213 (2009).

Sánchez-Navarro, I., Gámez-Pozo, A., González-Barón, M., Pinto-Marín, Á., Hardisson, D., López, R., Madero, R., Cejas, P., Mendiola, M., Espinosa, E. & Vara, J. Á. F. Comparison of gene expression profiling by reverse transcription quantitative PCR between fresh frozen and formalin-fixed, paraffin-embedded breast cancer tissues. *BioTechniques* **48**, 389–397 (2010).

Schleuter, D., Daufresne, M., Massol, F. & Argillier, C. A user's guide to functional diversity indices. *Ecological Monographs* **80**, 469–484 (2010).

Schlicker, A., Beran, G., Chresta, C. M., McWalter, G., Pritchard, A., Weston, S., Runswick, S., Davenport, S., Heathcote, K., Castro, D. A., Orphanides, G., French, T. & Wessels, L. F. A. Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines. *BMC Medical Genomics* **5**, 1–15 (2012).

Scott, D. W., Wright, G. W., Williams, P. M., Lih, C.-J., Walsh, W., Jaffe, E. S., Rosenwald, A., Campo, E., Chan, W. C., Connors, J. M., Smeland, E. B., Mottok, A., Braziel, R. M., Ott, G., Delabie, J., Tubbs, R. R., Cook, J. R., Weisenburger, D. D., Greiner, T. C., Glinsmann-Gibson, B. J., Fu, K., Staudt, L. M., Gascoyne, R. D. & Rimsza, L. M. Determining cell-of-origin subtypes of diffuse large B-cell lymphoma using gene expression in formalin-fixed paraffin-embedded tissue. *Blood* **123**, 1214–1217 (2014).

Sharif, S. & O'Connell, M. J. Gene Signatures in Stage II Colon Cancer: A Clinical Review. *Current Colorectal Cancer Reports* **8**, 225–231 (2012).

Shen, R., Mo, Q., Schultz, N., Seshan, V. E., Olshen, A. B., Huse, J., Ladanyi, M. & Sander, C. Integrative subtype discovery in glioblastoma using iCluster. *PLoS ONE* **7**, e35236 (2012).

Shen, R., Olshen, A. B. & Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25**, 2906–2912 (2009).

Siegmund, K. D., Marjoram, P., Woo, Y.-J., Tavaré, S. & Shibata, D. Inferring clonal expansion and cancer stem cell dynamics from DNA methylation patterns in colorectal cancers. *Proceedings of the National Academy of Sciences* **106**, 4828–4833 (2009).

Simons, C. C. J. M., Hughes, L. A. E., Smits, K. M., Khalid-de Bakker, C. A., Bruïne, A. P. de, Carvalho, B., Meijer, G. A., Schouten, L. J., Brandt, P. A. van den, Weijenberg, M. P. & Engeland, M. van. A novel classification of colorectal tumors based on microsatellite instability, the CpG island methylator phenotype and chromosomal instability: Implications for prognosis. *Annals of Oncology* **24**, 2048–2056 (2013).

Slattery, M. L., Herrick, J. S., Mullany, L. E., Valeri, N., Stevens, J., Caan, B. J., Samowitz, W. & Wolff, R. K. An evaluation and replication of miRNAs with disease stage and colorectal cancer-specific mortality. *International Journal of Cancer* **137**, 428–438 (2015).

Song, N., Pogue-Geile, K. L., Gavin, P. G., Yothers, G., Rim Kim, S., Johnson, N. L., Lipchick, C., Allegra, C. J., Petrelli, N. J., O'Connell, M. J., Wolmark, N. & Paik, S. Clinical outcome from oxaliplatin treatment in stage II/III colon cancer according to intrinsic subtypes: Secondary analysis of NASBP C-07/NRG oncology randomized clinical trial. *JAMA Oncology* **2**, 1162–1169 (2016).

Sorich, M. J., Wiese, M. D., Rowland, A., Kichenadasse, G., Mckinnon, R. A. & Karapetis, C. S. Extended RAS mutations and anti-EGFR monoclonal antibody survival benefit in metastatic colorectal cancer : a meta-analysis of randomized, controlled trials. *Annals of Oncology* **26**, 13–21 (2015).

Sottoriva, A., Kang, H., Ma, Z., Graham, T. A., Salomon, M. P., Zhao, J., Marjoram, P., Siegmund, K., Press, M. F., Shibata, D. & Curtis, C. A Big Bang model of human colorectal tumor growth. *Nature Genetics* **47**, 209–216 (2015).

Stange, D. E., Engel, F., Longerich, T., Koo, B. K., Koch, M., Delhomme, N., Aigner, M., Toedt, G., Schirmacher, P., Lichter, P., Weitz, J. & Radlwimmer, B. Expression of an ASCL2 related stem cell signature and IGF2 in colorectal cancer liver metastases with 11p15 . 5 gain. *Gut* **59**, 1236–1244 (2010).

Stein, D. E. Colorectal cancer. (2019). Available at: https://bestpractice.bmj.com/topics/en-gb/258.

Sturrock, P., Liebmann, J., Karam, A., Pieters, R. & Kurian, E. Colorectal cancer. In *Cancer Concepts: A Guidebook for the Non-Oncologist* (2015).

Suzuki, A., Matsushima, K., Makinoshima, H., Sugano, S., Kohno, T., Tsuchihara, K. & Suzuki, Y. Single-cell analysis of lung adenocarcinoma cell lines reveals diverse expression patterns of individual cells invoked by a molecular target drug treatment. *Genome Biology* **16**, 66 (2015).

Tan, A. C., Naiman, D. Q., Xu, L., Winslow, R. L. & Geman, D. Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics* **21**, 3896–3904 (2005).

Tang, H., Lv, W., Sun, W., Bi, Q. & Hao, Y. MiR-505 inhibits cell growth and EMT by targeting MAP3K3 through the AKT-NFκB pathway in NSCLC cells. *International Journal of Molecular Medicine* **43**, 1203–1216 (2019).

The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).

The GenePattern Team. ssGSEAProjection. *Broad Institute* (2013). Available at: http://software.broadinstitute.org/cancer/software/genepattern/modules/docs/ssGSEAProjection/4.

Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences* **99**, 6567–6572 (2002).

Tong, M., Zheng, W., Li, H., Li, X., Ao, L., Shen, Y., Liang, Q., Li, J., Hong, G., Yan, H., Cai, H., Li, M., Guan, Q. & Guo, Z. Multi-omics landscapes of colorectal cancer subtypes discriminated by an individualized prognostic signature for 5-fluorouracil-based chemotherapy. *Oncogenesis* **5**, e242 (2016).

Toyota, M., Ahuja, N., Ohe-Toyota, M., Herman, J. G., Baylin, S. B. & Issa, J.-P. J. CpG island methylator phenotype in colorectal cancer. *Medical Sciences* **96**, 8681–8686 (1999).

Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* **98**, 5116–5121 (2001).

Uchi, R., Takahashi, Y., Niida, A., Shimamura, T., Hirata, H., Sugimachi, K., Sawada, G., Iwaya, T., Kurashige, J., Shinden, Y., Iguchi, T., Eguchi, H., Chiba, K., Shiraishi, Y., Nagae, G., Yoshida, K., Nagata, Y., Haeno, H., Yamamoto, H., Ishii, H., Doki, Y., Iinuma, H., Sasaki, S., Nagayama, S., Yamada, K., Yachida, S., Kato, M., Shibata, T., Oki, E., Saeki, H., Shirabe, K., Oda, Y., Maehara, Y., Komune, S., Mori, M., Suzuki, Y., Yamamoto, K., Aburatani, H., Ogawa, S., Miyano, S. & Mimori, K. Integrated multiregional analysis proposing a new model of colorectal cancer evolution. *PLoS Genetics* **12**, e1005778 (2016).

U.S. Food & Drug Administration. Nucleic Acid Based Tests. (2019). Available at: https://www.fda.gov/medical-devices/vitro-diagnostics/nucleic-acid-based-tests.

Van Dongen, S. *A new cluster algorithm for graphs.* (Centrum voor Wiskunde en Informatica, 1998).

Vasaikar, S., Huang, C., Wang, X., Petyuk, V. A., Savage, S. R., Wen, B., Dou, Y., Zhang, Y., Shi, Z., Arshad, O. A., Gritsenko, M. A., Zimmerman, L. J., McDermott, J. E., Clauss, T. R., Moore, R. J., Zhao, R., Monroe, M. E., Wang, Y. T., Chambers, M. C., Slebos, R. J. C., Lau, K. S., Mo, Q., Ding, L., Ellis, M., Thiagarajan, M., Kinsinger, C. R., Rodriguez, H., Smith, R. D., Rodland, K. D., Liebler, D. C., Liu, T., Zhang, B., Ellis, M. J. C., Bavarva, J., Borucki, M., Elburn, K., Hannick, L., Vatanian, N., Payne, S. H., Carr, S. A., Clauser, K. R., Gillette, M. A., Kuhn, E., Mani, D. R., Cai, S., Ketchum, K. A., Thangudu, R. R., Whiteley, G. A., Paulovich, A., Whiteaker, J., Edward, N. J., Madhavan, S., McGarvey, P. B., Chan, D. W., Shih, I. M., Zhang, H., Zhang, Z., Zhu, H., Skates, S. J., White, F. M., Mertins, P., Pandey, A., Slebos, R. J. C., Boja, E., Hiltke, T., Mesri, M., Rivers, R. C., Stein, S. E., Fenyo, D., Ruggles, K., Levine, D. A., Oberti, M., Rudnick, P. A., Snyder, M., Tabb, D. L., Zhao, Y., Chen, X., Ransohoff, D. F., Hoofnagle, A., Sanders, M. E., Wang, Y., Davies, S. R., Townsend, R. R. & Watson, M. Proteogenomic analysis of human colon cancer reveals new therapeutic opportunities. *Cell* **177**, 1035–1049 (2019).

Veldman-Jones, M. H., Lai, Z., Wappett, M., Harbron, C. G., Barrett, J. C., Harrington, E. A. & Thress, K. S. Reproducible , Quantitative , and Flexible Molecular Subtyping of Clinical DLBCL Samples Using the NanoString nCounter System. *Clinical Cancer Research* **21**, 2367–2378 (2015).

Vermeulen, J., De Preter, K., Lefever, S., Nuytens, J., De Vloed, F., Derveaux, S., Hellemans, J., Speleman, F. & Vandesompele, J. Measurable impact of RNA quality on gene expression results from quantitative PCR. *Nucleic Acids Research* **39**, e63 (2011).

Vieira, A. F. & Schmitt, F. An update on breast cancer multigene prognostic tests—emergent clinical biomarkers. *Frontiers in Medicine* **5**, 1–12 (2018).

Wallden, B., Storhoff, J., Nielsen, T., Dowidar, N., Schaper, C., Ferree, S., Liu, S., Leung, S., Geiss, G., Snider, J., Vickery, T., Davies, S. R., Mardis, E. R., Gnant, M., Sestak, I., Ellis, M. J., Perou, C. M., Bernard, P. S. & Parker, J. S. Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. *BMC Medical Genomics* **8**, 54 (2015).

Wang, C., Mao, J., Redfield, S., Mo, Y., Lage, J. M. & Zhou, X. Systemic distribution, subcellular localization and differential expression of sphingosine-1-phosphate receptors in benign and malignant human tissues. *Experimental and Molecular Pathology* **97**, 259–265 (2014).

Wang, L., Zhu, J. & Zou, H. The doubly regularized support vector machine. *Statistica Sinica* **16**, 589–615 (2006).

Watanabe, T., Kobunai, T., Yamamoto, Y., Matsuda, K., Ishihara, S. & Nozawa, K. Chromosomal Instability (CIN) Phenotype, CIN High or CIN Low, Predicts Survival for Colorectal Cancer. *Journal of Clinical Oncology* **30**, 2256–2264 (2012).

Westwood, M., Asselt, T. van, Ramaekers, B., Whiting, P., Joore, M., Armstrong, N., Noake, C., Ross, J., Severens, J. & Kleijnen, J. KRAS mutation testing of tumours in adults with metastatic colorectal cancer: a systematic review and cost-effectiveness analysis. *Health Technology Assessment* **18**, 1–132 (2014).

Yu, J., Li, N., Wang, X., Ren, H., Wang, W., Wang, S., Song, Y., Liu, Y., Li, Y., Zhou, X., Luo, A., Liu, Z. & Jin, J. Circulating serum microRNA-345 correlates with unfavorable pathological response to preoperative chemoradiotherapy in locally advanced rectal cancer. *Oncotarget* **7**, 64233–64243 (2016).

Yu, K., Su, N., Le, T. D., Liu, L., Xu, T., Wang, H., Zhang, J., Gui, J., Zhang, W. & Li, J. miR-BaseConverter: an R/Bioconductor package for converting and retrieving miRNA name, accession, sequence and family information in different versions of miRBase. *BMC Bioinformatics* **19**, 514 (2018).

Yu, Y., Carey, M., Pollett, W., Green, J., Dicks, E., Parfrey, P., Yilmaz, Y. E. & Savas, S. The long-term survival characteristics of a cohort of colorectal cancer patients and baseline variables associated with survival outcomes with or without time-varying effects. *BMC Medicine* **17**, 1–12 (2019).

Yuan, Y. Spatial heterogeneity in the tumor microenvironment. *Cold Spring Harbor Perspectives in Medicine* **6**, a026583 (2017).

Zhan, T., Rindtorff, N. & Boutros, M. Wnt signaling in cancer. *Oncogene* **36**, 1461–1473 (2017).

Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M. C., Zimmerman, L. J., Shaddox, K. F., Kim, S., Davies, S. R., Wang, S., Wang, P., Kinsinger, C. R., Rivers, R. C., Rodriguez, H., Townsend, R. R., Ellis, M. J. C., Carr, S. A., Tabb, D. L., Coffey, R. J., Slebos, R.

J. C., Liebler, D. C. & the NCI CPTAC. Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**, 382–387 (2014).

Zhu, J., Chen, L., Zou, L., Yang, P., Wu, R., Mao, Y., Zhou, H., Li, R., Wang, K., Wang, W., Hua, D. & Zhang, X. MiR-20b, -21, and -130b inhibit PTEN expression resulting in B7-H1 over-expression in advanced colorectal cancer. *Human Immunology* **75**, 348–353 (2014).